Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit[*]

Sule Alan[†]

Teodora Boneva

Seda Ertac

Abstract

We show that grit, a skill that has been shown to be highly predictive of achievement, is malleable in childhood and can be fostered in the classroom environment. We evaluate a randomized educational intervention implemented in two independent elementary school samples. Outcomes are measured via a novel incentivized real effort task and performance in standardized tests. We find that treated students are more likely to exert effort to accumulate task-specific ability, and hence, more likely to succeed. In a follow up 2.5 years after the intervention, we estimate an effect of about 0.2 standard deviations on a standardized math test.
JEL Codes: C91; C93; D03; I28.

[*] We thank the editor, Lawrence Katz, and three anonymous referees. For comments and discussions, we are grateful to Orazio Attanasio, Marco Castillo, Thomas Crossley, Thomas Dohmen, Angela Duckworth, Armin Falk, James Heckman, John List, Gautam Rao, and many seminar and conference participants. Elif Kubilay, Nergis Zaim, Banu Donmez, Mert Gumren and Enes Duysak, as well as numerous other students provided excellent research assistance. This paper is part of a field project partially funded by the Turkish division of the ING Bank and Koç University. Boneva acknowledges financial support from the British Academy. Ertac acknowledges financial support from the TUBA-GEBIP and BAGEP programs. All errors are our own. The two RCTs evaluated in this paper are registered with the AEA Trial Registry (RCT ID: AEARCTR-0003317 and AEARCTR-0003327). Koç University IRB approval no: 2013.067.IRB.32.

[†] University of Essex, Bilkent University and J-PAL, salancrossley@gmail.com.

I Introduction

The growing literature on human capital accumulation has emphasized the importance of non-cognitive skills in explaining individual differences in achievement in various economic and social domains (Heckman *et al.*, 2006; Borghans *et al.*, 2008). These skills encompass a broad range of individual character traits, often measured via standardized questionnaires by psychologists and, more recently, via incentivized experimental elicitation techniques by economists. Non-cognitive skills such as patience, self-control and grit have been shown to be highly predictive of outcomes ranging from educational attainment, occupational and financial success to criminal activity and health outcomes; see Heckman *et al.* (2006); Almlund *et al.* (2011); Dohmen *et al.* (2011); Sutter *et al.* (2013); Heckman *et al.* (2011); Moffit *et al.* (2011); Castillo *et al.* (2011); Golsteyn *et al.* (2013). In fact, the predictive power of non-cognitive skills appears to rival that of cognitive skills (Roberts *et al.*, 2007; Kautz *et al.*, 2014). More importantly from a policy standpoint, there is now ample evidence suggesting that these important skills are malleable especially in the childhood period and can be fostered through educational interventions (Almlund *et al.*, 2011; Kautz *et al.*, 2014).[1]

Among these skills, "grit" is the focus of this paper. Grit is generally defined as perseverance toward a set goal and it is closely related to conscientiousness. Grit has been shown to be associated with college GPAs and educational attainment. It also predicts retention in different contexts: Grittier students are more likely to graduate from high school, grittier employees are more likely to keep their jobs, grittier soldiers are more likely to be retained in the army and grittier men are more likely to remain married; see Duckworth *et al.* (2007); Duckworth and Quinn (2009); Maddie *et al.* (2012); Eskreis-Winkler *et al.* (2014). Beliefs are likely to play an important role in producing "gritty" behavior. An individual will set ambitious performance goals and persevere in response to failures if her perceived productivity of effort is sufficiently high. While confidence about one's existing skills can be important in such decisions, optimistic beliefs about the role of effort in success are also likely to be crucial. The latter is related to the concept of "growth mindset" (see Dweck, 2006, Yeager and Dweck, 2012). An individual who holds this mindset believes that skills can be developed over time by exerting effort (e.g. by continued practice). Such an individual will be less discouraged by and more likely to persevere after early failures.

Given the central question of how to motivate individuals to work harder in educational and occupational settings, it is important to understand the nature of grit, and

[1] Well-known examples of early childhood and elementary school programs include the Perry Preschool Program (Heckman *et al.*, 2010, 2013), the Abecedarian Program (Campbell *et al.*, 2014; Garcia *et al.*, 2016), and the project STAR (Schanzenbach, 2006; Dee and West, 2011; Chetty *et al.*, 2011).

to explore ways of enhancing it. In this paper, we evaluate a randomized educational intervention that aims to foster grit in the classroom environment. We conjecture that an intervention that instills in children optimistic beliefs about the productivity of effort and encourages them to persevere through setbacks will increase the motivation to undertake and keep working at challenging but rewarding tasks, eventually resulting in higher achievement. The intervention involves a teacher training program that focuses on three interrelated ideas underlying grit: growth mindset, perseverance through failures, and goal-setting. The program is supported by a specifically designed curriculum to be implemented in class by students' own teachers. This curriculum consists of animated videos, mini case studies and classroom activities that highlight (i) the plasticity of the human brain against the notion of innately fixed ability, (ii) the role of effort in enhancing skills and achieving goals, (iii) the importance of a constructive interpretation of failures and therefore perseverance, and (iv) the importance of goal setting. Teachers in treated schools participate in a training seminar to learn how to implement the program. The materials are shaped by a multidisciplinary team of education consultants and elementary school teachers. The intervention also has a significant pedagogical component: teachers are encouraged to adopt a teaching philosophy that emphasizes the role of effort in everyday classroom practices, e.g. while giving performance feedback and interpreting test results.

We evaluate the impact of this program using two independent samples from a total of 52 state-run elementary schools in Istanbul, Turkey. Within each sample, the intervention is randomized across schools in which at least one teacher was willing to participate in the program. We measure the outcomes through a multi-faceted methodology that includes a novel incentivized real effort task, grades and objective test scores. The incentivized real effort task is designed to elicit core aspects of grit; challenge seeking, perseverance through setbacks, goal setting and the propensity to engage in effortful behavior to accumulate skill. Specifically, we elicit students' choices between a challenging high-reward and an easy low-reward task, and the dynamic response of this choice to negative performance feedback. The task also involves a temporal component, which allows us to observe skill accumulation in the challenging task through practice. In addition to experimental choices and outcomes, we administer standardized tests to measure mathematics and verbal (Turkish) skills. We also measure students' beliefs about the malleability of ability and the role of effort in achievement, as well as self-reported attitudes and behaviors regarding perseverance, using pre- and post-treatment questionnaires. We collect this information from over 3,200 4th grade students in total, by visiting 110 classrooms multiple times.

In both samples, our results reveal a striking impact of the intervention on students' behaviors and outcomes in the real effort task. In particular, we find treated students to be significantly more likely to opt for a difficult high-reward task than an easier low-reward alternative. Treated students are also significantly more likely to re-attempt the difficult task after receiving negative performance feedback. The design of our incentivized task additionally allows us to investigate whether treated students are more likely to aim for succeeding in the difficult task when they are given the opportunity to accumulate task-specific skill. When given time to acquire the skill needed to succeed in the difficult task, treated students are significantly more likely to set the goal of succeeding in the difficult task. They are also significantly more likely to achieve this goal. More specifically, they are about 8 to 10 percentage points more likely to actually succeed in the difficult task, and consequently, they collect about 16% to 26% higher rewards than students in the control group. These findings suggest that treated students are more likely to set ambitious goals, engage in skill accumulating activities, and end up with higher success as a result.

The positive effects we estimate in the incentivized task also extend to achievement outcomes. Although we do not estimate a significant treatment effect on subjective grades given by teachers, we find that treated students perform significantly better in an objective mathematics test. Tests conducted immediately after the program reveal a large treatment effect (about 0.31 standard deviations) on math and a smaller and less precise effect (about 0.13 standard deviations) on verbal performance. Particularly encouraging is that the program has remarkably persistent effects on math scores: In a follow-up conducted 2.5 (1.5) years after the implementation of the program in the first (second) sample, we estimate an effect of about 0.23 (0.19) standard deviations on an objective math test. For verbal scores, however, the estimated short-term effect seems to have dissipated in both samples. We find that the estimated treatment effects on both behavioral and achievement outcomes are remarkably similar across the two independent samples. The replicability and persistence of our results is encouraging and clears the path for a potential scale-up.

Our study relates to the growing number of studies that investigate the impact of growth mindset interventions on short-term academic achievement. These interventions are typically administered with the help of short videos that illustrate the plasticity of the brain and highlight the idea that intellectual ability is not fixed but can be developed (see, e.g., Dweck, 2006, Yeager and Dweck, 2012). While the early studies produced very promising results (e.g. Aronson *et al.*, 2002, Good *et al.*, 2003, Blackwell *et al.*, 2007), a recent meta-analysis concludes that the overall effects of such interventions are estimated to be weak, and that growth mindset interventions may only benefit students of low socioeconomic status or students who are academically at risk/low-achieving (Sisk *et al.*, 2018). Three recent studies that use large samples of students all reach this conclusion. Paunesku *et al.* (2015) evaluate the effect of a 45-minute mindset intervention in a sample of 9-12 graders and find that the intervention only has a positive impact on the end-of-semester GPA of students at risk of dropping out of high school. Similarly, Yeager *et al.* (2016b) investigate the impact of a two-period mindset intervention in a sample of 9th graders and find that the intervention only raises the end-of-semester GPA of previously low-achieving students. In a recent study, Yeager *et al.* (2018) randomly assign a 50-minute mindset intervention to 9th grade students in a representative sample of 65 U.S. public schools and find that the intervention significantly increases the end-of-year GPA of previously low-achieving students, while it has no effect on high-achieving students. Other prominent studies published to date include Sriram (2014), Yeager *et al.* (2014), Yeager *et al.* (2016a) and Bettinger *et al.* (2018). While some of these studies find positive effects, other studies do not find significant effects on achievement outcomes. Our study differs from these mindset interventions in three respects. First, our intervention is delivered by teachers and it is considerably more intense in terms of duration as well as content. Trained teachers spend twelve 2-hour sessions covering and discussing the material but the intervention does not merely consist of covering a curriculum. It also involves a significant pedagogical component. Specifically, teachers are encouraged to apply the ideas in everyday teaching and classroom activities. Our intervention aims to change children's beliefs and behaviors through the classroom practices of teachers, and places the focus more directly on encouraging actual perseverant behavior in class, in addition to introducing students to a set of ideas. This, in fact, also ensures that treated students are exposed to the concepts and ideas for the duration of an entire school year, and not only within the limited project hours. Second, we conduct a long-term follow-up for both of our samples with respect to objective test scores. Third, as part of this evaluation, we propose a novel incentivized task to measure the core aspects of grit: challenge seeking, perseverance after negative feedback, goal-setting and willingness to accumulate ability over time.

Our study also relates to the literature on how student coaching and goal-setting interventions affect student achievement in college. Bettinger and Baker (2014) test the effectiveness of individualized student coaching and find that having a personal coach significantly increases student retention. Oreopoulos and Petronijevic (2018) test the effectiveness of three different interventions (i.e. an online goal-setting exercise, a text message campaign and a personal coach) and find that while the personal coaching program has large effects on student achievement, the low-cost interventions relying on technology have no effects on academic outcomes. Dobronyi *et al.* (2017) test the effectiveness of two online goal-setting interventions, one of which includes a growth mindset component, and also find no evidence of an effect on student achievement or drop out. Oreopoulos *et al.* (2018) evaluate an online planning exercise aimed at increasing study time and find that while the intervention has some impact on the amount of time students study, it has no effect on academic outcomes.

