Swiss Leading House
Economics of Education • Firm Behaviour • Training Policies

Working Paper No. 129

Class Size in Early Grades, Student Grit and Later School Outcomes

Jana Gross, Simone Balestra and Uschi Backes-Gellner

Universität Zürich
IBW – Institut für Betriebswirtschaftslehre

UNIVERSITÄT BERN

Working Paper No. 129

Class Size in Early Grades, Student Grit and Later School Outcomes

Jana Gross, Simone Balestra and Uschi Backes-Gellner

September 2018 (first version: June 2017)

This paper was previously circulated under the title "Does Class Size Affect Student ‘Grit’? Evidence from a Randomised Experiment in Early Grades" (2017).

Please cite as:
"Class Size in Early Grades, Student Grit and Later School Outcomes." Swiss Leading House "Economics of Education" Working Paper No. 129, 2018 (first version 2017). By Jana Gross, Simone Balestra and Uschi Backes-Gellner.

The Swiss Leading House on Economics of Education, Firm Behavior and Training Policies is a Research Program of the Swiss State Secretariat for Education, Research, and Innovation (SERI).

www.economics-of-education.ch

Class Size in Early Grades, Student Grit and Later School Outcomes*

Jana Gross[†], Simone Balestra[‡] and Uschi Backes-Gellner[§]

Abstract

The increasing recognition of non-cognitive skills in economics has led many researchers to investigate how educational practices enhance these skills. In this paper, we focus on the non-cognitive skill known as 'grit', and we study the causal relation between class size and grit. Using data from follow-up surveys of Project STAR, we show that fourth-grade pupils who experienced small classes during early grades are 0.12 standard deviations higher in grit than their peers in regular classes. We also show that grit matters, because nearly half of the effect of smaller classes on test scores operates through grit. The effects of grit are far-reaching: students with higher grit have better grades at the end of compulsory schooling, are more likely to graduate from high school on time and are more likely to take a college entrance exam.

Keywords: class size, grit, non-cognitive skills.

JEL Classification: I21, J13, J24.

[*] We are grateful to Thomas Dohmen, Simon Janssen, Rafael Lalive, Edward Lazear, Carmit Segal, and Ludger Woessmann for their constructive comments. This study is partly funded by the Swiss State Secretariat for Education, Research and Innovation through its Leading House on the Economics of Education, Firm Behaviour and Training Policies.

[†] ETH Zurich, Switzerland. Telephone: +41 (0)44 633 8494; e-mail: jgross@ethz.ch. Address: Weinbergstrasse 56/58, CH-8006 Zurich, Switzerland. Jana Gross is the corresponding author.

[‡] University of St. Gallen, Switzerland. Telephone: +41 (0)71 224 2318; e-mail: simone.balestra@unisg.ch. Address: Rosenbergstrasse 51, CH-9000 St. Gallen, Switzerland.

[§] University of Zurich, Switzerland. Telephone: +41 (0)44 634 4281; e-mail: backes-gellner@business.uzh.ch. Address: Plattenstrasse 14, CH-8032 Zurich, Switzerland.

1 Introduction

Over the last decade, non-cognitive skills have emerged as one of the most important determinants of educational attainment (Heckman and Rubinstein, 2001; Segal, 2013) and labour market success (Heckman and Kautz, 2012; Heckman, Stixrund, and Urzua, 2006). Amongst all non-cognitive skills, the characteristic 'grit' has been shown as a strong predictors of both academic and job performance (Duckworth, Kirby, Tsukayama, Berstein, and Ericsson, 2010; Robertson-Kraft and Duckworth, 2014). 'Grit' entails two dimensions: consistency of interest and perseverance of effort (Duckworth, Peterson, Matthews, and Kelly, 2007). Individuals with high grit stay on track despite major failures or setbacks, and this ability to sustain persistence in effort and interest in long-term tasks positively impacts lifetime outcomes in various ways (Eskreis-Winkler, Duckworth, Shulman, and Beal, 2014). Importantly, researchers find no correlation between grit and cognitive skills, suggesting that the higher performance of gritty individuals is attributable to effort and interest rather than to cognitive ability (Culin, Tsukayama, and Duckworth, 2014).

Although research investigating the impact of grit on educational and labour market outcomes is well-developed, much less is known about whether modifying or shaping this skill is possible, and – if so – what educational practices allow individuals to do so (Duckworth, 2016; Segal, 2008). In this paper, we investigate the causal effect of class size in early grades on the development of grit. We focus on class size in early grades for two reasons. First, interventions in early childhood are most effective and show a consistent impact on adult life (Doyle, Harmon, Heckman, and Tremblay, 2009; Heckman, Pinto, and Savelyev, 2013). Second, reducing class size is a widespread educational practice aimed at improving student outcomes (Antecol, Eren, and Ozbeklik, 2016; Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011).

However, the actual impact of class size on student achievement remains debated, suggesting that test scores are likely not the only outcome on which we should focus (Schanzenbach, 2006). In smaller classes, students benefit from closer guidance and better mentoring, both of which increase not only student engagement (Dee, 2007) but also certain non-cognitive skills (Heckman and Mosso, 2014). For example, Dee and West (2011) use the National Education Longitudinal Study of 1988 to show that smaller eighth-grade classes are associated with improvements in several measures of school engagement. No paper so far analysed the impact on grit of smaller classes during early grades, leaving unanswered the question on whether such educational practices improve grit. To fill this research gap, we use data from a large-scale randomised experiment to study the causal relation between class size and grit.

