THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 134 2019 Issue 3

EVER FAILED, TRY AGAIN, SUCCEED BETTER: RESULTS FROM A RANDOMIZED EDUCATIONAL INTERVENTION ON GRIT[*]

Sule Alan
Teodora Boneva
Seda Ertac

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, and many other students provided excellent research assistance. This article 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 article are registered with the AEA Trial Registry (RCT ID: AEARCTR-0003317 and AEARCTR-0003327). Koç University IRB approval no: 2013.067.IRB.32.

The Quarterly Journal of Economics (2019), 1121–1162. doi:10.1093/qje/qjz006.
Advance Access publication on February 19, 2019.

I. Introduction

The growing literature on human capital accumulation has emphasized the importance of noncognitive skills in explaining individual differences in achievement in various economic and social domains (Heckman, Stixrud, and Urzua 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. Noncognitive skills such as patience, self-control, and grit have been shown to be highly predictive of outcomes ranging from educational attainment and occupational and financial success to criminal activity and health outcomes; see Heckman, Stixrud, and Urzua (2006); Almlund et al. (2011); Dohmen et al. (2011); Sutter et al. (2013); Heckman, Humphries, and Mader (2011); Moffit et al. (2011); Castillo et al. (2011); Golsteyn, Grönqvist, and Lindahl (2013). In fact, the predictive power of noncognitive skills appears to rival that of cognitive skills (Roberts et al. 2007; Kautz et al. 2014). More important 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 article. 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, Shulman, and Duckworth (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

[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 Project STAR (Schanzenbach 2006; Dee and West 2011; Chetty et al. 2011).

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 people to work harder in educational and occupational settings, it is important to understand the nature of grit and explore ways of enhancing it. In this article, we evaluate a randomized educational intervention that aims to foster grit in the classroom environment. We conjecture that an intervention that instills optimistic beliefs about the productivity of effort and encourages children 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. 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, for example, 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 multifaceted 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 posttreatment questionnaires. We collect this information from over 3,200 fourth-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 reattempt the difficult task after receiving negative performance feedback. The design of our incentivized task also 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 greater 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, the estimated short-term effect seems to have dissipated in both samples. We find that the estimated treatment effects on 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 article 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). Although the early studies produced very promising results (e.g., Aronson, Fried, and Good 2002; Good, Aronson, and Inzlicht 2003; Blackwell, Trzesniewski, and Dweck 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 9th–12th 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. (2016) investigate the impact of a two-period mindset intervention in a sample of ninth-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 ninth-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, Lee, and Jamieson (2016) and Bettinger et al. (2018). Although 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 and content. Trained teachers spend 12 two-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 thus focuses more directly on encouraging actual perseverant behavior in class, in addition to introducing students to a set of ideas. This also ensures that treated students are exposed to the concepts and ideas for the duration of an entire school year, not just within the limited project hours. Second, we conduct a long-term follow-up for both 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 (an online goal-setting exercise, a text message campaign, and a personal coach) and find that the personal coaching program has large effects on student achievement, while the low-cost interventions relying on technology have no effects on academic outcomes. Dobronyi, Oreopoulos, and Petronijevic (2017) test the effectiveness of two online goal-setting interventions, one of which includes a growth mindset component, and 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 although 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 noncognitive 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, noncognitive 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 noncognitive 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 article is organized as follows. Section II presents details on the design and implementation of the educational intervention and 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 postelementary schools to participate in extracurricular projects offered by the private sector, NGOs, the government, and international organizations. After being examined and endorsed by the ministry, these projects are made available to schools. Participation in these projects is at the discretion of teachers. The ministry allows up to five lecture hours a week for project-related classroom activities. The program we evaluate was implemented as an extracurricular project of this type.

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

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 anyone 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 two hours/week on the project over the course of 12 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 with 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 or fear of math and other challenging tasks. Visual materials and stories are supplemented with 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]

Volunteer 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. Instructors first introduced the concepts and their importance for academic achievement. They 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 in the materials as part of a teaching philosophy. To do this, they were given various pedagogical guidelines. These include praising students' effort and championing perseverant behavior and positive attitudes toward learning, rather than just praising good outcomes. Teachers were encouraged not to praise a successful student in a way that 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 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 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 in the curriculum materials and in pedagogical guidelines.

[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.