We show that a targeted educational intervention implemented by students' own teachers in the natural classroom environment, can produce remarkable effects on behaviors related to grit, and on success and payoffs in an incentivized real effort task. The effects extend to actual achievement outcomes and persist over time. Given the pivotal role of non-cognitive skills for academic achievement and labor market success (Duckworth *et al.*, 2007; Almlund *et al.*, 2011; Kautz *et al.*, 2014), this evidence is of utmost policy importance. Our results provide an affirmative answer to the question of whether grit is malleable, adding to the literature showing that preferences, non-cognitive skills and outcomes can be influenced through childhood interventions (e.g. Fryer, 2011; Bettinger *et al.*, 2012; Levitt *et al.*, 2016; Alan and Ertac, 2018; Kosse *et al.*, forthcoming).[2] Our intervention also highlights a particular low-cost way of fostering non-cognitive skills in the natural environment of the classroom. Being able to achieve such an impact in the school environment offers hope for reducing persistent achievement gaps observed in many countries, where educational policy actions aiming to enhance family inputs tend to face challenges in engaging families of low socioeconomic strata.

The paper is organized as follows. Section II presents details on the design and implementation of the educational intervention, and on the measurement of the different outcome variables of interest. Section III contains details on the data, while Section IV presents the results. Section V provides a brief discussion on potential channels and Section VI concludes. All Appendix material can be found in the Online Appendix.

II Design and Outcome Measurement

II.A. *Content of the Intervention*

The Turkish Ministry of Education encourages all elementary and post-elementary schools to participate in extra-curricular projects offered by the private sector, NGOs, the government, and international organizations. After being examined and endorsed by the

[2] Alan and Ertac (forthcoming) show that the intervention evaluated in the current paper also mitigates the well-documented gender gap in competition, while Alan and Ertac (2017) show that the intervention has an effect on patterns of altruism, such that there is less sympathy towards the unsuccessful.

Ministry, these projects are made available to schools. Participation in these projects is at the discretion of teachers. The Ministry allows up to 5 lecture hours per week for project-related classroom activities. The program we evaluate in this paper was implemented as an extra-curricular project of this type.

The program involves covering a specifically designed curriculum by children's own trained teachers. The curriculum consists of animated videos, mini case studies and classroom activities that highlight (i) the plasticity of the human brain against the notion of innate ability, (ii) the role of effort in enhancing skills and achieving goals, (iii) the importance of a constructive interpretation of setbacks and failures, and (iv) the importance of goal setting. The aim of the program is to expose students to a worldview in which any one of them can set goals in an area of their interest and can work toward these goals by exerting effort. The materials highlight the idea that to achieve goals, it is imperative to avoid interpreting immediate failures as a lack of innate ability or intelligence. This worldview embraces any productive area of interest, whether it be music, art, science or sports. While the target concepts of the educational materials were determined by the scientific team, specific contents (e.g. scripts) were shaped with input from an interdisciplinary team of education psychologists, a group of voluntary elementary school teachers, children's story writers and media animation artists, according to the age and cognitive capacity of the students. A minimum of 10 sessions were recommended to the teachers to complete the curriculum. Most teachers reported that they spent at least 2 hours/week on the project over the course of twelve weeks.

To give an example of the material covered, in an animated video, two students who hold opposite views on the malleability of ability engage in a dialog. The student who believes that ability is innate and therefore there is no scope for enhancing ability through effort, points out that the setbacks she experiences are reminders of the fact that she is not intelligent. Following this remark, the student who holds the opposite view replies that she knows that setbacks are usually inevitable on the way to success; she interprets them as opportunities to learn, and therefore, they do not discourage her. The video contains further conversations between these two students on similar ideas such as the importance of sustained effort in achieving one's long-term goals. Training materials also include stories in the form of mini-case studies containing similar ideas in different contexts. In addition to material about the malleability of abilities, the intervention contains materials that highlight the importance of goal-setting, and address issues that tend to hinder perseverance, such as fear of failure, fear of math and other challenging tasks. Visual materials and stories are supplemented by classroom activities created and supervised by teachers, based on general suggestions and guidelines put forward in the teacher training seminars. For example, in a large number of schools, students prepared colorful posters that contain famous phrases of renowned individuals pertaining to the importance of grit and perseverance. These posters were exhibited in these schools in the week during which the lives of famous scientists and explorers in history were covered as part of the life sciences curriculum.[3]

Volunteering teachers who were assigned to the treatment group participated in a training seminar to learn how to implement the program. The seminar was carried out over the course of one day. In the seminar, instructors first introduced the concepts and their importance for academic achievement. They then guided the teachers through the materials and suggested classroom activities with the help of education consultants. The seminar was structured in an interactive manner and instructors aimed to actively engage the teachers in different activities to exemplify the different concepts. In addition to receiving detailed instructions on how to cover the curriculum, teachers were encouraged to adopt the ideas put forward in the materials as part of a teaching philosophy. To do this, they were given various pedagogical guidelines. These include praising students' effort, championing perseverant behavior and positive attitudes toward learning, rather than just praising good outcomes. Teachers were also encouraged not to praise a successful student in a way which would imply that the student possesses superior innate ability. Rather they were advised to highlight the role of effort in success. In this sense, the intervention is not merely a set of materials to be covered in a specified period of time, but rather an attempt to change the mindset of children by changing the classroom practices of the teachers. To assess how successful this attempt was, we conducted an anonymous survey among teachers at the end of the academic year and asked about their views on the ideas put forward in the materials. More than 95% of all teachers report that they agree with the ideas conveyed by the training and 93% report having implemented the program. It is important to stress that the intervention is not prescriptive in nature. Because we were concerned about the optimality of perseverance in different contexts at the design stage, we took great care to avoid normative propositions regarding gritty behavior both in the curriculum materials and in pedagogical guidelines.

II.B. Evaluation Design

Turkey has a two-tier education system where the children from middle and higher socioeconomic strata tend to attend well-resourced private schools. Because our sample covers only state-funded schools in remote areas of Istanbul, it predominantly repre-

[3] Oversight of the ministry and the input received from independent school teachers in preparation of the materials ensured that all activities and reading materials complemented the existing curricula. A summary of the curriculum can be found in Online Appendix C.

sents Turkey's lower socioeconomic segment. The program we evaluate is the second arm of a two-arm randomized-controlled trial initiated in Spring 2013. It was implemented as two independent studies, giving us two independent evaluation samples. In both samples, the intervention was randomized across schools in which at least one teacher stated his/her willingness to participate in the program.

In the first study, we randomly allocated 15 schools to Initial Treatment (IT), 10 schools to Control-then-Treatment (CT) and the remaining 12 schools to Pure Control (PC). As soon as the baseline data were collected in Spring 2013, the first arm of the RCT, referred to as the "patience" arm, was implemented. This involved training the teachers in the initial treatment (IT) group to cover a curriculum that aims to encourage forward-looking behavior. In May-June 2013, we collected our first follow-up data and measured the effect of the patience treatment on the intertemporal choices of children.[4] In Fall 2013, our IT group received the "grit" intervention, while the CT group (9 schools) received the "patience" intervention. Note that the IT group had now received two treatments (grit and patience) combined. The CT group never received the grit intervention and remained as the "patience only" treatment. The results of the evaluation of the patience arm with respect to children's intertemporal decisions and behavioral conduct are reported in Alan and Ertac (2018). In the current paper, we compare treated students (in the 15 IT schools that received grit+patience) and control students (in 9 CT schools that received patience only and 12 PC schools) when using this sample (Sample 1 henceforth) to evaluate the effect of the grit intervention. Notice, however, that the design of this study does not allow us to evaluate the effect of the grit intervention in isolation. Even though we show that the patience treatment has no effect on grit related outcomes by comparing CT and PC (see Tables A.1, A.2 and A.3 in Online Appendix A), we cannot rule out the effect of dynamic complementarities.[5] The second study, which was implemented in the school year 2015-2016 and essentially provides a replication sample, resolves this issue.

In the second study, we randomly assign the same grit intervention across a new set of state schools in Istanbul. This sample (Sample 2 henceforth) consists of 16 schools (8 treatment, 8 control). While the intervention followed the same procedures (same curricular materials and teacher training approach), there are a couple of important differences in the way the study was conducted. These changes were made to alleviate potential issues with the design of the first study, which were due to logistical constraints. First, in the replication study the treatment schools were not subject to the

[4] After this follow-up we lost 1 CT school.

[5] The estimated treatment effect of the patience intervention on test scores is negative and very imprecisely estimated. We note that our estimates from Sample 2 help us rule out that the estimated effects of the grit intervention on test scores are materially affected by any potential effects of the patience treatment.

patience treatment. This allows us isolate the effect of the grit intervention. Second, we administer objective math and verbal tests, not just at follow-up but also at baseline. These tests measure students' math and verbal (Turkish) performance, two core skills that are of utmost importance for students' further academic endeavors (Altonji *et al.* 2012; Hodara 2013; Aucejo and James 2019).[6]

In both studies, the randomization was performed in the following way. First, the Istanbul Directorate of Education sent the official documentation of the program to all elementary schools in designated districts of Istanbul. The teachers in these schools were then contacted in random sequence and offered to participate in the program. Teachers were informed that upon participation they would be assigned to different training phases within the coming two academic years. All teachers who agreed to participate were promised to eventually receive all training materials and to participate in training seminars, but they were not told when within the next two academic years they would receive the treatment, until the random assignment was completed. The promise of the training offer was made to the teacher and not to current students, i.e. while children in control groups would never receive the training as they move on to middle school after year 4, their teachers would, albeit at a later time.

Once a teacher stated a willingness to participate, we assigned their school into treatment or control. The sample generated with this design contains schools in which at least one teacher stated their willingness to participate in the program. Therefore, the estimated impact of the program is the average treatment effect on the treated and is not readily generalizable to the population. However, in the study Sample 1 approximately 60% of the contacted teachers accepted our offer and the most common reason for non-participation was being "busy with other projects, although happy to participate in this program at a later date" (about 20%). The rest of the non-participation was due to "impending transfer to a school in another city, with a willingness to participate if the program is implemented there" (about 5%), and "not being in a position to participate due to private circumstances" (about 10%). In study Sample 2, acceptance of the training offer reached 80%. Given these numbers, we conjecture that the external validity of our results is strong.

In Sample 1, baseline data were collected in Spring 2013, the first intervention (patience) was implemented in Spring 2013, and the grit intervention was implemented in Fall 2013. In Sample 2, the baseline data were collected in Spring 2015, the intervention (grit only) was implemented in Fall 2015. We note that the school year in Turkey starts in mid-September and finishes in early-June. In both samples, treated teachers spent

[6] Another difference between Sample 1 and Sample 2 is that the students in Sample 2 are about 6 months younger than the students in Sample 1. This is because of an unexpected educational reform implemented in 2012 that lowered the age at which children start school.

about 12 weeks in the beginning of the school year to cover the curriculum we designed. In Sample 1, the incentivized experiments were conducted in May 2014, towards the end of the school year. By that time, students had been exposed to the trained teacher for almost the entire academic year. In Sample 2, the experiments were carried out in January 2016, shortly after the teachers had covered the twelve-week curriculum.