To identify causal effects, we use data from 'Project STAR and Beyond', a 1990 follow-up survey of the Tennessee Student-Teacher Achievement Ratio experiment (Project STAR). Project STAR is a large-scale experiment that randomly assigned students and teachers to classes of different size. After the experiment, annual follow-up studies continued collecting students' information to investigate the long-term effects of class size. Our analysis centres on the outcomes collected in the fourth grade study on student behaviour, the Student Participation Questionnaire (SPQ). The SPQ provides personal attributes necessary for building our outcome variable as closely as possible to the original grit scale developed by Duckworth et al. (2007).

We first provide descriptive evidence suggesting that students assigned to smaller classes during Project STAR score higher on the grit scale than their counterparts in the control group. Second, to further investigate this relation, we estimate several econometric specifications to study the causal effect of class size on grit. Our results show that, on average, pupils in smaller classes are about 0.12 standard deviations higher in grit than their peers in regular classes. Whilst controlling for student and fourth grade characteristics barely affects the effect size, it improves the precision of the coefficient of interest. The effect size is not only statistically significant but also important in magnitude: for example, the class size effect on pupils' fourth grade test scores is only a third of the grit effect in the same year.[1]

To deepen scholarly understanding of the effect of class size on grit, we perform sub-sample analyses according to ethnicity and socio-economic status. Our results show that student's grit in smaller classes positively increases for both non-white students and students with a lower socio-economic status. These results are supported by Lazear (2001) who suggests that, for a given class size reduction, the increase in educational outcome (broadly defined) is larger for disadvantaged students. Following this theoretical intuition, Jepsen and Rivkin (2009) show empirically that class size effects are indeed more important for disadvantaged students. Our results show that in terms of grit, non-white pupils also benefit more from smaller classes than the average student. Moreover, as previous work shows that boys tend to be more disruptive and more likely to lose focus during instructional time, we conduct separate analyses for boys and girls (Bertrand, Pan, and Kamenica, 2013). Our results support this hypothesis, showing that boys develop grit more in smaller classes than girls.

To shed light on the importance of grit in determining student achievement and later outcomes, we present a causal mediation analysis (Imai, Keele, and Yamamoto, 2010). The purpose of this analysis is to understand to what extent the class size effect on school outcomes operates through grit. The results are also useful for better understanding the underlying mechanism between class size, grit and school outcomes. Our analysis suggests that the well-documented class size effect on test scores and later school outcomes is driven mainly by grit. This pathway is also suggested by Cunha and Heckman (2008), who show that non-cognitive skills promote the formation of cognitive skills, not the other way about.

We perform several robustness checks. First, we test the sensitivity of our results to alternative specifications of the dependent variable by using an extended grit scale. Second, we deal with (non-

[1] Estimation taken from Schanzenbach (2006), who also uses data from Project STAR and Beyond.

random) attrition by computing bounds of the treatment effect and by imputing grit for students who did not participate in the SPQ. Third, we control for teacher rating standards by adding teacher fixed effects to our models. Overall, all the robustness checks indicate that neither the specification of the dependent variable nor attrition patterns threaten the validity of our analyses.

The remainder of this study proceeds as follows. Section 2 gives an overview of the literature on both grit and class size effects on non-cognitive skills. Section 3 presents the data set, the econometric approach, and some descriptive statistics. Section 4 presents and discusses the results and section, including a series of robustness checks and sensitivity analyses. Section 5 analyses the impact of grit on later school outcomes and how the class size effect operates through grit in determining such outcomes. Finally, section 6 concludes.

2 Literature

A wide body of evidence recognises the increasing importance of non-cognitive skills (Chamorro-Premuzic and Furnham, 2003; Duckworth and Seligman, 2006; Finn and Achilles, 1999; Heckman, Stixrund, and Urzua, 2006; McCrae and John, 1992). Non-cognitive skills are strong predictors of educational, academic and other lifetime outcomes (Heckman and Kautz, 2012; Heckman and Rubinstein, 2001; Kautz, Heckman, Diris, ter Weel, and Borghans, 2014; Segal, 2013). Moreover, Borghans, Duckworth, Heckman, and Weel (2008) show that non-cognitive skills become equally relevant to cognitive skills, if not more so, in predicting later academic performance.

In this strand of the literature, the discussion of the non-cognitive skill grit is relatively new, because grit itself is a concept that entered the academic discussion only in 2007, with Duckworth et al. (2007). Grit is the ability to maintain effort and interest over time, even in the face of major setbacks. The two main attributes of grit constitute persistence of effort and consistency of interest. Gritty individuals show high stamina in both diligent work and passion when pursuing long-term goals. Whereas others perceive failure as a signal to change course, gritty individuals stay on course (Duckworth et al., 2007).

Gritty individuals gain higher achievement in both their working environment and private life (Culin, Tsukayama, and Duckworth, 2014; Duckworth et al., 2007). In line with Duckworth et al. (2007), Duckworth et al. (2010), show that gritty individuals are higher educated than their less gritty peers. Moreover, as gritty individuals are more likely to graduate from high school on time (Eskreis-Winkler et al., 2014), the importance of grit clearly starts in school, not afterwards. However, this effect is not an ability effect, because grit is not correlated with IQ (Duckworth et al., 2007; Duckworth and Quinn, 2009).