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 represents Turkey's lower socioeconomic segment. The program we evaluate is the second arm of a two-arm randomized controlled trial (RCT) 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 their 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 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, and the CT group (nine 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 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 article, 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) to evaluate the effect of the grit intervention. Notice 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 Online Appendix Tables A.1, A.2, and A.3), we cannot rule out the effect of dynamic complementarities.[5] The second study, which

[4] After this follow-up, we lost one 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 Sam-

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) 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 how 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 patience treatment. This allows us to 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, Blom, and Meghir 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 a chance 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, that is, while children in control groups would not 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 gener-

ple 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.

[6] Another difference between Sample 1 and Sample 2 is that the students in Sample 2 are about six 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.

ated 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 Sample 1 approximately 60% of the contacted teachers accepted our offer, and the most common reason for nonparticipation was being "busy with other projects, although happy to participate in this program at a later date" (about 20%). The rest of the nonparticipation 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 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, and 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 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, toward the end of the school year. By that time, students had been exposed to the trained teacher for almost the whole academic year. In Sample 2, the experiments were carried out in January 2016, shortly after the teachers had covered the 12-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 easy tracking of students when they change schools, 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 sixth grade for Sample 1 and fifth grade in Sample 2. We matched the lists with our elementary school data based on student and elementary school name; 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 in Sample 1 (p-value = .883) and in Sample 2 (p-value = .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 Online Appendix Table A.4, 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, for example, in baseline verbal test scores. We use a number of baseline variables as covariates in the regressions to correct for potential imbalances and use inverse probability weights to account for possible differential attrition. Details of the evaluation designs for each study sample are given in Table I.

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 extracurricular 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 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. In the first visit, children go through five rounds of a mathematical

[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.

TABLE I
DESIGN

	Sample 1			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 (2.5 years)	Mar ’16 (2.5 years)	Mar ’16 (2.5 years)	Jun ’17 (1.5 years)	Jun ’17 (1.5 years)

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 round 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 that sum to 100 within 1.5 minutes.[8] The rewards consist of gifts of value to children of this age group. These include fun stationery items, small puzzles, skipping ropes, flying discs, 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: (i) the "four-gift game," which yields four gifts in the case of success and zero in the case of failure, and (ii) the "one-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 four-gift game is more difficult than the one-gift game. In particular, in the one-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 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) are significantly lower in value than the rewards in the actual task, and children are aware of this. In addition, information about actual rewards they receive from this task is not revealed until the end of the first visit. In the main five-round part of the experiment, subjects are distributed two booklets of five pages each, the four-gift game booklet and the one-gift game booklet. Each booklet contains five pages that correspond to the 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

[8] Note that while Sample 1 students are given 1 minute, 30 seconds 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 Online Appendix B for examples of the two types of task.

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 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 three pairs were correctly found. As mentioned students have the opportunity to switch back and forth between the two types of tasks as the rounds progress.

These 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 second round onward, 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 in exactly one 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 four-gift (difficult) game or the one-gift (easy) game at that point. They are told that they will have access to an “exercise booklet”, which contains examples and practice questions that have a similar difficulty level to the four-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 second visit at the end of the first 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. 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 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 we deliberately avoided wording/terminology that was frequently used in the intervention (e.g., grit, quitting, challenge) during the data collection. The experiments were labeled 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 they 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, 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 (1,899) in Sample 1 were present on the day of testing and consented to participate, while 91% (1,360 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, Raven, and Court 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 0 and standard deviation 1. 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 five-point scale (1: very low, 5: very high). The success variable asks teachers to rate the students' overall academic performance on a five-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 0 and a standard deviation of 1.[11]

[10] We collected experimental outcomes in all but four classrooms in Sample 1, which we could not visit due to scheduling constraints toward 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.