Acknowledging the importance of a long-term follow-up, we launched two separate data collection efforts, one covering Sample 1 in March 2016 and the second covering Sample 2 in June 2017. The first one involved revisiting the students in Sample 1 when they were in grade 6, approximately 2.5 years after the intervention and giving them math and verbal tests based on the official grade 6 curriculum. The second one involved revisiting Sample 2 students when they were in grade 5, approximately 1.5 years after the program, with the same purpose (math and verbal tests based on the grade 5 curriculum). Because there is no central database in Turkey that allows tracking students easily when they change schools, to be able to conduct the follow-up, we enlisted elementary school headmasters' help in getting a list of schools that their students usually go to in the neighborhood.[7] After locating these middle schools, we obtained a list of the students enrolled in 6th grade for Sample 1 and 5th grade in Sample 2. We then matched the lists with our elementary school data based on student and elementary school name, and with this method we were able to track about 55% (60%) of the students in the original Sample 1 (Sample 2). We note that attrition is balanced across treatment and control groups both in Sample 1 (p-value=0.883) as well as in Sample 2 (p-value=0.935). To conduct the tests at follow-up, we visited these middle schools and found the students distributed across different classrooms. We identified the students who were part of our study and assembled them in a separate room in which they took the tests. As we show in Table A.4 in Online Appendix A, we do not find any significant differences in student characteristics in our follow-up data for Sample 1. For Sample 2, while most characteristics are well-balanced, we detect some differences, e.g. in baseline verbal test scores. We use a number of baseline variables as covariates in the regressions to correct for potential imbalances and also use inverse probability weights to account for possible differential attrition. Details of the evaluation designs for each study sample are given in Table I.

[Insert Table I here.]

Note that our control group was also subject to a number of placebo treatments at the time of our study. These treatments were all ministry-approved extra-curricular

[7] Turkey has a two-tier education system where the children of middle and higher socioeconomic strata tend to attend well-resourced private schools. Because our sample covers only state-funded schools in remote areas of Istanbul, it predominantly represents Turkey's low socioeconomic segment. In this segment, most families send their children to the closest state school in their catchment area.

projects (e.g. on environment sensitivity, health and hygiene), similar to the current intervention in terms of teacher involvement and types of activities but unlikely to have affected the outcomes we study. These placebo treatments allow us to rule out various potential mechanisms as we discuss later in Section V.

II.C. Experimental Outcomes: A Real Effort Task

We estimate the effect of the intervention on students' behaviors and outcomes in an incentivized experimental task designed to measure several aspects of grit. Our design requires two different visits to the same classroom, a week apart from each other. In the first visit, children go through five rounds of a mathematical real effort task. In particular, they are presented with a grid which contains different numbers where the goal is to find pairs of numbers that add up to 100. At the end of the five rounds, one of the rounds is selected at random and subjects get rewarded based on their performance in that round. Rewards depend on meeting a performance target. In all the tasks we present to the children, the target is to find three pairs of numbers which sum to 100 within 1.5 minutes.[8] The rewards consist of gifts of value to children of this age group. These include fun stationary items, small puzzles, skipping ropes, frisbees, small balls and keychains. We carefully selected the items to reflect what was currently trendy and sought-after among children of this age.

Before each round starts, subjects have the chance to choose between two different types of tasks for that round: (1) the "4-gift game", which yields four gifts in the case of success and zero in the case of failure, and (2) the "1-gift game", which yields one gift in the case of success and zero in the case of failure. Although in both games the goal is to find at least three pairs of numbers adding to 100, the 4-gift game is more difficult than the 1-gift game. In particular, in the 1-gift game the grid is smaller, and the matching pairs are easier to identify.[9]

Before the five periods start, all subjects are given a large grid that contains many matching numbers and they are given two minutes to find as many pairs of numbers that add to 100 as possible. This is intended to both familiarize the children with the task before they make decisions, and measure task-specific ability. The rewards are such that children get a small gift for each pair they can find. These small gifts (e.g. a regular pencil, single hairpin etc.) are significantly lower in value than the gifts used as rewards in the actual task, and children are aware of this. In addition, information

[8] Note that while Sample 1 students are given 1.5 minutes for each round, Sample 2 students are given 1 minute 45 seconds for each round. We chose to give Sample 2 students more time because they were on average younger than Sample 1 students. See Section III for more details on the characteristics of the two samples.

[9] See Appendix B for examples of the two types of task.

about actual rewards they receive from this task is not revealed until the end of the 1st visit. In the main 5-round part of the experiment, subjects are distributed two booklets of 5 pages each, the 4-gift game booklet and the 1-gift game booklet. Each booklet contains 5 pages that correspond to the 5 rounds of the relevant type of game. In addition, subjects are distributed a choice sheet. Before a typical round starts, subjects are instructed to circle their game of choice for the upcoming round in their choice sheet, and then get ready to open the relevant page of their booklet of choice. They are then given 1.5 minutes to find as many matching number pairs as they can. All students are instructed to fold their arms once the 1.5 minutes are over. During this time, experimenters go around the class and circle either “Succeeded” or “Failed” on the students’ sheets for that round, based on whether at least 3 pairs were correctly found. As mentioned above, students have the opportunity to switch back and forth between the two types of tasks as the rounds progress.

The above procedures, whereby students work on their task of choice in each round have one exception. In the first round, the students’ choices are implemented with 50% chance, and with 50% chance they play the difficult game irrespective of their choice. This allows us to obtain a sample of students playing the difficult task in the first round that is free from selection. From the 2nd round onwards, students are completely free in their choices, and their choices are implemented with 100% chance.

After the five rounds are completed, we inform the children that we will visit their classrooms once more, in exactly a week’s time. The children are told that they will play the game one more time during this second visit, and that they need to decide now whether they would like to play the 4-gift (difficult) game or the 1-gift (easy) game at that point. They are also told that they will have access to an "exercise booklet" which contains examples and practice questions that have a similar difficulty level to the 4-gift game. Just as in the first round, to get a subsample to play the difficult game free of selection, the students’ choices are implemented with 50% chance, and with 50% chance they play the challenging game in the next visit. Students are aware of this procedure when they make their choices. They are also informed about which game they are going to play in the 2nd visit at the end of the 1st visit. Actual rewards from the first visit are not revealed until after all the choices have been made for the second visit. In total, the first visit takes two lecture hours.

In the second visit, children perform the task they chose at the end of the first visit or the difficult task, depending on whether the difficult task was imposed for them or not. They again have 1.5 minutes to find pairs of numbers that add up to 100. The game is played for one round, and rewards are based on performance during that round. The reward basket in the second visit contains the same array of items that were used as rewards in the first visit. Full instructions are given in Online Appendix B.

To minimize potential Pygmalion/experimenter demand effects, we made sure that teachers were not present in the classroom during the data collection. Students were made aware that no information on their choices/outcomes would be shared with their teachers. We did not inform students that the data collection was in any way related to the educational material they had been exposed to, and deliberately avoided wording/terminology that was frequently used in the intervention (e.g. grit, quitting, challenge) during the data collection. The experiments were labelled as 'games' in which the students could earn rewards. It was repeatedly emphasized that there was no right or wrong decision in these games, that everyone was different and each student was free to do as he/she pleased. Finally, the use of strong incentives, as advocated in the experimental economics literature, helps minimize potential Pygmalion effects (Hertwig and Ortmann 2001). The rewards children could earn in the experimental tasks were of significant value to them. Overall, we were very careful to take precautions at the design stage to minimize potential Pygmalion effects and we believe it is very unlikely that teachers' or experimenters' expectations could have altered students' behavior in the experimental tasks. Similarly, to prevent potential demand or Hawthorne effects operating on test scores, teachers were given no information about the study design, and neither the teachers nor the students knew that we would be conducting standardized tests at any point in time. We therefore do not expect teachers to have changed their teaching to prepare their students for the tests. Our longer-term measurements, which were conducted after children moved on to middle school, provide further reassurance, since children are not in the same environment anymore and are taught by different teachers for each subject.

III Data and Baseline Information

The treatment was randomized across 36 schools in Sample 1 (15 treatment, 21 control) and 16 schools in Sample 2 (8 treatment and 8 control). The number of students who were officially registered in the classrooms that were part of the trial at the beginning of the school year was 2,575 in Sample 1 (in 68 classrooms) and 1,499 in Sample 2 (in 42 classrooms). The average number of students officially registered in each classroom in the beginning of the school year is 38 in Sample 1 and 36 in Sample 2. In the classrooms in which the data collection was conducted, 79% of the students (1899) in Sample 1 were present on the day of testing and consented to participate, while 91%

(1360 students) were present and consented in Sample 2.[10] We estimate the treatment effects separately for each study sample.

For both samples, the baseline data contain rich information on student characteristics. In addition to collecting information on demographic variables such as gender and age, we administer a Raven's progressive matrices test to obtain a measure of cognitive ability (Raven *et al.*, 2004). Moreover, we measure students' risk tolerance using a version of the Gneezy and Potters (1997) risky allocation task. We also conducted surveys before and after the intervention to gather information on students' (i) baseline beliefs about the malleability of ability, and (ii) attitudes and behaviors related to grit and perseverance. The questions measuring grit are based on the Duckworth and Quinn (2009) grit scale and elicit self-reported gritty behaviors, while questions that elicit beliefs about the malleability of abilities (mindset) are based on Dweck (2006); see Online Appendix D for the full set of questions. To obtain the aggregate measures we are interested in, we extract the first principal component from the students' responses to these questionnaire items, and normalize the variables to have mean zero and standard deviation one. Finally, we also have information on the students' academic success and their families' socioeconomic status (SES), obtained through teacher surveys. For these, teachers are asked to rate the wealth level of the students' family on a 5-point scale (1: very low, 5: very high). The success variable asks teachers to rate the students' overall academic performance on a 5-point scale (1: very low, 5: very high). Both samples contain measures of prior academic achievement. These are grades (for Sample 1 and 2) and standardized test scores (for Sample 2) in two core subjects, mathematics and Turkish. For the purpose of the analysis, we normalize them to have a mean of zero and a standard deviation of one.[11]

We use these baseline measures to assess the samples' balance across treatment status. Table II provides the balance tests for Sample 1 and Sample 2. In Sample 1, we do not observe any statistically significant differences in student characteristics, test scores or beliefs. In Sample 2, most characteristics, test scores and beliefs are also balanced, although there are some significant differences across treatment and control. We use a number of baseline variables as covariates in the estimation of the average treatment effects to increase the precision of our estimates and to account for potential imbalances in baseline covariates which are predictive of our outcome measures.

Next, we investigate whether students with different treatment status have different

[10] We collected experimental outcomes in all but four classrooms in Sample 1, which we could not visit due to scheduling constraints towards the end of the school year. We visited all classrooms in Sample 2. Differences in absenteeism across the two samples reflect the fact that Sample 1 classrooms were visited in May (almost at the end of the school year) and Sample 2 in January.