Alan, Boneva, and Ertac (2015) present evidence that grit might be malleable in early childhood. They evaluate a large-scale randomized educational intervention that aims to improve children's time preferences and grit. They show that children that receive the treatment perform better than the control group in both experimental tasks and school outcomes. Although Alan, Boneva, and Ertac (2015) do not measure grit explicitly, they suggest that educational activities implemented in a natural classroom environment can affect children's behaviours related to goal-setting and perseverance. The question thus remains whether and, if so, how grit can be changed by more widespread educational policies such as reductions in class size.

Two studies are close to the present investigation. First, Schanzenbach (2006) provides an overview of the academic literature using the Project STAR experiments. She also analyses the class size effect on the following personality traits: self-concept, motivation, and listening. These three traits come from the Self-Concept, ad Motivational Inventory (SCAMIN), which was given at the end of each year during the STAR experiment. By contrast, we use measures of non-cognitive skills elicited by the SPQ after Project STAR, in the fourth-grade follow-up survey. The advantage of using the SPQ is that it measures a broader battery of non-cognitive skills than the SCAMIN, allowing us to build a grit scale. Second, in their working paper[2] version, Dee and West (2011) analyse class size effects on three aggregated measures of non-cognitive skills: initiative, effort, and non-participatory behaviour. While these three composite scores are also taken from the SPQ, we do not use them for our grit scale. We instead use single items from the SPQ directly.

Many behavioural traits fall under the category of non-cognitive skills. However, despite overlapping key determinants with other non-cognitive skills, grit is distinctive. First, Duckworth and Gross (2014) argue that grit entails the ability to work assiduously toward a single long-term goal despite setbacks, whereas self-control is the ability to regulate behaviour when pursuing a goal despite attractive alternatives. Second, despite overlapping areas of achievement for both grit and conscientiousness, grit differs in its emphasis on long-term stamina (Duckworth et al., 2007). Similarly, whilst both grit and hardiness provide attributes of the motivation to work hard despite stress or failure, hardiness does not necessarily entail consistency of interest amongst long-term goals (Maddi, 2006).

The main challenge for out project lies in formalising the concept of grit because the first 12-item grit scale was introduced in 2007 by Duckworth et al. (2007). Duckworth and Quinn (2009) improve on the initial grit scale by showing that their eight-item scale is as meaningful as the 12-item scale. Moreover, Duckworth also developed a different eight-item scale[3] specifically for children, a scale on which we base our grit measures. However, as our data were collected before the grit scale was fully developed, we are not able to perfectly replicate the original grit scale. Nevertheless, we build our scale as closely as possible to Duckworth's scale for children, and we test the relevance of our scale through robustness checks.

[2] See Dee and West (2008).

[3] Publicly available on her website http://angeladuckworth.com/research/.

3 Data and Empirical Strategy

3.1 Data, Measures and Descriptive Statistics

Although the grit scale was both formalised and developed within the last ten years, we use data that were collected before Duckworth et al. (2007) defined grit as a characteristic. We use data from follow-up surveys of the Tennessee Student-Teacher Achievement Ratio experiment (Project STAR), a large-scale, randomised class-size experiment that took place between 1985 and 1989. The 1990 follow-up study provides student behavioural attributes that we use to develop our outcome variable, grit. Project STAR originally involved 11,600 students from kindergarten through third grade.[4] It was commissioned by the Tennessee state legislature and implemented by a consortium of Tennessee universities and the Tennessee State Department of Education.

The experiment randomly assigned kindergarten pupils to small classes (target enrolment between 13 and 17 students) or regular classes (target enrolment between 22 and 26 students, either with or without a teacher's aide). The class-type assignments of pupils and teachers were maintained through the third grade. Children and teachers entering the study after kindergarten were also randomly assigned to one of the treatments. Although the data cover only one state and one cohort, the experiment included a heterogeneous set of schools from across Tennessee, including large and small, urban and rural, and wealthy and poor districts. Consequently, the schools included in the data represent most of the educational conditions that exist in the United States (Krueger and Whitmore, 2001).

From fourth grade on, all students returned to a regular class size. After the Project STAR experiment, annual follow-up studies continued collecting students' information to investigate the long-term effects of having smaller classes at an early age. Our analysis centres on the outcomes collected in the fourth grade follow-up study on student behaviour, the 'Student Participation Questionnaire' (SPQ).[5] At the end of fourth grade (summer 1990) teachers had to rate about ten randomly selected students who participated in Project STAR on 31 items, 29 of which cluster into the following categories: effort, non-participatory behaviour, initiative taking and valuing school outcomes. For each item, teachers had to rate the occurrence of a specific behaviour from 'never' to 'always' on a five-point Likert scale.[6] Of the initial 11,600 pupils, roughly 7,300 attended fourth grade and, of those 7,300, 2,200 were randomly selected for participating in the SPQ.

To build our outcome variable, we follow Duckworth's eight-item scale for children as closely as possible. Two elements define a grit scale: consistency of interest and perseverance of effort. Consistency of interest means, amongst other things, that an individual is able to stay focused on

[4] For detailed information about the experiment, see Finn and Achilles (1990); Folger and Breda (1989) and Word, Johnston, Bain, Fulton, Zaharias, Achilles, Lintz, Folger, and Breda (1990).