TABLE II

Mean Comparisons of Pretreatment Variables

	Sample 1			Sample 2		
	Control	Treatment		Control	Treatment	
	mean	mean	Difference	mean	mean	Difference
	[std. dev.]	[std. dev.]	(*p*-value)	[std. dev.]	[std. dev.]	(*p*-value)
	(1)	(2)	(3)	(4)	(5)	(6)
Beliefs (survey)	0.03	−0.02	−0.05	−0.02	0.02	0.04
	[1.00]	[1.00]	(.64)	[1.02]	[0.98]	(.64)
Grit (survey)	−0.01	0.01	0.03	0.05	−0.07	−0.12
	[1.01]	[0.99]	(.85)	[0.98]	[1.02]	(.23)
Gender (male = 1)	0.53	0.51	−0.01	0.50	0.52	0.03
	[0.50]	[0.50]	(.46)	[0.50]	[0.50]	(.27)
Age	10.02	10.03	0.01	9.43	9.46	0.03
	[0.44]	[0.48]	(.64)	[0.53]	[0.47]	(.47)
Raven	0.02	−0.02	−0.03	0.08	−0.11	−0.19*
	[1.00]	[1.00]	(.83)	[0.97]	[1.03]	(.10)
Risk tolerance	2.60	2.51	−0.09	2.17	2.21	0.05
	[1.49]	[1.52]	(.52)	[1.51]	[1.67]	(.84)
Wealth	2.86	2.75	−0.11	2.61	2.68	0.08
	[0.94]	[1.02]	(.46)	[1.09]	[0.93]	(.68)
Success in school	3.41	3.28	−0.13	3.42	3.30	−0.12
	[1.05]	[1.12]	(.14)	[1.05]	[1.14]	(.37)
Class size	37.17	42.51	5.34	35.13	39.98	4.85
	[8.20]	[9.62]	(.14)	[5.52]	[8.36]	(.14)
Math test score	0.05	−0.04	−0.09	0.00	−0.00	−0.01
	[0.97]	[1.02]	(.57)	[1.03]	[0.97]	(.94)
Verbal test score	0.08	−0.07	−0.15	0.10	−0.13	−0.23**
	[0.92]	[1.05]	(.43)	[0.97]	[1.03]	(.03)
Task ability	4.88	4.78	−0.10	3.68	3.94	0.26
	[2.39]	[2.32]	(.64)	[2.19]	[2.12]	(.13)
N	1,132	1,443		816	683	

Notes. Columns (1), (2), (4), and (5) display the means of the pretreatment 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. * $p < .10$, ** $p < .05$, *** $p < .01$. The variables beliefs (about the malleability of skills) and grit are extracted factors from questionnaire items in the pretreatment student survey. The Raven score is measured using a progressive Raven's matrices test (Raven, Raven, and Court 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.

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 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 Table II, the number of pairs found in this task (referred to as "task ability") is not different across treatment status in either sample.

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 that equals 1 if school j is in the treatment group and 0 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 or more correct pairs. In the case of payoffs, the equation is estimated via OLS. 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, and 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 (2019), 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 nonzero intracluster 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 rerandomize 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 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. For the sake of brevity, all tables in this section present the estimated treatment effects without presenting the coefficient estimates of the covariates.

1. First Visit. In the first visit, students are asked to choose between the one-gift (easy) game and the four-gift (difficult) game in the five main rounds of the experiment. With the exception of the first round, in which some students are randomly selected to

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

do the difficult task irrespective of their choice, students perform the task of their choice before moving on to the next round. In both samples the vast majority of students was successful on the easy task, but this was 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 regardless 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, whereas for Sample 2 the corresponding number is 20%. Given that the difficult task yields four gifts in the case of success and the easy task only yields one 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 5 in Sample 1 (Sample 2). Although 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 with 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 1 and the effect also persists until round 5, 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]

[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, for example, because they had to go to the bathroom.

TABLE III

Treatment Effect on Choice of Difficult Task

	Difficult round 1 (1)	Difficult round 2 (2)	Difficult round 3 (3)	Difficult round 4 (4)	Difficult round 5 (5)	Difficult all (6)	After failure (7)	Next week (8)
Panel 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	.004	.034	.002	.009	.003	.006	.051	.000
Control Mean	0.67	0.54	0.43	0.42	0.40	0.24	0.40	0.45
N	1,889	1,884	1,885	1,882	1,886	1,862	642	1,858
Panel 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	.017	.002	.004	.006	.005	.013	.109	.004
Control mean	0.67	0.51	0.35	0.30	0.26	0.16	0.50	0.41
N	1,354	1,351	1,351	1,350	1,354	1,335	585	1,349

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. * $p < .10$, ** $p < .05$, *** $p < .01$. The outcome variable in columns (1)–(5) is a dummy variable that equals 1 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 1 if the student chooses the difficult task in all five rounds. The outcome variable in column (7) is a dummy variable that equals 1 if the student chooses to do the difficult task in the second 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 that equals 1 if the student chooses to do the difficult task for the following week. Treatment is a dummy variable that equals 1 if the student attends a school that 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.