[11] We do not have baseline standardized test scores for Sample 1. In fact, one motivation for replicating the intervention was to obtain an objective measure of achievement at baseline.

task-specific ability at the beginning of the incentivized experiment. As explained in Section II, at the beginning of the first visit, there is an initial round where students are asked to find as many pairs as possible in a large grid of numbers. This round allows us to measure the students' task-specific skill level. As can be seen in the last row of Table II, the number of pairs found in this task (referred to as "task ability") is not different across treatment status in either sample.

[Insert Table II here.]

Finally, we note that students' choices in the experimental task correlate with baseline test scores. Specifically, choosing the difficult task in all five rounds and choosing the difficult task for the second visit are positively correlated with math and verbal scores at baseline (see Online Appendix Table A.5).

IV Results

IV.A. Estimation of Treatment Effects

To test the null hypothesis that the program had no impact on the experimental outcome y^E, we estimate the average treatment effect conditioning on baseline covariates:

$$y_{ij}^E = \alpha_0 + \alpha_1 T_j + X_{ij}^{'} \gamma + \varepsilon_{ij}$$

where T_j is a dummy variable which equals 1 if school j is in the treatment group and zero otherwise, and X_{ij} is a vector of observables for student i in school j that are potentially predictive of the outcome measures we study. The estimated $\hat{\alpha}_1$ is the average treatment effect on the treated. When estimating the treatment effect on experimental choices and outcomes, we control for task ability, gender, the Raven score, baseline beliefs and test scores, and risk tolerance as well as a dummy variable for whether the student has any inconsistent data entries.

Estimates are obtained via a logit regression when the outcome considered is binary. This is the case for students' choices between the difficult and the easy task, and for their success/failure in meeting the performance target. The binary outcome variable "success" is defined as finding three correct pairs or more. In the case of payoffs, the above equation is estimated via ordinary least squares. The outcome variable "payoff" takes the value 0 if the target of finding three pairs is not met, 1 if the easy game is played and the target is met, and 4 if the difficult game is played and the target is met. To test the null hypothesis that the program had no impact on test scores y^T, we estimate the average treatment effect using the same specification and control for gender, the Raven score, class size, baseline beliefs and test scores in the estimation.[12]

In all empirical analyses, standard errors are clustered at the school level, which is the unit of randomization. To account for the small number of clusters we also run permutation tests and provide exact p-values. As highlighted by Young (forthcoming), using permutation inference is important in the context of clustered RCT designs. Given that we randomized treatment at the school level, regression model errors will not be independent within clusters because the outcome variables have non-zero intra-cluster correlation while the treatment assignment is mechanically correlated within clusters (Cameron and Miller, 2015). We use a Fisherian randomization inference, which constitutes a test of a sharp null (no effect, rather than no average treatment effect). The procedure is straightforward to implement. Since we have perfect information on the exact randomization procedure of our study (school-level clustered randomization design), we re-randomize the treatment assignment 1,000 times and calculate the Fisher exact p-values. We use the coefficient estimate as the randomization statistic. The corresponding p-values are presented in an extra row at the bottom of the results tables.

IV.B. Treatment Effect on Choices and Outcomes in the Real Effort Task

In the following, we examine the effect of treatment on students' choices and outcomes in the incentivized real effort task. Section IV.B.1. presents the results for the first visit while Section IV.B.2. presents the results of the second visit. For the sake of brevity, all tables in this section present the estimated treatment effects without presenting the coefficient estimates of the covariates.

IV.B.1. First Visit

In the first visit, students are asked to choose between the 1-gift (easy) game and the 4-gift (difficult) game in each of the five main rounds of the experiment. With the exception of the first round, in which some students are randomly selected to do the difficult task irrespective of their choice, students perform the task of their choice before moving on to the next round. While in both samples the vast majority of students is successful on the easy task, this is not the case for the difficult task. Given that we randomly selected a subset of students to do the difficult task in the first round irrespective of their choice, this allows us to obtain an estimate of the empirical success rate on the difficult task free from selection. In Sample 1, 29% of the students for whom the difficult task is imposed are successful on the difficult task, while for Sample 2 the corresponding

[12] Note that all experimental results are robust to excluding all individuals from the estimation for whom we have inconsistent data entries (9%), e.g. doing the easy task when difficult is imposed, or actually playing a different game than they planned for (see Tables A.6 and A.7 in Online Appendix A).

number is 20%. Given that the difficult task yields 4 gifts in the case of success while the easy task only yields 1 gift, the expected payoff from the two tasks was about equal.

Table III presents the estimated treatment effect on students' choice of task difficulty during the five rounds of the first visit (columns 1-5).[13] The presented estimates are average marginal effects from logit regressions in which we regress the choice of task difficulty on a treatment dummy and a set of covariates. The first finding to note is that in both samples, the proportion of students in the control group who attempt the difficult task declines visibly through the rounds. While in both samples about 67% of the control group students attempt the difficult task in the first round, only 40% (26%) attempt the difficult task in round five in Sample 1 (Sample 2). While a similar trend can be observed for treated students, we note that in all five rounds treated students are significantly more likely to attempt the difficult task compared to control group students. In Sample 1, students are 10 percentage points more likely to choose the difficult task in the first round and this effect persists until the fifth round in which students are 9 percentage points more likely to choose the difficult task. Similarly, students in Sample 2 are also 10 percentage points more likely to choose the difficult task in round one and the effect also persists until round five, in which they are 13 percentage points more likely to attempt the difficult task. In fact, treated students are about 9 and 12 percentage points more likely to choose the difficult task in all of the five rounds in Sample 1 and Sample 2, respectively (see column 6).[14]

Next, we estimate the effect of treatment on task choice in round two for those students who failed at the imposed difficult task in the first round. Given that we randomly chose a subset of students to perform the difficult task in the first round, we can analyze how treatment affects task choice after failure in a sample that is free from selection. In Table III column 7, we show that treated students who failed at the imposed difficult task in round one are significantly more likely to want to re-attempt the difficult task in round two, despite the fact that there are no significant differences in task ability (in visit 1) across treatment and control group students who failed at the imposed difficult task (Sample 1 p-value=0.47; Sample 2 p-value=0.24). Success in the imposed difficult task, while not exogenous, is balanced across treatment and control group students (Sample 1 p-value=0.59, Sample 2 p-value=0.15) in the first visit. The estimated difference between treatment and control group students is striking. Among the students who failed at the imposed difficult task in Sample 1, treated students are 15 percentage points more likely to re-attempt the difficult task in the subsequent round (permutation

[13] In the first round, when the difficult task is not imposed, we take the task that students actually played as their choice.

[14] Regarding the small fluctuations from round to round, we note that these arise because some students did not complete all rounds, e.g. because they had to go to the bathroom.

p-value=0.05). The corresponding estimate in Sample 2 is also 15 percentage points (permutation p-value=0.11). Note that if we perform this estimation with all students for whom the difficult task was imposed we obtain similar results. When we restrict the sample to those students who were randomly selected to do the difficult task in the first round irrespective of their choice, we find that treated students are 11 and 16 percentage points more likely to choose the difficult task for the subsequent round in Sample 1 and Sample 2, respectively. While we cannot rule out that there are unobserved differences between treatment and control group students who failed at the imposed difficult task, these results strongly suggest that the intervention affects the way in which students react to negative feedback.

As explained in Section II.C., at the end of the first visit we let the students know that we will come back exactly one week later and that they will play the same game for one additional round. We also inform them that if they like, they can take a study booklet covering numerous examples of the difficult game and study/practice over the week using this booklet. We emphasize that this is entirely voluntary. We then collect their decisions on which type of task they would like to do in the following week. After we collect these choices, students are informed whether they will have to play the difficult game in the following week, or the game of their choice. The purpose of this exercise is to see whether the treatment generates "goal-setting" behavior in the form of a commitment to improve task related ability in the six days before the second visit. We predict that students who believe that ability in this task is malleable through sustained effort and perseverance are more likely to set the goal of succeeding in the difficult game and therefore more likely to commit to playing the difficult game. This is exactly what we see in the last column of Table III. Treated students are estimated to be 14 percentage points more likely to plan to play the difficult game in the following week in Sample 1 (permutation p-value=0.000), and 18 percentage points more likely to plan to play the difficult game in Sample 2 (permutation p-value=0.004).

[Insert Table III here.]

We now turn to the question of whether treatment affects students' experimental outcomes, namely, success and payoffs. Table IV column 1 presents the estimated treatment effects on success in round 1 of the first visit for the sample which was forced to play the difficult game. This particular round is designed in such a way that allows us to estimate the treatment effect on success in the difficult game free of selection. As mentioned above, we find no significant treatment effect on success rates in either sample (permutation p-values are 0.59 and 0.15 for Sample 1 and 2 respectively). This is also generally true for payoffs in all rounds: the estimated treatment effects on payoffs in all five rounds are not statistically different from zero, with the exception of the first round in Sample 2, which is positively significant at the 5% level.

Treated students set the goal of succeeding on the difficult task in the second week - but did they actually achieve this goal? This is the question we explore in the next subsection.

[Insert Table IV here.]

IV.B.2. Second Visit

The temporal component of our experimental task serves a very important purpose for our study. While it is unlikely that students can improve their ability on a task within five rounds of only 1.5 minutes, it may well be that students can accumulate task-specific ability when given sufficient time. Ability accumulation takes time and effort, and the amount of time and effort required to master a task varies according to the characteristics of the task. In this specific real effort task, we chose to give students one week, with the conjecture that it would be sufficient for motivated students to work through the exercises provided in the study booklet and that such effort would lead to a higher probability of success in the second visit.

As in the first round of the first visit, a random subset of students were asked to do the difficult task during the second visit, irrespective of their choice. This allows us to investigate whether the treatment affects the probability of success in the difficult task in the second visit. Table V presents the estimated treatment effects on outcomes of the second visit. The first column presents the treatment effects on success obtained from the sample on which the difficult task is imposed, while columns 2-5 present the treatment effects on payoffs. For the latter, we estimate treatment effects on the entire sample as well as conditional on whether the difficult task was imposed in the class or not. Looking at the first column for both samples, we see that treated students are about 8 (10) percentage points more likely to succeed in the difficult game in Sample 1 (Sample 2). These effects are statistically significant. The increased success rate is also reflected in payoffs: we estimate a statistically significant 16% and 26% treatment effect on payoffs in Sample 1 and Sample 2, respectively (0.30 and 0.45 more gifts for treated students in Sample 1 and Sample 2, respectively). Note also that the estimated effects are similar for the imposed and unimposed samples. Considering the combined payoffs of both visits (the last column), we estimate 12% higher payoffs in Sample 1, and 21% higher payoffs in Sample 2 relative to their respective control groups.

A natural question is whether there is a type of student for whom the treatment was particularly successful. Presumably, treatment may have a differential impact on students with different task-related ability levels. For example, the treatment might be effective in pushing a potentially able but reluctant student into planning to do the difficult task and in encouraging her to study. It may encourage a student with low ability to study hard as well. Since the performance technology is conducive to ability accumulation, we might also observe increased success rates in the second week for these students. Our analyses, however, do not reveal any systematic heterogeneity in treatment effects with respect to gender, task ability and cognitive ability.[15]

[Insert Table V here.]