[5] The SPQ was also collected in eighth grade but – unfortunately – in a different form. Specifically, the eighth-grade SPQ had fewer items than the fourth grade one, preventing us from constructing a comparable grit scale in the eighth grade.

[6] Appendix Table A.2 shows the full questionnaire.

a given goal. Similarly, perseverance of effort means, amongst other things, that an individual is persistent when confronted with a given task. From all items on the SPQ, we choose only those that match Duckworth's grit scale. Given that the SPQ was collected in 1990 and that the concept of grit was developed only in the 2000s, we are able to replicate only five of the eight items on Duckworth's grit scale for children.

Appendix Table A.1 shows a detailed overview comparing our grit five-item scale to Duckworth's eight-item grit scale. The items 'student doesn't take initiative, must be helped to get started and kept going on work' and 'student does more than assigned work' indicate an individual's ability to remain focused on a given goal without becoming distracted, thus representing the attribute *consistency of interest*. Similarly, the following three items represent the attribute *perseverance of effort*: 'student gets discouraged and stops trying when encounters an obstacle', 'student is persistent when confronted with difficult problems', and 'student tries to finish difficult assignments'. Factor analysis supports the validity of our concept of grit by showing that all five elements belong to the same underlying concept, with $\alpha = 0.87$, well beyond the commonly accepted threshold of 0.7. This analysis suggests that the combination of the five elements can identify the latent variable 'grit'. To construct the grit scale, we sum all five elements and standardise the composite scale, as is common in this strand of the literature (Alan, Boneva, and Ertac, 2015). We assess the relevance of our grit scale in section 5, where we show that grit not only explains most of the increase in fourth grade test scores but is also highly correlated with long-term achievement at the end of high school.

The main difference between Duckworth's grit scale that from the SPQ is who completes the questionnaire. While Duckworth's scale is a self-reported questionnaire, the SPQ is completed by teachers. This difference has both advantages and disadvantages. Self-reported questionnaires are problematic because individuals tend to answer in a 'socially desirable' way (McDonald, 2008; Paulhus, 1991). Similarly, individuals also are prone to 'extreme responding', which might call into question the validity of the self-report measure (Gerhards and Gravert, 2015; Paulhus and Vazire, 2007). Finally, teachers' reports are likely to produce a better relative measure of grit than pupils themselves, because teachers observe the behaviour of the entire class rather than just a sub-group of colleagues or friends.

Relying on teachers' evaluations has also some challenges, because teachers might be biased as well (Elder, 2010). For example, the reports provided by teachers teaching smaller classes might systematically differ from the reports provided by teachers teaching regular classes, for reasons other than differences in students' actual behaviour. As students in smaller classes are more likely to obtain more attention from their teachers, this might result in the teachers being more likely to notice students' actions and behaviours.

This bias could exist if a pupil who was assigned to a small class during the STAR experiment receives the same teacher in fourth grade. The data shows that for our sample of 2,188 children 49

Table 1: Descriptive Statistics

	Small (1)	Regular (2)	Difference (1) – (2) (3)
Grit scale (z-score)	0.103	–0.055	0.16***
Average class size during STAR	15.46	22.75	–7.29***
Student characteristics:			
Boy	0.486	0.507	–0.02
Free-lunch eligible	0.510	0.489	0.02
Non-white	0.251	0.222	0.03
Fourth grade characteristics:			
Male teacher	0.058	0.057	0.00
White teacher	0.826	0.847	–0.02
Share of non-white classmates	0.260	0.225	0.03
Observations	726	1,462	2,188

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Ordinary least squares models with robust standard errors clustered at the classroom level. Regressions in column 3 include school-by-entry-wave fixed effects.
Project STAR and Beyond database.

have the same teacher in fourth grade as in one of the previous grades. Of those 49 children, only four belong to the treatment group, corresponding to 0.5 per cent of the entire treated population. It is thus unlikely that the measures and results are either biased or driven by these four children, and excluding them from the sample has no impact on our results. In sum, it appears safe to assume that the treatment during the STAR experiment does not affect how correct teachers' assessments are. However, the possibility of teacher bias is still present and we attempt to tackle this issue in our robustness checks by including teacher fixed effects in the regressions.

Table 1 presents descriptive statistics for our sample of 2,188 pupils, divided into treatment (column 1) and control (column 2) groups. Column 3 shows the difference between the first two columns and shows whether that difference is statistically significant. Three elements of Table 1 are relevant to our research design. First, pupils assigned to smaller classes were actually put in smaller classes. The average difference between small and regular class size, about seven pupils, is highly significant. For each pupil, we calculate his or her class size as the average number of pupils in a class throughout Project STAR. Second, descriptive evidence already suggests that pupils assigned to smaller classes tend to score higher in grit than their counterparts in regular classes. The difference in grit scale is highly significant, with an effect size of roughly 0.16 standard deviations. Third, the treatment and control groups are balanced according to both student and teacher characteristics, suggesting that no observable characteristic influenced the assignment of treatments.