Next we estimate the effect of treatment on task choice in round 2 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 1 are significantly more likely to want to reattempt the difficult task in round 2, even though 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 = .47; Sample 2 p-value = .24). Success in the imposed difficult task, while not exogenous, is balanced across treatment and control group students (Sample 1 p-value = .59, Sample 2 p-value = .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 reattempt the difficult task in the subsequent round (permutation p-value = .05). The corresponding estimate in Sample 2 is also 15 percentage points (permutation p-value = .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. Although we cannot rule out unobserved differences between treatment and control group students who failed at the imposed difficult task, these results strongly suggest that the intervention affects how 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 an 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 with it. 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

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
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 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	.593	.951	.650	.705	.294	.560
Control mean	0.29	1.33	0.99	1.35	1.20	1.26
N	917	1,878	1,866	1,874	1,870	1,872
Panel 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	.147	.036	.903	.382	.828	.410
Control mean	0.20	0.77	0.65	1.01	0.92	1.00
N	750	1,350	1,350	1,349	1,348	1,350

Notes. Reported estimates in column (1) are average marginal effects from a logit regression. Reported estimates in columns (2)–(6) are obtained via OLS regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$. The outcome variable in column (1) is a dummy variable that 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 that equals 1 if the student attends a school that 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.

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 = .000), and 18 percentage points more likely to plan to play the difficult game in Sample 2 (permutation p-value = .004).

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 a way that allows us to estimate the treatment effect on success in the difficult game free of selection. As mentioned, we find no significant treatment effect on success rates in either sample (permutation p-values are .59 and .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 0, 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.

2. Second Visit. The temporal component of our experimental task serves a very important purpose for our study. Although 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, and 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. 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

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
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel 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	.016	.004	.026	.058	.002	.488	.073
Control mean	0.47	1.82	1.87	1.75	3.10	0.62	0.55
N	1,101	1,969	1,101	868	1,710	1,868	1,567
Panel 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	.040	.008	.049	.012	.009	.012	.012
Control mean	0.45	1.70	1.81	1.55	2.61	0.47	0.50
N	878	1,350	878	472	1,248	1,344	1,266

Notes. Reported estimates in columns (1), (6), and (7) are average marginal effects from logit regressions. Estimates in columns (2)–(5) are obtained via OLS regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$. The outcome variable in column (1) is a dummy that 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 that indicates whether the student makes the payoff-maximizing choice in visit 1 and visit 2, respectively. Treatment is a dummy variable that equals 1 if the student attends a school that 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.

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 encouraging her to study. It may encourage a student with low ability to study hard as well. Because 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, or cognitive ability.[15]

IV.C. Are Choices Payoff-Maximizing?

An important question regarding an intervention of this sort is whether being gritty is good for everyone, that is, 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.

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

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, because effort costs are unobservable. However, the choices and outcomes of treated students in the second visit suggest, through revealed preference, that these choices might also be utility maximizing for this group.

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

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 teacher-given grades may be unaffected by the intervention is that teachers in elementary school tend to apply

TABLE VI

Treatment Effect on Grades Given by Teacher

	Sample 1		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	.623	.863	.992	.982
Control mean	0.06	0.05	0.10	0.09
N	2,237	2,233	1,404	1,404

Notes. Estimates are obtained via OLS regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$. The dependent variables are the students' math and verbal grades at follow-up, which were given by the teacher. Treatment is a dummy variable that equals 1 if the student attends a school that has been treated with the grit intervention. Controls include gender, the Raven score, class size, baseline beliefs, and test scores.

TABLE VII

Treatment Effect on Standardized Test Scores

	Sample 1		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	.044	.572	.008	.105	.026	.625
Control mean	−0.09	0.02	−0.06	0.01	−0.02	0.06
N	1,040	1,036	1,347	1,350	781	778

Notes. Estimates are obtained via OLS regressions. Standard errors are clustered at the school level (unit of randomization) and reported in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$. 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 that equals 1 if the student attends a school that has been treated with the grit intervention. Controls include gender, the Raven score, class size, baseline beliefs, and test scores.

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 = .008) on standardized math scores and 0.13 (permutation p-value = .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 = .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 = .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. Online Appendix Table A.12 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.