IV.C. Are Choices Payoff-Maximizing?

An important question as regards an intervention of this sort is whether being gritty is good for everyone, i.e. whether it is optimal for children to always set challenging goals, persevere in the case of setbacks and engage in costly skill accumulation activities. Certain endeavors might not be worth the time and effort if they are unachievable or if the costs of perseverance required for success are so high that they outweigh the potential gains. In general, perseverance is more likely to pay off when the performance technology is conducive to skill accumulation and the costs of effort or investment are not too high.

To get some insight into this question, we investigate to what extent individual choices of task difficulty are payoff-maximizing in expectation. More specifically, we first obtain an individual measure of each student's probability of success in each task given the student's baseline characteristics, using the empirical distribution of success. We then calculate the student's expected payoff from choosing the difficult task and compare that with the expected payoff from choosing the easy task. Once we have an estimate of which task choice would be payoff-maximizing for each student, we compare this payoff-maximizing choice to the student's actual task choice.

In Sample 1, treated students are no more likely to choose the payoff-maximizing task in the first round of the first visit (Table V, column 6) but they are 8 percentage points more likely to choose the payoff-maximizing task for the second visit (Table V, column 7). In Sample 2, students are more likely to choose the payoff-maximizing task in both visits. In particular, treated students are 6 percentage points more likely to make the payoff-maximizing choice in visit 1, and 8 percentage points more likely to make the payoff-maximizing choice in visit 2. Overall, we conclude that treated students were more likely to make decisions that were payoff-maximizing in expectation. Note that it is difficult to make statements about utility (rather than payoffs) in this context, since effort costs are unobservable. However, the choices and outcomes of treated students in the 2nd visit suggests, through revealed preference, that these choices might also be utility-maximizing for this group.

[15] Full results on heterogeneity are available upon request.

Overall, the estimated effects using our behavioral measure are strong and also robust to linear probability estimation and estimation without baseline covariates; see Tables A.8 and A.9 for the former, and Tables A.10 and A.11 for the latter in Online Appendix A.

IV.D. Treatment Effect on Test Scores

The implication of a change in beliefs regarding the malleability of skills can be far-reaching. For one thing, a student who used to think that there is not much one can do to excel in an area, whether that be related to art or science, may now be convinced that all it takes is goal-setting and hard work. If this is the case, we may be able to see improvements in other domains where sustained effort results in better outcomes. The obvious outcome to look at in this regard is school grades. For this purpose, we collect official grades (given by the teacher) that reflect the students' math and verbal performance at the end of the school year. Because of the possibility that teachers' assessments may have been affected by the treatment in an unknown way, we decided to also administer standardized tests (math and verbal) in both samples.

We find no significant impact of treatment on average teacher-given grades in either sample (Table VI). Anecdotal evidence from conversations with out-of-sample teachers suggests that the reason why teacher-given grades may be unaffected by the intervention is that teachers in elementary school tend to apply a relative grading scheme with a stable distribution. On the contrary, we find remarkably large and significant treatment effects on standardized test scores (Table VII). In the first (short-term) follow-up in Sample 2, which was conducted in January 2016, we detect a significant treatment effect of 0.31 (permutation p-value=0.008) on standardized math scores and 0.13 (permutation p-value=0.105) on standardized verbal scores. In the second follow-up in Sample 2, which we administered approximately 1.5 years after the intervention, we still find a positive and significant treatment effect of 0.19 standard deviations for math (permutation p-value=0.026) and a positive albeit insignificant effect for verbal scores. Similarly, for Sample 1 where we have data from a 2.5 year follow-up, we find that the treatment has a persistent effect on standardized math performance. In particular, the treatment raises student achievement in the standardized math test by 0.23 standard deviations (permutation p-value=0.044). Again, we find a positive albeit insignificant result for performance on the verbal test for this sample, suggesting that the results for Turkish performance are fading over time. Table A.12 in Online Appendix A provides the estimated treatment effects on test scores in which we only use baseline test scores as controls. We note that we lose precision when we exclude the rich set of control variables in those regressions. Specifically, the long-run effects on math test scores for

Sample 2 are less precisely estimated and no longer significant at conventional levels.

[Insert Table VI here.]

[Insert Table VII here.]

Compared to other estimates in the literature, our short-term effect on math scores is large. To put these effect sizes in perspective, we note that Schanzenbach (2006), in a review of the existing evidence on the project STAR, concludes that being randomly assigned to a small class raises student test scores by 0.15 standard deviations. Note, however, that while we estimate a large effect immediately after the program implementation, the estimated effects after 2.5 years following the implementation are smaller, and more in line with the literature. Note also that we deliberately target low socioeconomic status students for whom interventions of this type have been shown to be most effective (see Sisk *et al.*, 2018).

The differential effect of the treatment on math and verbal scores is also consistent with the literature. A recent review article by Fryer (2017) which summarizes the lessons learned from close to 200 randomized field experiments in education notes that educational interventions in general tend to be more effective at increasing math achievement relative to reading achievement (e.g., Hoxby and Murarka, 2009, Abdulkadiroglu *et al.*, 2011, Dobbie and Fryer, 2011, Angrist *et al.*, 2012, Fryer, 2014). As noted in the review article, there are different theories that may explain the disparity in treatment effects by subject area. Firstly, it may be that reading scores are influenced to a great extent by the language spoken outside the classroom, which is why they may be harder to influence through targeted interventions in the school environment (Rickford, 1999, Charity *et al.*, 2004). Secondly, research in developmental psychology has suggested that the critical period for language development occurs early in life, while the critical period for developing higher cognitive functions extends to adolescence (e.g., Hopkins and Bracht, 1975, Newport, 1990, Knudson *et al.*, 2006).

Finally, our results also relate to the literature on the importance of teacher quality for student achievement (Rivkin *et al.*, 2005, Hanushek, 2011, Hanushek and Rivkin, 2012). Previous studies have shown that teachers affect later-life outcomes of students both through influencing their test scores as well as their non-cognitive skills (see, e.g., Chetty *et al.*, 2011, Chetty *et al.*, 2014, Jackson, 2018). Consistently with this, educational policymakers in many countries provide professional development programs for teachers (Popova *et al.* 2016). We relate to this literature by showing that a program based on training teachers has the potential to raise both students' test scores as well as students' non-cognitive skills as measured through a behavioral task. To the extent that changes in non-cognitive skills are persistent and lead to better life outcomes, it is plausible to expect that the impact of the program on students' non-cognitive skills may spill over to other important life outcomes in the long-run.

IV.E. Multiple Hypotheses and Replication

We estimate the effect of the treatment on multiple outcomes (several experimental as well as achievement outcomes). This may raise the issue of multiple hypotheses testing. Tables A.13 and A.14 in Online Appendix A provide Romano-Wolf p-values along with the original ones. For the purpose of this analysis, we group our main outcome measures into two blocks, namely (i) achievement outcomes and (ii) survey and experimental outcomes, and perform the analysis separately for each block. The results confirm that the precision of all our estimated treatment effects survives this adjustment, i.e., none of the estimated effects switch from being statistically significant to insignificant. This test is conservative in our specific context, since it does not account for the fact that we replicated our study using an independent sample of schools. As we report above, the intervention has yielded both qualitatively and quantitatively similar results in the replication sample. Figure I shows this in visual clarity. While all significant treatment effects in Sample 1 appear as significant in Sample 2, all insignificant findings replicate in the same manner. While the rate of false positives depends both on the observed significance level and the statistical power of an experiment, which we report in Online Appendix Tables A.15 and A.16, an independent replication like the one we have dramatically increases the chances that the original finding is true (see Maniadis *et al.*, 2014). This is especially important in our setting in which attrition rates lower the power of our design.

[Insert Figure I here.]

V Discussion

While our research design does not allow us to disentangle all possible channels through which the intervention may have affected outcomes, we can provide some suggestive evidence on which channels may potentially be important and which channels are unlikely to have played a role. One potential channel may be beliefs regarding the malleability of ability through effort. It may be that the intervention shifted the beliefs about the productivity of effort toward more optimism, resulting in more perseverant behavior and higher resilience against setbacks. Consistently with this mechanism, we estimate a significant treatment effect on students' self-reported beliefs about the malleability of skills as well as their self-reported levels of grit. The estimated treatment effect on students' beliefs about the malleability of skills is 0.35 standard deviations in Sample 1 and 0.33 standard deviations in Sample 2, while the estimated effect on students' self-reported grit is 0.29 standard deviations in Sample 1 and 0.35 standard deviations in Sample 2 (see Table A.17 in Online Appendix A). Figures II and III present the visible location shift in these survey-based measures. These results provide evidence, albeit suggestive, that the program may have generated the estimated effects by influencing students' beliefs about the malleability of skills/the productivity of effort.

[Insert Figure II here.]

[Insert Figure III here.]

In addition to beliefs about the malleability of ability, other beliefs and behaviors may have been affected by the treatment, and therefore could have played a role in mediating the effects. Beliefs about students' own ability, i.e., their self-confidence is one alternative belief channel that could, for example, lead to ambitious goal setting. We should note, however, that our intervention does not aim to increase self-confidence about ability but rather students' optimism about the future performance they can achieve through exerting effort. A child who is not particularly confident about her ability (for example after having experienced a failure) may still be optimistic about her future performance, if she thinks that she can improve her performance by exerting effort, as emphasized by the intervention. Nevertheless, we did consider this channel at the design stage and collected both baseline and follow-up information on students' self-assessment of their own math and verbal ability, as well as how smart they believe they are relative to others. We then constructed an aggregate measure of self-confidence by extracting a factor from these survey questions. Using this measure, we find that the treatment had no effect on students' self-confidence in their own ability (p-value=0.81). In terms of other attitudes and behaviors, we consider students' attitude towards risk and patience, which may have been affected by the treatment and may have mediated our estimated treatment effects. Risk tolerant individuals may be more likely to undertake challenging tasks, and patient individuals may be more willing to work towards goals whose payoffs will come later, as is usually the case in education as well as in our behavioral task. We do not estimate statistically significant treatment effects on either risk tolerance (p-value=0.52) or patience (p-value=0.97).[16]

Recall that the program was implemented by teachers within the allotted extracurricular hours. An alternative channel may be that the implementation of the program leads to more intensive student-teacher interaction, which in turn results in higher test scores. However, we did rule out this potential channel at the design stage by making sure that our control teachers were also engaged in Ministry-approved extracurricular

[16] The latter result comes from Sample 2, where we can estimate the effect of the pure grit treatment on patience measured by a convex time budget task adapted from Andreoni and Sprenger (2012) and used in Alan and Ertac (2018). In this task, children are asked to make an intertemporal consumption allocation in which waiting pays off, and patience is measured by the amount allocated to the earlier date.

projects. These involved similar levels of classroom activity and student-teacher interaction. Besides the program on patience, whose effects we can rule out, these 'placebo' projects were related to the environment, dental care and hygiene, which are unlikely to affect the outcomes we are interested in.