To further examine the relation between class size and grit, Figure 1 depicts the distribution of grit for the treatment group and control group, respectively. The figure clearly shows that pupils assigned to smaller classes have higher grit than their counterparts in the control group. Overall, both

Figure 1: KERNEL DENSITY OF GRIT BY TREATMENT GROUP

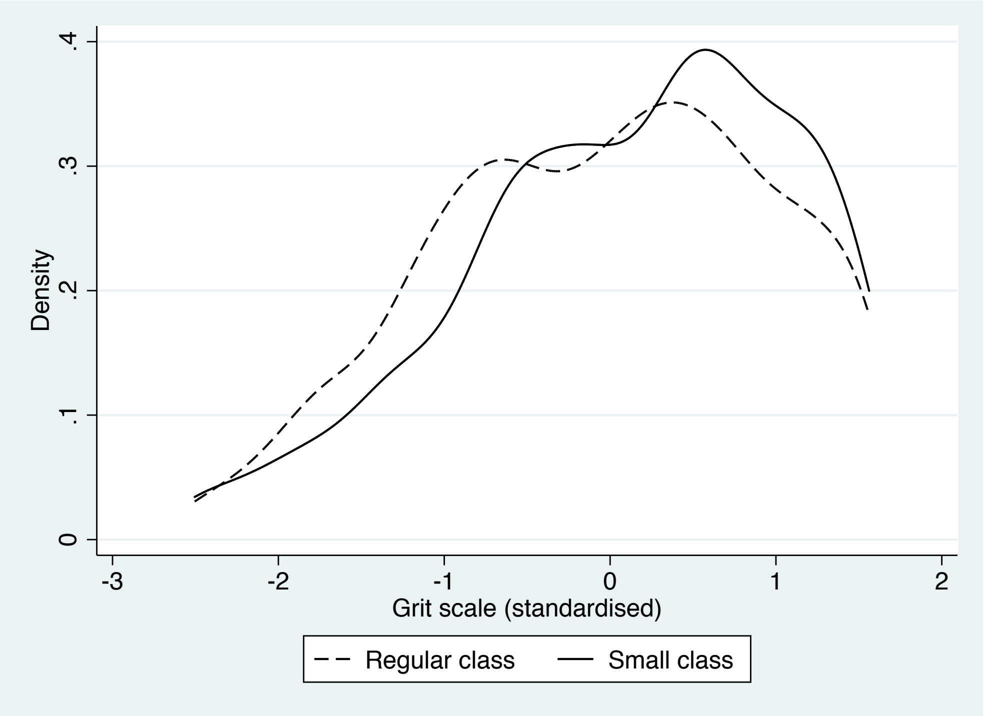


Table 1 and Figure 1 suggest that smaller classes tend to produce grittier pupils. However, we need to demonstrate that this result is not due to selective attrition or other drivers, such as achievement level. We deal with these issues in sections 4 and 5.

3.2 Econometric Models

To investigate whether class size has a causal effect on how gritty a pupil is, we start by estimating the following regression equation:

$$Grit_{iwcs} = \alpha + \beta \cdot Small_{wcs} + X'_{iwcs}\gamma + Z'_{wcs}\delta + \zeta_{ws} + \varepsilon_{iwcs} \tag{1}$$

where $Grit_{iwcs}$ is the forth-grade standardised grit scale of pupil i that entered Project STAR in wave w and was assigned to class c at school s. $Small_{wcs}$ is a dummy variable indicating whether the student was assigned to a small class upon entering Project STAR. X_{iwcs} is a vector of student characteristics, including dummy variables for gender, ethnicity and number of years of free-lunch eligibility. Similarly, Z_{wcs} is a vector of fourth-grade characteristics, including teacher gender, teacher ethnicity and the share of white classmates. Given that the randomisation was done at entry *within* schools, we also include school-by-entry-wave fixed effects (ζ_{ws}). Adding these fixed effects ensures the independence between treatment assignment and other variables.

We estimate equation 1 by ordinary least squares (OLS). Given that $Small_{csw}$ is randomly assigned, β provides an unbiased estimate of the intention-to-treat effect of class size on grit. However, we are interested in the average treatment effect of the treated of class size on grit. As Krueger (1999) and others report, pupils who were assigned to small classes had varying numbers of classmates resulting from mobility and enrolment differences across schools. Likewise, pupils in the regular classes had variable class sizes. We thus consider a structural model that takes actual class size into account. Specifically, we estimate the following system of equations by two-stage-least-squares (TSLS):

$$\begin{aligned} Grit_{iwcs} &= \pi_0 + \pi_1 \cdot CS_{wcs} + X'_{iwcs}\pi_2 + Z'_{wcs}\pi_3 + \zeta_{ws} + e_{iwcs} \\ CS_{wcs} &= \iota_0 + \iota_1 \cdot Small_{wcs} + X'_{iwcs}\iota_2 + Z'_{wcs}\iota_3 + \eta_{ws} + u_{iwcs} \end{aligned} \tag{2}$$

where the variable CS_{wcs} is the average class size pupils had during Project STAR. Treatment assignment constitutes an ideal instrument for class size, not only because the assignment is random but also because its effect on grit runs entirely through class size. In this set-up we use only variation in class size due to initial assignment to a regular or small class to provide variation in actual class size in the grit equation. The coefficient of interest in equation 2 is π_1, which represents the local average treatment effect of a one-pupil increase in class size on the outcome grit.

4 Results

This section comprises three parts: the first part shows the causal relation between class size and the non-cognitive skill, grit. The second part covers the heterogeneous effects of class size, and the third part provides robustness checks.