Compared with 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 Project STAR, concludes that being randomly assigned to a small class raises student test scores by 0.15 standard deviations. Note, however, that although we estimate a large effect immediately after the program implementation, the estimated effects 2.5 years following the implementation are smaller and more in line with the literature. Note also that we deliberately target low SES 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) summarizes the lessons learned from close to 200 randomized field experiments in education and 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. First, 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, Scarborough, and Griffin 2004). Second, 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, Hanushek, and Kain 2005; Hanushek 2011; Hanushek and Rivkin 2012). Previous studies have shown that teachers affect later-life outcomes of students through influencing their test scores and their noncognitive skills (see, e.g., Chetty et al. 2011; Chetty, Friedman, and Rockoff 2014; Jackson 2018). Consistently with this, educational policymakers in many countries provide professional development programs for teachers (Popova, Evans, and Arancibia 2016). We relate to this literature by showing that a program based on training teachers has the potential to raise students' test scores and their noncognitive skills as measured through a behavioral task. To the extent that changes in noncognitive skills are persistent and lead to better life outcomes, it is plausible to expect that the impact of the program on students' noncognitive 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. Online Appendix Tables A.13 and A.14 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, that is, 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. Although the

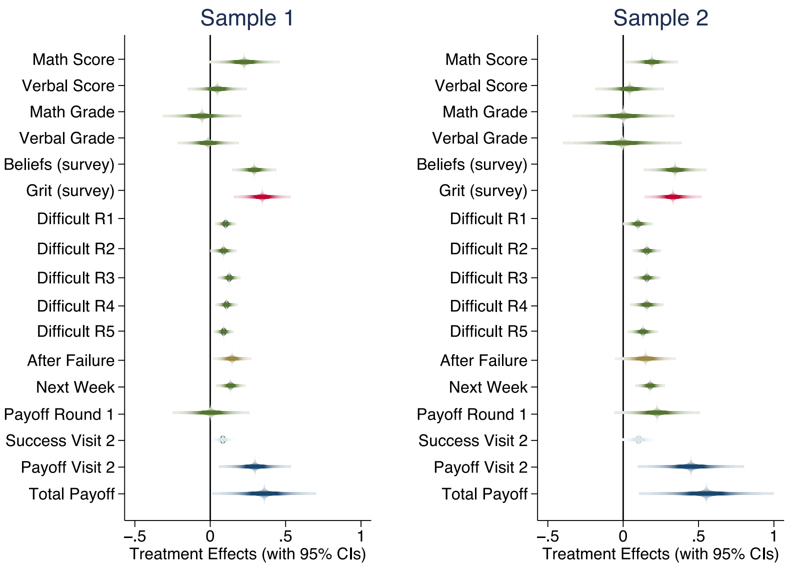

FIGURE I

Estimated Treatment Effect Coefficients

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 0. 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.

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, Tufano, and List 2014). This is especially important in our setting in which attrition rates lower the power of our design.

V. Discussion

Although 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 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 Online Appendix Table A.17). 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.

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, that is, their self-confidence, is one alternative belief channel that could lead to ambitious goal setting. We should note, however, that our intervention does not aim to increase self-confidence about ability but 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 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 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

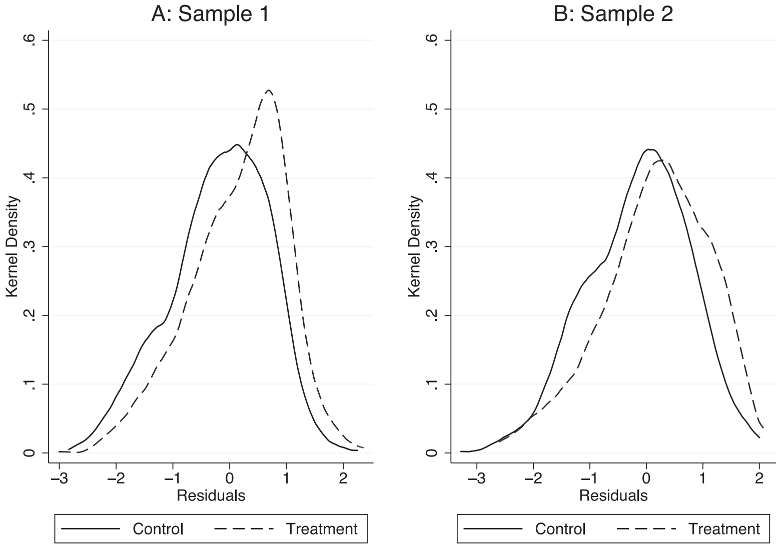

FIGURE II

Effect of Treatment on Self-Reported Malleability Beliefs

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

find that the treatment had no effect on students' self-confidence in their ability (p-value = .81). In terms of other attitudes and behaviors, we consider students' attitude toward risk and patience, which may have been affected by the treatment and may have mediated our estimated treatment effects. Risk-tolerant people 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 and in our behavioral task. We do not estimate statistically significant treatment effects on either risk tolerance (p-value = .52) or patience (p-value = .97).[16]

[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.