We re-emphasize that the evidence we document in this section is only suggestive and by no means gives an exhaustive account of all possible channels. In fact, there are a couple of alternative channels we cannot rule out with our design. One is the role of peer effects. Peer effects have been studied recently in the laboratory in the context of perseverance (Gerhards and Gravert, 2016, Buechel *et al.*, 2018). In our context, students in treated classrooms may change their beliefs and behaviors in response to changes in their classmates' beliefs and behaviors, amplifying the effects of the intervention. The intervention may also create a classroom culture where gritty behavior becomes a norm, which may further strengthen the effects. Similarly, our intervention may also be effective in producing long-lasting effects because of autoproductive dynamics (see, e.g., Yeager and Walton, 2011 for a discussion). Attributing realized success to high effort might create a self-fulfilling cycle of more effort and more success. Improving grit may therefore impact learning in persistent ways. These dynamics of course may interact with peer effects in unknown ways. We leave exploring these interesting channels to future research.

VI Conclusion

Using two independent study samples, we evaluate a large-scale randomized educational intervention that aims to enhance grit in the classroom environment. We estimate the effect of treatment on students' (i) behaviors and outcomes in an incentivized behavioral task and (ii) grades and performance in standardized tests after the implementation of the intervention. We find significant treatment effects of the intervention on students' behaviors and outcomes in the task, which are remarkably similar across the two independent samples. In both samples, treated students are significantly more likely to set challenging goals, to engage in skill accumulation activities, and to accumulate more skill and obtain higher payoffs as a result. Moreover, the intervention also has a large positive impact on students' objective math performance. This effect persists after 2.5 years after the implementation of the program. The effects we report may persist further into adolescence and adulthood, especially since realizations of success attributed to high effort might create a productive cycle of further effort and further success.

From the policy perspective, the paper contributes to the ongoing debate about the malleability of non-cognitive skills and the role of educational programs in enhancing individual achievement through interventions specifically targeting those skills (Almlund *et al.*, 2011; Kautz *et al.*, 2014). Our results provide an affirmative answer to the question of malleability within the context of an important non-cognitive skill, and highlight a particular low-cost alternative that can be implemented to foster it in the natural environment of the classroom. Being able to achieve such an impact in the school environment offers hope for reducing persistent achievement gaps observed in many countries, where many educational policy actions aiming to improve family inputs face challenges in engaging families of low socioeconomic strata.

University of Essex, Bilkent University and J-PAL
University of Oxford
Koç University

Supplementary Material

An Online Appendix to this article as well as the data and code replicating the tables and figures in this article can be found at *The Quarterly Journal of Economics* online.

References

Abdulkadiroglu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J. and Pathak, P. A. (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots. *Quarterly Journal of Economics*, **126** (2), 699–748.

Alan, S. and Ertac, S. (2017). Belief in Hard Work and Altruism: Evidence from a Randomized Experiment. Working Paper.

— and — (2018). Fostering Patience in the Classroom: Results from a Randomized Educational Intervention. *Journal of Political Economy*, **126** (5), 1865–1911.

— and — (forthcoming). Mitigating the Gender Gap in the Willingness to Compete: Evidence from a Randomized Field Experiment. *Journal of the European Economic Association*, **https://doi.org/10.1093/jeea/jvy036**.

Almlund, M., Duckworth, A. L., Heckman, J. J. and Kautz, T. D. (2011). Personality Psychology and Economics. *In E. Hanushek, S. Machin, and L. Woessman, eds., Handbook of the Economics of Education*, pp. 1–181.

Altonji, J. G., Blom, E. and Meghir, C. (2012). Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers. *Annual Review of Economics*, **4** (1), 185–223.

Andreoni, J. and Sprenger, C. (2012). Estimating Time Preferences from Convex Budgets. *American Economic Review*, **102** (7), 3333–3356.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A. and Walters, C. R. (2012). Who Benefits from KIPP? *Journal of Policy Analysis and Management*, **31** (4), 837–860.

Aronson, J., Fried, C. B. and Good, C. (2002). Reducing the Effects of Stereotype Threat on African American College Students by Shaping Theories of Intelligence. *Journal of Experimental Social Psychology*, **38**, 113–125.

Aucejo, E. M. and James, J. (2019). The Path to College Education: The Role of Math and Verbal Skills. Working Paper.

Bettinger, E., Ludvigsen, S., Rege, M., Solli, I. F. and Yeager, D. (2018). Increasing Perseverance in Math: Evidence from a Field Experiment in Norway. *Journal of Economic Behavior and Organization*, **146**, 1–15.

Bettinger, E. P. and Baker, R. B. (2014). The Effects of Student Coaching: An Evaluation of a Randomized Experiment in Student Advising. *Educational Evaluation and Policy Analysis*, **36** (1), 3–19.

—, Long, B. T., Oreopoulos, P. and Sanbonmatsu, L. (2012). The Role of Simplification and Information: Evidence from the FAFSA Experiment. *Quarterly Journal of Economics*, **127** (3), 1205–1242.

Blackwell, L. S., Trzesniewski, K. H. and Dweck, C. S. (2007). Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention. *Child Development*, **78** (1), 246–263.

Borghans, L., Duckworth, A. L., Heckman, J. J. and ter Weel, B. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, **43** (4), 972–1059.

Buechel, B., Mechtenberg, L. and Petersen, J. (2018). If I Can Do It, So Can You! Peer Effects on Perseverance. *Journal of Economic Behavior and Organization*, **155**, 301–314.

Cameron, A. C. and Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, **50** (2), 317–372.

Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E. and Pan, Y. (2014). Early Childhood Investments Substantially Boost Adult Health. *Science*, **343** (6178), 1478–1485.

Castillo, M., Ferraro, P. J., Jordan, J. L. and Petrie, R. (2011). The Today and Tomorrow of Kids: Time Preferences and Educational Outcomes of Children. *Journal of Public Economics*, **95** (11), 1377–1385.

Charity, A. H., Scarborough, H. S. and Griffin, D. M. (2004). Familiarity with School English in African American Children and Its Relation to Early Reading Achievement. *Child Development*, **75** (5), 1340–1356.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Whitmore Schanzenbach, D. and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, **125** (4), 1593–1660.

—, — and Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, **104** (9), 2633–2679.

Dee, T. S. and West, M. R. (2011). The Non-Cognitive Returns to Class Size. *Educational Evaluation and Policy Analysis*, **33** (1), 23–46.

Dobbie, W. and Fryer, R. (2011). Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, **3** (3), 158–187.

Dobronyi, C. R., Oreopoulos, P. and Petronijevic, U. (2017). Goal Setting, Academic Reminders, and College Success: A Large-scale Field Experiment. NBER Working Paper 23738.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J. and Wagner, G. G. (2011). Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences. *Journal of the European Economic Association*, **9** (3), 522–550.

Duckworth, A. L., Peterson, C., Matthews, M. D. and Kelly, D. R. (2007). Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology*, **92** (6), 1087–1101.

— and Quinn, P. D. (2009). Development and Validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*, **91** (2), 166–174.

Dweck, C. (2006). *Mindset: The New Psychology of Success*. New York, NY: Random House.

Eskreis-Winkler, L., Shulman, E. P. and Duckworth, A. L. (2014). Survivor Mission: Do Those Who Survive Have a Drive to Thrive at Work? *Journal of Positive Psychology*, **9** (3), 209–218.

Fryer, R. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quarterly Journal of Economics*, **126** (4), 1755–1798.

— (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence From Field Experiments. *Quarterly Journal of Economics*, **129** (3), 1355–1407.

— (2017). The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. *In A. V. Banerjee and E. Duflo, eds., Handbook of Field Experiments*, **2**, 95–322.

Garcia, J. L., Heckman, J. J., Leaf, D. E. and Prados, M. J. (2016). The Life-cycle Benefits of an Influential Early Childhood Program. NBER Working Paper 22993.

Gerhards, L. and Gravert, C. (2016). Because Of You I Did Not Give Up - How Peers Affect Perseverance. Working Papers in Economics 659, University of Gothenburg.

GNEEZY, U. and POTTERS, J. (1997). An Experiment on Risk Taking and Evaluation Periods. *Quarterly Journal of Economics*, **112** (2), 631–645.

GOLSTEYN, B. H., GRÖNQVIST, H. and LINDAHL, L. (2013). Adolescent Time Preferences Predict Lifetime Outcomes. *Economic Journal*, **124** (580), 739–761.

GOOD, C., ARONSON, J. and INZLICHT, M. (2003). Improving Adolescents' Standardized Test Performance: An Intervention to Reduce the Effects of Stereotype Threat. *Journal of Applied Developmental Psychology*, **24** (6), 645–662.

HANUSHEK, E. A. (2011). The Economic Value of Higher Teacher Quality. *Economics of Education Review*, **30**, 466–479.

— and RIVKIN, S. G. (2012). The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics*, **4** (1), 31–57.

HECKMAN, J. J., HUMPHRIES, J. E. and MADER, N. S. (2011). The GED. *In E. Hanushek, S. Machin and L. Woessmann, eds., Handbook of the Economics of Education*, **3**, 423–484.

—, MOON, S. H., PINTO, R., SAVELYEV, P. A. and YAVITZ, A. (2010). The Rate of Return to the HighScope Perry Preschool Program. *Journal of Public Economics*, **94** (1), 114–128.

—, PINTO, R. and SAVELYEV, P. (2013). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*, **103** (6), 2052–2086.

—, STIXRUD, J. and URZUA, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, **24** (3), 411–482.

HERTWIG, R. and ORTMANN, A. (2001). Experimental Practices in Economics: A Methodological Challenge for Psychologists. *Behavioral and Brain Sciences*, **24** (3), 383–451.

HODARA, M. (2013). Improving Students' College Math Readiness: A Review of the Evidence on Postsecondary Interventions and Reforms. CAPSEE Working Paper.

HOPKINS, K. D. and BRACHT, G. (1975). Ten-Year Stability of Verbal and Nonverbal IQ Scores. *American Education Research Journal*, **12** (4), 469–477.

HOXBY, C. M. and MURARKA, S. (2009). Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement. NBER Working Paper 14852.

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.

Kautz, T., Heckman, J. J., Diris, R., ter Weel, B. and Borghans, L. (2014). Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success. Paris, France: Organization for Economic Co-operation and Development.

Knudson, E. I., Heckman, J. J., Cameron, J. L. and Shonkoff, J. P. (2006). Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce. *Proceedings of the National Academy of Sciences*, **103** (27), 10155–10162.

Kosse, F., Deckers, T., Falk, A., Pinger, P. and Schildberg-Hörisch, H. (forthcoming). The Formation of Prosociality: Causal Evidence on the Role of the Social Environment. *Journal of Political Economy*.

Levitt, S. D., List, J. A., Neckermann, S. and Sadoff, S. (2016). The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, **8** (4), 183–219.

Maddie, S. R., Matthews, M. D., Kelly, D. R., Villarreal, B. and White, M. (2012). The Role of Hardiness and Grit in Predicting Performance and Retention of USMA Cadets. *Military Psychology*, **24** (1), 19–28.

Maniadis, Z., Tufano, F. and List, J. A. (2014). One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *American Economic Review*, **104**, 277–290.

Moffit, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H. L., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M. and Caspi, A. (2011). A Gradient of Childhood Self-control Predicts Health, Wealth, and Public Safety. *Proceedings of the National Academy of Sciences*, **108** (7), 2693–2698.