4.1 The Causal Effect of Class Size on Grit

Table 2 shows the causal effects of class size on grit and is divided into two parts. The first part (columns 1 to 3) presents OLS estimates with different specifications. The first specification (column 1) controls for school-by-entry-wave fixed effects. The second specification (column 2) also controls for student characteristics, including gender, free-lunch eligibility and ethnicity. Column 3 shows our preferred estimate, which additionally controls for fourth grade characteristics, including teacher's gender, teacher's ethnicity and the share of non-white classmates. The second part (columns 4 and 5) shows the two-stage least squares estimates, using the assignment to small class as an instrument for the average class size during Project STAR.

Column 1 shows highly significant ($p < 0.01$) differences between small and regular classes. On average, students in smaller classes were 0.142 standard deviations higher in grit than their peers in regular classes. When we control for student and fourth grade characteristics (columns 2 and 3), the effect slightly decreases but remains highly significant. According to our preferred estimate (column 3), students assigned to smaller classes during Project STAR are 0.123 standard deviations higher in grit in fourth grade, compared to students in regular classes.[7] This effect is both statistically significant

[7] Using the factor loadings extracted from the factor analysis as dependent variable does not change the results. Results are

Table 2: EFFECT OF CLASS SIZE ON GRIT SCALE, OLS AND TSLS

	Ordinary Least Squares			Two Stage Least Squares	
	Grit scale (z-score) (1)	Grit scale (z-score) (2)	Grit scale (z-score) (3)	Average class size in STAR (4)	Grit scale (z-score) (5)
Assigned to small	0.142***	0.122***	0.123***	–7.244***	
	(0.047)	(0.046)	(0.046)	(0.100)	
Average class size in STAR					–0.017***
					(0.006)
School-by-entry-wave FE	YES	YES	YES	YES	YES
Student characteristics	NO	YES	YES	YES	YES
Fourth grade characteristics	NO	NO	YES	YES	YES
R^2	0.112	0.189	0.189	0.850	0.190
Observations	2,188	2,188	2,188	2,188	2,188

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors clustered at the classroom level are in parentheses. Student characteristics include gender, years of free-lunch eligibility and ethnicity. Fourth grade characteristics include teacher's gender, teacher's ethnicity and the share of non-white classmates. Project STAR and Beyond database.

and relevant in magnitude. The grit effect is twice as large as the test score effect in the same year (Schanzenbach, 2006) and also slightly larger than the effect of class size on other non-cognitive skills found in the related literature. For example, Dee and West (2011) use the National Education Longitudinal Study to study the impact of class size on student engagement in school, estimating effect sizes ranging between 0.05 and 0.09 standard deviations. We find that the impact of class size on grit is larger than the effects that Dee and West found for engagement, a finding that underlines the importance of grit relative to other non-cognitive skills.

The first stage of the TSLS model (column 4) shows that the number of students in small classes is about seven students fewer than in regular classes, confirming the descriptive statistics of the average class size in Table 1. The estimates in column 5 corroborate the findings in columns 1 to 3. When the average class size increases by one student, students' grit decreases by 0.017 standard deviations ($p < 0.01$). If the class size increases by seven students, as in Project STAR, grit decreases by 0.119 standard deviations. This estimate is very close to the one we obtain with the OLS models. This closeness is due to the high implementation fidelity of the treatments in Project STAR, as also shown by the R-squared of the first stage in column 4.

To understand how each grit attribute is affected by class size reduction, we estimate the relation between class size reduction and the single items on the grit scale. Appendix Table A.3, which provides the results, is divided into two panels. All specifications are controlled for school-by-entry-wave fixed effects, pupil and fourth grade characteristics. Panel A shows the OLS estimates, and panel B presents the TSLS estimates. In panel A, smaller classes have a positive impact on four of the five items. Pupils in a small class try harder to finish difficult assignments by 0.079 standard deviations ($p < 0.10$).

available upon request.

When pupils face difficult problems (e.g. in take-home assignments), smaller classes increase pupils' persistence by 0.144 standard deviations ($p < 0.01$). In addition, pupils' initiative increases by 0.101 standard deviations in small classes ($p < 0.05$). Students' engagement in doing more than merely the assigned work increases by 0.112 standard deviations ($p < 0.05$), showing the increase in effort spent on both in-class work and homework. When we use the TSLS approach (panel B), the estimates are in line with the findings in panel A, and the significance level remains the same relative to OLS.

4.2 Heterogeneous Effects

To further examine the relation between smaller classes and grit, we regress class size assignment on grit for different sub-groups, according to gender, socio-economic status and ethnicity. Table 3 presents the results for gender (columns 1 and 2), free-lunch eligibility (column 3) and ethnicity (column 4). As is common in the literature, we use free-lunch eligibility as a proxy for disadvantaged socio-economic background. We use two panels: panel A shows the OLS results, and panel B presents the TSLS estimates. Both specifications control for student characteristics, school-by-entry-wave fixed effects and fourth grade characteristics.

Table 3: SUB-SAMPLE ANALYSIS, OLS AND TSLS

	Grit scale (z-score)			
	Boys	Girls	Free-lunch	Non-white
	(1)	(2)	(3)	(4)
A. *OLS*				
Assigned to small	0.179***	0.092*	0.141**	0.284***
	(0.069)	(0.054)	(0.071)	(0.111)
B. *TSLS*				
Average class size in STAR	–0.025***	–0.013*	–0.020**	–0.042***
	(0.009)	(0.007)	(0.009)	(0.015)
School-by-entry-wave FE	YES	YES	YES	YES
Student characteristics	YES	YES	YES	YES
Fourth grade characteristics	YES	YES	YES	YES
Observations	1,094	1,094	1,085	507

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors clustered at the classroom level are in parentheses. Student characteristics include gender, years of free-lunch eligibility and ethnicity. Fourth grade characteristics include teacher's gender, teacher's ethnicity and the share of non-white classmates. Project STAR and Beyond database.