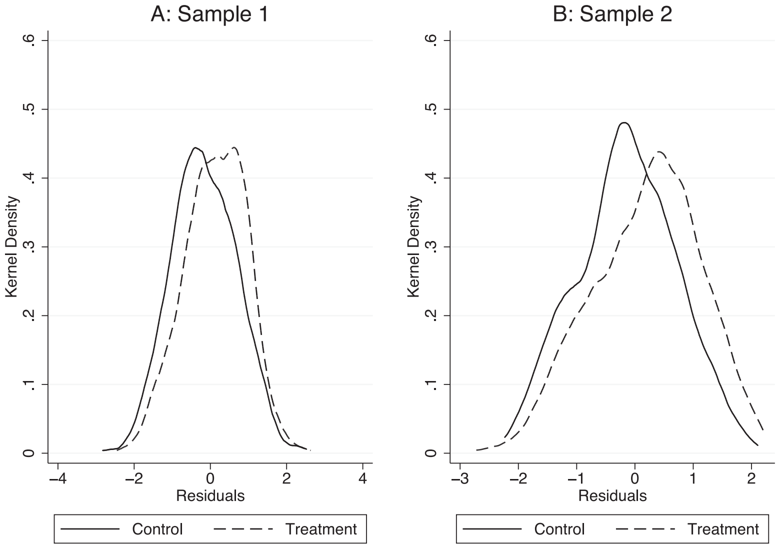

FIGURE III

Effect of Treatment on Self-Reported Grit

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

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 ruled out this potential channel at the design stage by making sure that our control teachers were also engaged in ministry-approved extracurricular 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 reemphasize 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, Mechtenberg, and Petersen 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 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, engage in skill-accumulation activities, and 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 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 article contributes to the ongoing debate about the malleability of noncognitive 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 noncognitive 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 Abdul Latif Jameel Poverty Action Lab
University of Oxford
Koç University

Supplementary Material

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online. Data and code replicating tables and figures in this article can be found in Alan et al. (2019), in the Harvard Dataverse, doi:10.7910/DVN/SAVGAL.

References

Abdulkadiroglu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak, "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots," *Quarterly Journal of Economics*, 126 (2011), 699–748.

Alan, Sule, Teodora Boneva, and Seda Ertac, "Replication Data for: 'Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit'," Harvard Dataverse (2019), doi: 10.7910/DVN/SAVGAL.

Alan, Sule, and Seda Ertac, "Belief in Hard Work and Altruism: Evidence from a Randomized Experiment," Unpublished Manuscript, University of Essex, 2017.

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

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

Almlund, Mathilde, Angela L. Duckworth, James J. Heckman, and Tim D. Kautz, "Personality Psychology and Economics," in *Handbook of the Economics of Education*, E. Hanushek, S. Machin, and L. Woessman, eds. (Amsterdam: North-Holland, 2011), 1–181.

Altonji, Joseph G., Erica Blom, and Costas Meghir, "Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers," *Annual Review of Economics*, 4 (2012), 185–223.

Andreoni, James, and Charles Sprenger, "Estimating Time Preferences from Convex Budgets," *American Economic Review*, 102 (2012), 3333–3356.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters, "Who Benefits from KIPP?," *Journal of Policy Analysis and Management*, 31 (2012), 837–860.

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

Aucejo, Esteban M., and Jonathan James, “The Path to College Education: The Role of Math and Verbal Skills,” Working Paper, 2019.

Bettinger, Eric, Sten Ludvigsen, Mari Rege, Ingeborg F. Solli, and David Yeager, “Increasing Perseverance in Math: Evidence from a Field Experiment in Norway,” *Journal of Economic Behavior and Organization*, 146 (2018), 1–15.

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

Bettinger, Eric P., Bridget T. Long, Philip Oreopoulos, and Lisa Sanbonmatsu, “The Role of Simplification and Information: Evidence from the FAFSA Experiment,” *Quarterly Journal of Economics*, 127 (2012), 1205–1242.

Blackwell, Lisa S., Kali H. Trzesniewski, and Carol S. Dweck, “Implicit Theories of Intelligence Predict Achievement across an Adolescent Transition: A Longitudinal Study and an Intervention,” *Child Development*, 78 (2007), 246–263.

Borghans, Lex, Angela L. Duckworth, James J. Heckman, and Bas ter Weel, “The Economics and Psychology of Personality Traits,” *Journal of Human Resources*, 43 (2008), 972–1059.

Buechel, Berno, Lydia Mechtenberg, and Julia Petersen, “If I Can Do It, So Can You! Peer Effects on Perseverance,” *Journal of Economic Behavior and Organization*, 155 (2018), 301–314.