Newport, E. L. (1990). Maturational Constraints on Language Learning. *Cognitive Science*, **14** (11), 11–28.

Oreopoulos, P., Patterson, R. W., Petronijevic, U. and Pope, N. G. (2018). Lack of Study Time is the Problem, but What is the Solution? Unsuccessful Attempts to Help Traditional and Online College Students. NBER Working Paper 25036.

— and Petronijevic, U. (2018). Student Coaching: How Far Can Technology Go? *Journal of Human Resources*, **53** (2), 299–329.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S. and Dweck, C. S. (2015). Mind-set Interventions Are a Scalable Treatment for Academic Underachievement. *Psychological Science,* **26** (6), 784–793.

Popova, A., Evans, D. K. and Arancibia, V. (2016). Training Teachers on the Job: What Works and How to Measure It. World Bank Group, Policy Research Working Paper 7834.

Raven, J., Raven, J. and Court, J. H. (2004). Manual for Raven's Progressive Matrices and Vocabulary Scales. *San Antonio, TX: Harcourt Assessment*.

Rickford, J. R. (1999). African American Vernacular English: Features, Evolution, Educational Implication. Malden, MA: Blackwell Publishers.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, **73** (2), 417–458.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A. and Goldberg, L. R. (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.

Schanzenbach, D. (2006). What Have Researchers Learned from Project STAR? *Brookings Papers on Education Policy*, **9**, 205–228.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L. and Macnamara, B. N. (2018). To What Extent and Under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses. *Psychological Science*, **29** (4), 549–571.

Sriram, R. (2014). Rethinking Intelligence: The Role of Mindset in Promoting Success For Academically High-Risk Students. *Journal of College Student Retention,* **15** (4), 515–536.

Sutter, M., Kocher, M. G., Glätze-Rützler, D. and Trautmann, S. T. (2013). Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior. *American Economic Review*, **103** (1), 510–531.

Yeager, D. S. and Dweck, C. S. (2012). Mindsets that Promote Resilience: When Students Believe that Personal Characteristics Can Be Developed. *Educational Psychologist*, **47** (4), 302–314.

—, Hanselman, P., Walton, G. M., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C., Hinojosa, C., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang Hooper, S., Murray, J., Carvalho, C., Hahn, R., Ferguson, R., Duckworth, A. and Dweck, C. S. (2018). Where and For Whom Can a Brief, Scalable Mindset Intervention Improve Adolescents' Educational Trajectories? Unpublished Manuscript.

—, Johnson, R., Spitzer, B. J., Trzesniewski, K. H., Powers, J. and Dweck, C. S. (2014). The Far-Reaching Effects of Believing People Can Change: Implicit Theories of Personality Shape Stress, Health, and Achievement During Adolescence. *Journal of Personality and Social Psychology*, **106** (6), 867–884.

—, Lee, H. Y. and Jamieson, J. (2016a). How to Improve Adolescent Stress Responses: Insights From an Integration of Implicit Theories of Biopsychosocial Models. *Psychological Science*, **27** (8), 1078–1091.

—, Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., Lee, H. Y., O'Brien, J., Flint, K., Roberts, A., Trott, J., Greene, D., Walton, G. M. and Dweck, C. S. (2016b). Using Design Thinking to Improve Psychological Interventions: The Case of the Growth Mindset During the Transition to High School. *Journal of Educational Psychology*, **108** (3), 374–391.

— and Walton, G. M. (2011). Social-Psychological Interventions in Education: They're Not Magic. *Review of Educational Research*, **81** (2), 267–301.

Young, A. (forthcoming). Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Result. *Quarterly Journal of Economics*, **https://doi.org/10.1093/qje/qjy029**.

Tables

Table I Design

	A: Sample 1			*B: Sample 2*	
	Patience+Grit (15 schools)	Patience (9 schools)	Control (12 schools)	Grit (8 schools)	Control (8 schools)
Baseline Data Collection	Mar ’13	Mar ’13	Mar ’13	May ’15	May ’15
Patience Training	Spring ’13	Fall ’13	-	-	-
Grit Training	Fall ’13	-	-	Fall ’15	-
Short-Run Follow-up Data Collection	May ’14	May ’14	May ’14	Jan ’16	Jan ’16
Long-Run Follow-up Data Collection	Mar ’16	Mar ’16	Mar ’16	Jun ’17	Jun ’17
	(2.5 years)	(2.5 years)	(2.5 years)	(1.5 years)	(1.5 years)

Table II Mean Comparisons of Pre-Treatment Variables

	A: Sample 1			*B: Sample 2*		
	Control Mean [SD]	Treatment Mean [SD]	Difference (p-value)	Control Mean [SD]	Treatment Mean [SD]	Difference (p-value)
Beliefs (survey)	0.03	-0.02	-0.05	-0.02	0.02	0.04
	[1.00]	[1.00]	(0.64)	[1.02]	[0.98]	(0.64)
Grit (survey)	-0.01	0.01	0.03	0.05	-0.07	-0.12
	[1.01]	[0.99]	(0.85)	[0.98]	[1.02]	(0.23)
Gender (Male=1)	0.53	0.51	-0.01	0.50	0.52	0.03
	[0.50]	[0.50]	(0.46)	[0.50]	[0.50]	(0.27)
Age	10.02	10.03	0.01	9.43	9.46	0.03
	[0.44]	[0.48]	(0.64)	[0.53]	[0.47]	(0.47)
Raven	0.02	-0.02	-0.03	0.08	-0.11	-0.19*
	[1.00]	[1.00]	(0.83)	[0.97]	[1.03]	(0.10)
Risk Tolerance	2.60	2.51	-0.09	2.17	2.21	0.05
	[1.49]	[1.52]	(0.52)	[1.51]	[1.67]	(0.84)
Wealth	2.86	2.75	-0.11	2.61	2.68	0.08
	[0.94]	[1.02]	(0.46)	[1.09]	[0.93]	(0.68)
Success in School	3.41	3.28	-0.13	3.42	3.30	-0.12
	[1.05]	[1.12]	(0.14)	[1.05]	[1.14]	(0.37)
Class Size	37.17	42.51	5.34	35.13	39.98	4.85
	[8.20]	[9.62]	(0.14)	[5.52]	[8.36]	(0.14)
Math Test Score	0.05	-0.04	-0.09	0.00	-0.00	-0.01
	[0.97]	[1.02]	(0.57)	[1.03]	[0.97]	(0.94)
Verbal Test Score	0.08	-0.07	-0.15	0.10	-0.13	-0.23**
	[0.92]	[1.05]	(0.43)	[0.97]	[1.03]	(0.03)
Task Ability	4.88	4.78	-0.10	3.68	3.94	0.26
	[2.39]	[2.32]	(0.64)	[2.19]	[2.12]	(0.13)
N	1,132	1,443		816	683	

Notes. Columns 1-2 and 4-5 display the means of the pre-treatment variables in the control and treatment groups for Samples 1 and 2, respectively. Standard deviations are displayed in brackets. Columns 3 and 6 show the estimated difference in means which is obtained from regressing the variable of interest on the treatment dummy. Standard errors are clustered at the school level (unit of randomization) and p-values are reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The variables beliefs (about the malleability of skills) and grit are extracted factors from questionnaire items in the pre-treatment student survey. The Raven score is measured using a progressive Raven's matrices test (Raven *et al.*, 2004). Task ability refers to the student's performance in the ability measuring round of the experiment. Risk tolerance is elicited using the incentivized Gneezy and Potters (1997) task. The student's wealth and success in school is based on reports by teachers (scale: 1-5). Students' math and verbal baseline test scores are normalized (mean 0, standard deviation 1). For Sample 1, these test scores refer to the grades given to the students by their teachers, while for Sample 2 they refer to the students' performance on the standardized tests we administer.

Table III Treatment Effect on Choice of Difficult Task

	Difficult Round 1	Difficult Round 2	Difficult Round 3	Difficult Round 4	Difficult Round 5	Difficult All	After Failure	Next Week
A: Sample 1								
Treatment	0.102^{***}	0.088^{**}	0.126^{***}	0.108^{***}	0.089^{***}	0.088^{***}	0.145^{***}	0.135^{***}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)
Permutation p-value	0.004	0.034	0.002	0.009	0.003	0.006	0.051	0.000
Control Mean	0.67	0.54	0.43	0.42	0.40	0.24	0.40	0.45
N	1889	1884	1885	1882	1886	1862	642	1858
B: Sample 2								
Treatment	0.098^{**}	0.157^{***}	0.157^{***}	0.157^{***}	0.131^{***}	0.121^{***}	0.149^{*}	0.179^{***}
	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.08)	(0.04)
Permutation p-value	0.017	0.002	0.004	0.006	0.005	0.013	0.109	0.004
Control Mean	0.67	0.51	0.35	0.30	0.26	0.16	0.50	0.41
N	1354	1351	1351	1350	1354	1335	585	1349

Notes. Reported estimates are average marginal effects from logit regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The outcome variable in columns 1-5 is a dummy variable which equals one if the student chooses to do the difficult task in the respective round of the first visit, while the outcome variable in column 6 equals one if the student chooses the difficult task in all five rounds. The outcome variable in column 7 is a dummy variable which equals 1 if the student chooses to do the difficult task in the 2nd round of the first visit; estimates are obtained for students for whom the difficult task was imposed in round 1 and who failed to meet the target. The outcome variable in column 8 is a dummy which equals 1 if the student chooses to do the difficult task for the following week. Treatment is a dummy variable which equals 1 if the student attends a school which has been treated with the grit intervention. Controls include task ability, gender, the Raven score, baseline beliefs and test scores, and risk tolerance as well as a dummy variable for whether the student has some inconsistent data entries.

Table IV Treatment Effect on Success and Payoffs in the First Visit

	Success Round 1	Payoff Round 1	Payoff Round 2	Payoff Round 3	Payoff Round 4	Payoff Round 5
A: Sample 1						
Treatment	0.023	0.006	0.027	0.029	0.101	0.053
	(0.04)	(0.09)	(0.06)	(0.07)	(0.09)	(0.09)
Permutation p-value	0.593	0.951	0.650	0.705	0.294	0.560
Control Mean	0.29	1.33	0.99	1.35	1.20	1.26
N	917	1878	1866	1874	1870	1872
B: Sample 2						
Treatment	0.045	0.225**	0.012	0.081	0.009	0.062
	(0.03)	(0.10)	(0.09)	(0.08)	(0.05)	(0.08)
Permutation p-value	0.147	0.036	0.903	0.382	0.828	0.410
Control Mean	0.20	0.77	0.65	1.01	0.92	1.00
N	750	1350	1350	1349	1348	1350

Notes. Reported estimates in column 1 are average marginal effects from a logit regression. Reported estimates in columns 2-6 are obtained via Ordinary Least Squares regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The outcome variable in column 1 is a dummy variable which equals 1 if the student was successful in meeting the target. Estimates in column 1 are obtained for students for whom the difficult task was imposed. The outcome variable in columns 2-6 is the student's payoff in the respective round. Estimates are obtained for all students. Treatment is a dummy variable which equals 1 if the student attends a school which has been treated with the grit intervention. Controls include task ability, gender, the Raven score, baseline beliefs and test scores, and risk tolerance as well as a dummy variable for whether the student has some inconsistent data entries.