As expected, boys profit more than girls in smaller classes. Boys experience a significant increase in grit of 0.179 standard deviations compared to 0.092 standard deviations for girls. Although this difference is only marginally significant ($p = 0.10$), it is in line with the literature showing that despite more regular class attendance, boys are more likely to suffer from behavioural problems (Finn and Rock, 1997). A higher pupil-teacher interaction allows teachers to better control interpersonal issues in smaller classes, and they can spend more time on in-class material than on class management (Blatchford, Bassett, and Brown, 2011; Finn and Achilles, 1999). For example, external distractions are reduced, thereby allowing boys in particular to focus on in-class work (Finn, Pannozzo, and Achilles, 2003). As a result, boys might develop persistence in effort, resulting in a significant higher increase in their being gritty than for girls.

In terms of ethnicity, Project STAR schools had a larger minority fraction than did other schools in Tennessee. In addition, most minority students in Project STAR were black[8] (Krueger and Whitmore, 2001). Therefore, a large fraction of non-white students are free-lunch eligible. Not surprisingly, non-white students in smaller classes are 0.284 standard deviations grittier than their peers in regular classes ($p < 0.01$), and low-income students profit by an increase of 0.141 standard deviations in grit in smaller classes ($p < 0.05$). While the difference in effect size between white and non-white pupils is significant ($p = 0.01$), the difference between free-lunch eligible pupils and the other pupils is not.

Our results are in line with previous findings that non-white students from lower socio-economic backgrounds profit the most from educational reforms (Finn, Fulton, Zaharias, and Nye, 1989; Krueger and Whitmore, 2001). Fredriksson, Öckert, and Oosterbeek (2016) provide an explanation for this finding, which might fit also our case. Fredriksson, Öckert, and Oosterbeek (2016) show that parents react to class size and conclude that an increase in class size causes high income parents to help their children. They also show that only low-income children find their teachers harder to follow when taught in a larger class, which might explain why we find larger effects for non-white and free lunch eligible children.

4.3 Robustness Checks

To check the robustness of our estimates and test the sensitivity of our results, we perform several robustness checks. First, we test the sensitivity of our results to alternative specifications of the dependent variable. We extend our grit scale to elements that partially represent grit, but that could also overlap with other personality traits, time preferences or parental effect. We then perform our main analysis using this extended grit scale, which includes the following six new elements: whether the pupil completes homework on time, whether the pupil completes in-class work, whether the pupil tries to do work well, whether the pupil approaches assignments with effort, whether the pupil discusses the subject matter with the teacher outside of class and whether the pupil comes late to class (reversed).

Second, we might be concerned that teacher rating standards bias the results. Different teachers have may have different rating standards and if this residual heterogeneity correlates with the treatment variable we could have omitted variables bias. To examine this possibility, we estimate our main models including (fourth grade) teacher fixed-effects. The teacher fixed effects control for heterogeneous teacher rating standards.

[8] Black students constitute more than 98 per cent of the non-white sub-sample.

A third potential concern is attrition, which is a threat to both internal and external validity. In Project STAR and its follow-up studies, attrition is relatively high (Hanushek, 1999). If outcome data are missing for some pupils, one concern is that the potential outcomes for those observed in the treatment group will differ from the potential outcomes for those observed in the control group. Even if attrition is not different across treatments, departures could yield analytic samples that vary significantly from the population of interest, limiting the external validity of the estimated effects.

In the data set for fourth grade, we have two types of attrition. First, for some students we do not have data on fourth grade, something that could happen as a result of either mobility or grade retention. From the initial sample of roughly 11,600 pupils, we have data for only 63 per cent (7,324 pupils). Second, of the fourth grade population, 10 pupils per class were randomly selected for the SPQ survey. Thus we have complete information for about 2,200 pupils, corresponding to 30 per cent of the total fourth grade population. Although no data exists on students who did not participate in either fourth grade or the SPQ before the collection of grit information, we can look for evidence of non-random attrition by examining differences in attrition rates across treatments and in observable characteristics across treatments.

To do so, we first regress an indicator of whether a pupil participated in fourth grade on the treatment dummy, and an interaction between the treatment dummy and the pupils' characteristics. Then we perform the same analysis for the pupils who participated in the SPQ, conditional on our having fourth grade data for them.

Appendix Table A.4 presents our attrition analysis. For fourth grade data, we find that attrition rates are balanced between the treatment and control groups. This finding suggests that pupils in smaller classes were neither more nor less likely than pupils in regular classes to have progressed to fourth grade. However, boys in the control group appear slightly more likely to have left the sample by fourth grade ($p < 0.10$). This attrition pattern is not new, as it results from boys (especially those in regular classes) being more likely to repeat grades (Konstantopoulos, 2008). For this reason, we observe slightly fewer boys in the control group than in the treatment group.