Cameron, A. Colin, and Douglas L. Miller, “A Practitioner’s Guide to Cluster-Robust Inference,” *Journal of Human Resources*, 50 (2015), 317–372.

Campbell, Frances, Gabriella Conti, James J. Heckman, Seong H. Moon, Rodrigo Pinto, Elizabeth Pungello, and Yi Pan, “Early Childhood Investments Substantially Boost Adult Health,” *Science*, 343 (2014), 1478–1485.

Castillo, Marco, Paul J. Ferraro, Jeffrey L. Jordan, and Ragan Petrie, “The Today and Tomorrow of Kids: Time Preferences and Educational Outcomes of Children,” *Journal of Public Economics*, 95 (2011), 1377–1385.

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

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR,” *Quarterly Journal of Economics*, 125 (2011), 1593–1660.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood,” *American Economic Review*, 104 (2014), 2633–2679.

Dee, Thomas S., and Martin R. West, “The Non-Cognitive Returns to Class Size,” *Educational Evaluation and Policy Analysis*, 33 (2011), 23–46.

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

Dobronyi, Christopher R., Philip Oreopoulos, and Uros Petronijevic, “Goal Setting, Academic Reminders, and College Success: A Large-scale Field Experiment,” NBER Working Paper no. 23738, 2017.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner, “Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences,” *Journal of the European Economic Association*, 9 (2011), 522–550.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly, “Grit: Perseverance and Passion for Long-Term Goals,” *Journal of Personality and Social Psychology*, 92 (2007), 1087–1101.

Duckworth, Angela L., and Patrick D. Quinn, "Development and Validation of the Short Grit Scale (Grit-S)," *Journal of Personality Assessment*, 91 (2009), 166–174.

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

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

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

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

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

Garcia, Jorge L., James J. Heckman, Duncan E. Leaf, and Maria J. Prados, "The Life-Cycle Benefits of an Influential Early Childhood Program," NBER Working Paper no. 22993, 2016.

Gerhards, Leonie, and Christina Gravert, "Because of You I Did Not Give Up—How Peers Affect Perseverance," Working Papers in Economics 659, University of Gothenburg, 2016.

Gneezy, Uri, and Jan Potters, "An Experiment on Risk Taking and Evaluation Periods," *Quarterly Journal of Economics*, 112 (1997), 631–645.

Golsteyn, Bart H. H., Hans Grönqvist, and Lena Lindahl, "Adolescent Time Preferences Predict Lifetime Outcomes," *Economic Journal*, 124 (2013), 739–761.

Good, Catherine, Joshua Aronson, and Michael Inzlicht, "Improving Adolescents' Standardized Test Performance: An Intervention to Reduce the Effects of Stereotype Threat," *Journal of Applied Developmental Psychology*, 24 (2003), 645–662.

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

Hanushek, Eric A., and Steven G. Rivkin, "The Distribution of Teacher Quality and Implications for Policy," *Annual Review of Economics*, 4 (2012), 31–57.

Heckman, James J., John Eric Humphries, and Nicholas S. Mader, "The GED," in *Handbook of the Economics of Education* vol. 3, E. Hanushek, S. Machin and L. Woessmann, eds. (Amsterdam: North-Holland: 2011), 423–484.

Heckman, James J., Seong H. Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz, "The Rate of Return to the HighScope Perry Preschool Program," *Journal of Public Economics*, 94 (2010), 114–128.

Heckman, James J., Rodrigo Pinto, and Peter Savelyev, "Understanding the Mechanisms through which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103 (2013), 2052–2086.

Heckman, James J., Jora Stixrud, and Sergio Urzua, "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24 (2006), 411–482.

Hertwig, Ralph, and Andreas Ortmann, "Experimental Practices in Economics: A Methodological Challenge for Psychologists," *Behavioral and Brain Sciences*, 24 (2001), 383–451.

Hodara, Michelle, "Improving Students' College Math Readiness: A Review of the Evidence on Postsecondary Interventions and Reforms," CAPSEE Working Paper, 2013.

Hopkins, Kenneth D., and G. Bracht, "Ten-Year Stability of Verbal and Nonverbal IQ Scores," *American Education Research Journal*, 12 (1975), 469–477.

Hoxby, Caroline M., and Sonali Murarka, "Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement," NBER Working Paper no. 14852, 2009.

Jackson, C. Kirabo, "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes," *Journal of Political Economy*, 126 (2018), 2072–2107.

Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans, *Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success*, (Paris: Organisation for Economic Co-operation and Development, 2014).

Knudson, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff, “Economic, Neurobiological, and Behavioral Perspectives on Building America’s Future Workforce,” *Proceedings of the National Academy of Sciences*, 103 (2006), 10155–10162.

Kosse, Fabian, Thomas Deckers, Armin Falk, Pia Pinger, and Hannah Schildberg-Hörisch, “The Formation of Prosociality: Causal Evidence on the Role of the Social Environment,” *Journal of Political Economy* (forthcoming).

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff, “The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance,” *American Economic Journal: Economic Policy*, 8 (2016), 183–219.

Maddie, Salvatore R., Michael D. Matthews, Dennis R. Kelly, Brandilynn Villarreal, and Marina White, “The Role of Hardiness and Grit in Predicting Performance and Retention of USMA Cadets,” *Military Psychology*, 24 (2012), 19–28.

Maniadis, Zacharias, Fabio Tufano, and John A. List, “One Swallow Doesn’t Make a Summer: New Evidence on Anchoring Effects,” *American Economic Review*, 104 (2014), 277–290.

Moffit, Terrie E., Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, Hona Lee Harrington, Renate Houts, Richie Poulton, Brent W. Roberts, Stephen Ross, Malcolm R. Sears, W. Murray Thomson, and Avshalom Caspi, “A Gradient of Childhood Self-control Predicts Health, Wealth, and Public Safety,” *Proceedings of the National Academy of Sciences*, 108 (2011), 2693–2698.

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

Oreopoulos, Philip, Richard W. Patterson, Uros Petronijevic, and Nolan G. Pope, “Lack of Study Time Is the Problem, but What Is the Solution? Unsuccessful Attempts to Help Traditional and Online College Students,” NBER Working Paper no. 25036, 2018.

Oreopoulos, Philip, and Uros Petronijevic, “Student Coaching: How Far Can Technology Go?,” *Journal of Human Resources*, 53 (2018), 299–329.

Paunesku, David, Gregory M. Walton, Carissa Romero, Eric N. Smith, David S. Yeager, and Carol S. Dweck, “Mind-set Interventions Are a Scalable Treatment for Academic Underachievement,” *Psychological Science*, 26 (2015), 784–793.

Popova, Anna, David K. Evans, and Violeta Arancibia, “Training Teachers on the Job: What Works and How to Measure It,” World Bank Group, Policy Research Working Paper 7834, 2016.

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

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

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, “Teachers, Schools, and Academic Achievement,” *Econometrica*, 73 (2005), 417–458.

Roberts, Brent W., Nathan R. Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R. Goldberg, “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 (2007), 313–345.

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

Sisk, Victoria F., Alexander P. Burgoyne, Jingze Sun, Jennifer L. Butler, and Brooke N. Macnamara, “To What Extent and Under Which Circumstances

Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses," *Psychological Science*, 29 (2018), 549–571.

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

Sutter, Matthias, Martin G. Kocher, Daniela Glätze-Rützler, and Stefan T. Trautmann, "Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior," *American Economic Review*, 103 (2013), 510–531.

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

Yeager, D. S., P. Hanselman, G. M. Walton, R. Crosnoe, C. Muller, E. Tipton, B. Schneider, and C. Hulleman, et al., "Where and for Whom Can a Brief, Scalable Mindset Intervention Improve Adolescents' Educational Trajectories?," unpublished manuscript, 2018.

Yeager, David S., Rebecca Johnson, Brian J. Spitzer, Kali H. Trzesniewski, Joseph Powers, and Carol S. Dweck, "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 (2014), 867–884.

Yeager, David S., Hae Y. Lee, and Jeremy Jamieson, "How to Improve Adolescent Stress Responses: Insights from an Integration of Implicit Theories of Biopsychosocial Models," *Psychological Science*, 27 (2016), 1078–1091.

Yeager, David S., Carissa Romero, Dave Paunesku, Christopher S. Hulleman, Barbara Schneider, Cintia Hinojosa, Hae Y. Lee, and Joseph O'Brien, et al., "Using Design Thinking to Improve Psychological Interventions: The Case of the Growth Mindset during the Transition to High School," *Journal of Educational Psychology*, 108 (2016), 374–391.

Yeager, David S., and Gregory M. Walton, "Social-Psychological Interventions in Education: They're Not Magic," *Review of Educational Research*, 81 (2011), 267–301.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Result," *Quarterly Journal of Economics*, 134 (2019), 557–598.