Table V Treatment Effect on Success and Payoffs in the Second Visit

	Success	Payoff			Total Payoff	Maximizing Choice	
	Imposed	All	Imposed	Not Imposed	All	Visit 1	Visit 2
A: Sample 1							
Treatment	0.084***	0.297***	0.323**	0.245**	0.359***	0.017	0.078*
	(0.03)	(0.09)	(0.13)	(0.10)	(0.13)	(0.02)	(0.04)
Permuted p-value	0.016	0.004	0.026	0.058	0.002	0.488	0.073
Control Mean	0.47	1.82	1.87	1.75	3.10	0.62	0.55
N	1101	1969	1101	868	1710	1868	1567
B: Sample 2							
Treatment	0.103**	0.450***	0.399**	0.576***	0.552***	0.064***	0.082***
	(0.04)	(0.12)	(0.17)	(0.11)	(0.15)	(0.02)	(0.03)
Permuted p-value	0.040	0.008	0.049	0.012	0.009	0.012	0.012
Control Mean	0.45	1.70	1.81	1.55	2.61	0.47	0.50
N	878	1350	878	472	1248	1344	1266

Notes. Reported estimates in columns 1, 6 and 7 are average marginal effects from logit regressions. Estimates in columns 2-5 are obtained via Ordinary Least Squares regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The outcome variable in column 1 is a dummy which equals 1 if the student was successful in meeting the target. The outcome in columns 2-4 is the student's payoff in visit 2. The sample used in the analysis either contains all observations ("All"), the observations for whom the difficult game was imposed ("Imposed") or for whom it was not imposed ("Not Imposed"). The outcome variable in column 5 is the sum of the average payoff in visit 1 and the payoff in visit 2. the outcome variable in columns 6 and 7 is a dummy variable which indicates whether the student makes the payoff-maximizing choice in visit 1 and visit 2, respectively. Treatment is a dummy variable which equals 1 if the student attends a school which has been treated with the grit intervention. Controls include task ability, gender, the Raven score, baseline beliefs and test scores, and risk tolerance as well as a dummy variable for whether the student has some inconsistent data entries.

Table VI Treatment Effect on Grades Given by Teacher

	A: Sample 1		*B: Sample 2*	
	Math Grade	Verbal Grade	Math Grade	Verbal Grade
Treatment	-0.054	-0.013	0.002	-0.006
	(0.10)	(0.07)	(0.11)	(0.13)
Permutation p-value	0.623	0.863	0.992	0.982
Control Mean	0.06	0.05	0.10	0.09
N	2237	2233	1404	1404

Notes. Estimates are obtained via Ordinary Least Squares regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The dependent variables are the students' math and verbal grades at follow-up, which were given by the teacher. Treatment is a dummy variable which equals 1 if the student attends a school which has been treated with the grit intervention. Controls include gender, the Raven score, class size and baseline beliefs and test scores.

Table VII Treatment Effect on Standardized Test Scores

	A: Sample 1		*B: Sample 2*			
	Math Score Long-Run	Verbal Score Long-Run	Math Score Short-Run	Verbal Score Short-Run	Math Score Long-Run	Verbal Score Long-Run
Treatment	0.225**	0.046	0.311***	0.126*	0.190***	0.043
	(0.09)	(0.07)	(0.09)	(0.06)	(0.06)	(0.08)
Permutation p-value	0.044	0.572	0.008	0.105	0.026	0.625
Control Mean	-0.09	0.02	-0.06	0.01	-0.02	0.06
N	1040	1036	1347	1350	781	778

Notes. Estimates are obtained via Ordinary Least Squares regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The dependent variables are the students' math and verbal standardized test scores at follow-up. The long-run follow-up data for Sample 1 was collected 2.5 years after the intervention. For Sample 2, the short-run and the long-run follow-up data were collected immediately after the implementation of the intervention and 1.5 years after the intervention, respectively. Treatment is a dummy variable which equals 1 if the student attends a school which has been treated with the grit intervention. Controls include gender, the Raven score, class size, baseline beliefs and test scores.

Figures

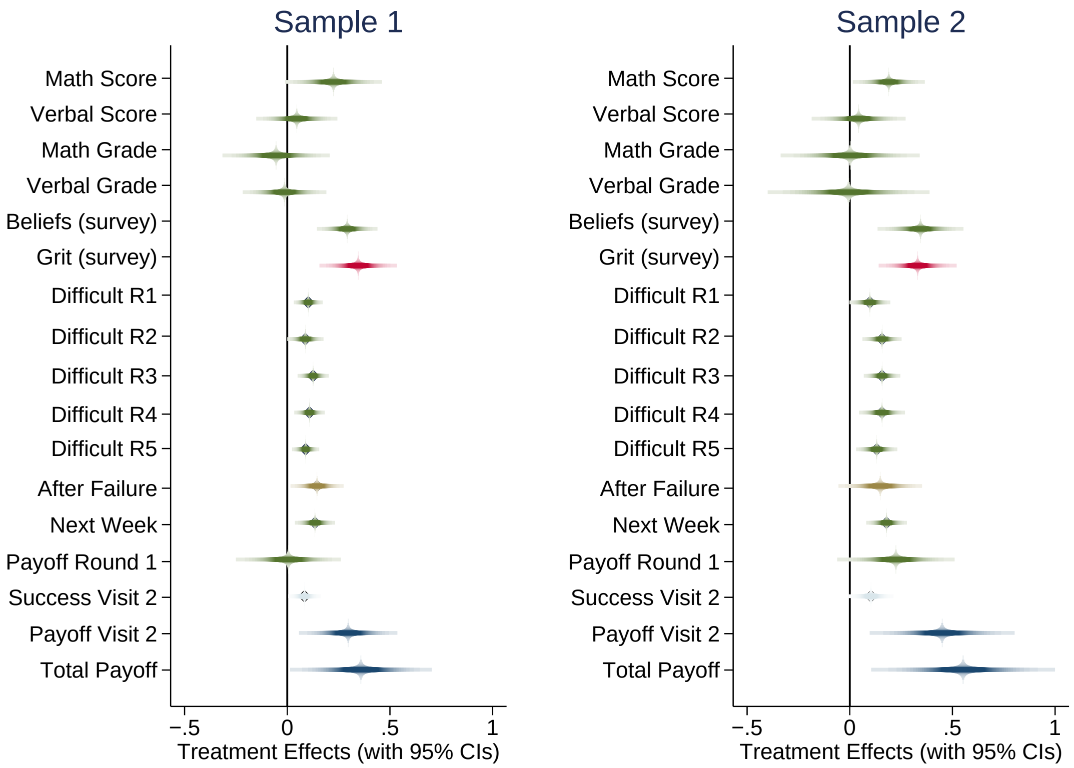

Figure I Estimated Treatment Effect Coefficients.

Notes. The figure depicts the estimated treatment effects and their 95% confidence intervals (see Tables III-VII and Online Appendix Table A.17). Confidence intervals are based on standard errors clustered at the school level (unit of randomization). The vertical line indicates a treatment effect of zero. The first four outcomes are long-run test scores and grades, respectively, followed by the standardized survey constructs of beliefs (growth mindset) and grit. The remaining outcomes come from the incentivized task. *Difficult R1-R5*: Binary choice of difficult task (rounds 1-5). *After Failure*: Binary choice of difficult task in round 2 conditional on failing in round 1 (for sample in which the difficult task was imposed in round 1). *Next Week*: Binary choice of difficult task for week 2. *Payoff Round 1*: Payoff in round 1, week 1. *Success Visit 2*: Success rate in visit 2 (for sample in which difficult task was imposed in visit 2). *Payoff Visit 2*: Payoff in visit 2. *Total Payoff*: Total payoff from both visits.

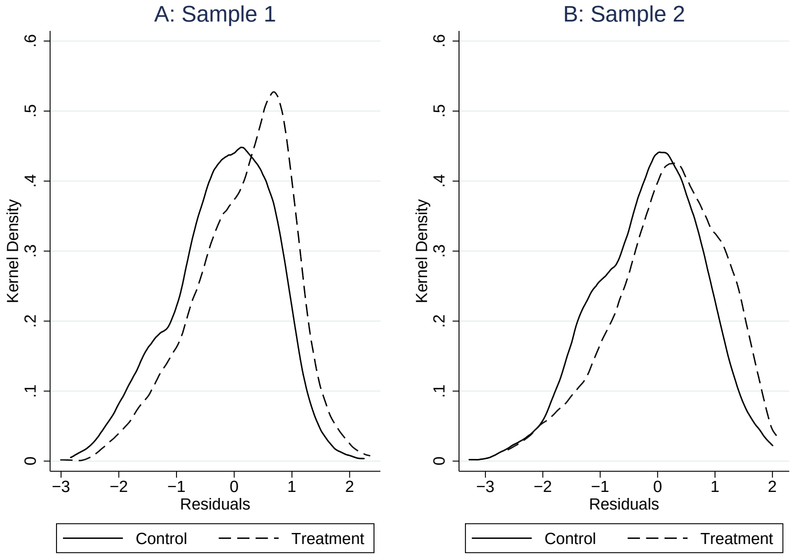

Figure II Effect of Treatment on Self-Reported Malleability Beliefs.

Notes. This figure displays the distribution of self-reported beliefs about the malleability of skills at follow-up which cannot be explained by baseline covariates. Residuals are calculated on the basis of the regressions presented in Online Appendix Table A.17.

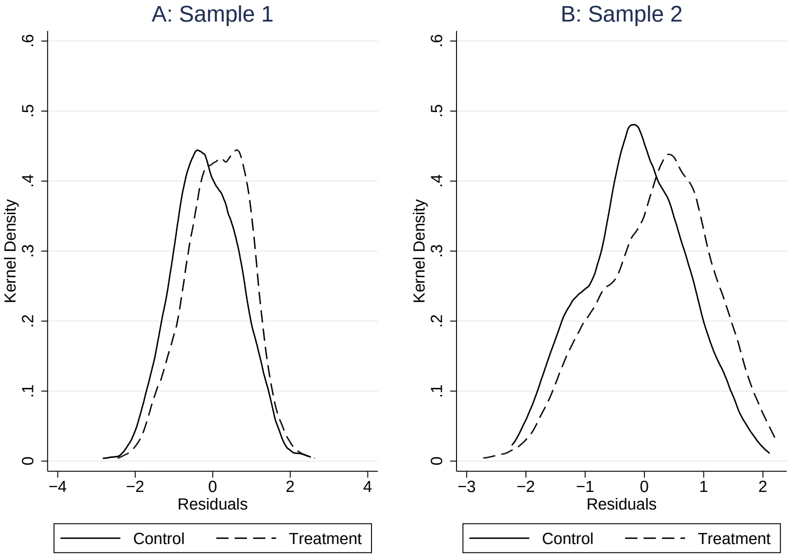

Figure III Effect of Treatment on Self-Reported Grit.

Notes. This figure displays the distribution of self-reported grit at follow-up which cannot be explained by baseline covariates. Residuals are calculated on the basis of the regressions presented in Online Appendix Table A.17.