As for SPQ participation (conditional on our having fourth grade data), we find no particular pattern of non-random attrition for pupil gender, free-lunch eligibility or ethnicity. However, we observe that pupils assigned to smaller classes during Project STAR were significantly more likely to be selected for the SPQ. This difference in attrition patterns is worrisome, because it could create a situation in which the treatment and control groups differ with respect to unobserved characteristics. If so, the internal validity of our results could be undermined. To deal with this potential problem, we perform two robustness checks. First, we compute the bounds of the treatment effect based on Lee's trimming approach (Lee, 2009). This approach applies to research designs such as ours, in which the regressor of interest is assumed to be exogenous and the dependent variable is missing in a potentially non-random manner. Lee's approach yields the tightest bounds on average treatment effects consistent with the observed data, and these bounds can be further tightened if we include baseline characteristics.[9]

The second strategy for dealing with attrition is to impute grit scales for pupils who did not participate in the SPQ. We adopt a worst-case scenario and predict the grit of pupils who left the control group as if they had been assigned to smaller classes. Conversely, we predict grit scales for pupils who left the treatment group as if they had received no treatment. This imputation technique should lead to both an increase in average grit for the control group and a decrease in average grit for the treatment groups.

Table 4 presents the results of our robustness checks. It shows that using the extended grit scale does not change the main result of the paper (column 2). Although including additional elements yields slightly less precise coefficients and moderately smaller effect sizes, the effect on grit of reducing class size is always positive and significant – a finding that likely indicates that the additional items are adding noise to the grit measure.

Column 2 shows how the effect of smaller classes changes when we include the teacher fixed effects. Both the intention-to-treat effect and the second stage effect are smaller in magnitude, as Dee and West (2008) also find for their outcomes. However, in contrast to Dee and West (2008), the impact of smaller classes on grit remains significant at the five per cent level. That finding indicates that even by keeping teachers' rating standards constant, the effect of smaller classes on grit remains.

In column 3, following Lee's trimming approach, we compute the treatment effect bounds. This procedure estimates a lower bound and an upper bound of the true effect of being assigned to smaller classes. The bounds are computed through using all the students participating in fourth grade (7,324), not only those selected for the SPQ. We find that the lower bound is positive but not statistically different from zero. The upper bound is positive and significant with a magnitude of 0.3 standard deviations. We conclude that, under reasonable assumptions and using the information available in the data, the effect of smaller classes on grit is zero at worst.

Finally, in column 4 we impute the missing grit scales using the worst-case scenario described previously. The idea is to assume that those who did not participate in the SPQ were the grittier students of the control group and the least gritty pupils of the treatment group. Doing so reduces the causal effect of smaller classes on grit scale in terms of magnitude (by about two-thirds) but not in terms of significance. This finding suggests that even in a worst-case scenario the effect of smaller classes on grit remains positive and significant.

One final concern could be that small classes have significant murky effects on all kinds of student skills, and our measure for grit is just picking up some of these effects. If this was the case, our grit scale would be no better that a random mix of items measured by the SPQ. To address this concern, we

[9] The key assumption is a monotonicity restriction on how the assignment to treatment affects selection, a restriction that is implicitly assumed in standard formulations of the sample selection problem (Lee, 2009).

Table 4: Robustness Checks

	Extended grit scale (z-scores) (1)	Grit scale (z-scores) (2)	Grit scale (z-scores) (3)	Imputed grit (z-scores) (4)
A. *OLS*				
Assigned to small	0.116**	0.095**		0.054***
	(0.046)	(0.048)		(0.016)
B. *TSLS*				
Average class size in STAR	–0.016**	0.013**		–0.007***
	(0.006)	(0.007)		(0.002)
C. *Lee bounds*				
Lower bound of small			0.032	
			(0.056)	
Upper bound of small			0.295***	
			(0.062)	
School-by-entry-wave FE	YES	YES	NO	YES
Student characteristics	YES	YES	YES	YES
Fourth grade characteristics	YES	NO	NO	NO
Teacher FE	NO	YES	NO	NO
Observations	2,188	2,188	7,324	7,324

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors are in parentheses (clustered at the classroom level in columns 1 and 2, bootstrapped with 500 repetitions in column 3, clustered at the school-by-entry-wave level in column 4). Student characteristics include gender, years of free-lunch eligibility and ethnicity. Fourth grade characteristics include teacher's gender, teacher's ethnicity and the share of non-white classmates.
Project STAR and Beyond database.

perform the following permutation test: we select five elements at random from all the non-cognitive items on the SPQ and construct a new index in the same way we construct grit. Then we estimate the effect of smaller classes on the random index. We repeat this procedure 2,000 times and plot the effect sizes for each simulation. Appendix Figure A.1 shows the results of this permutation exercise, where the solid line indicates the effect of small on our grit scale. We find that our grit scale is more affected by class size than most other randomly selected indices of non-cognitive skills. In detail, 66 times a random index yields effect sizes larger than that of grit. Therefore, the effect on grit is larger than 96.7 per cent of the other random five-element indices. The result of this permutation test supports the view that grit really is a distinct skill affected by class size.

5 External Validation of the Grit Measure and Long-Term Effects

The previous section shows that grit, as measurable with the SPQ data, can be modified through class size. However, we have yet to demonstrate whether grit – or more specifically the grit-approximation we use – actually determines other outcomes relevant for policy-makers, education professionals and families. Given that the literature indicates that grit drives both educational and labour market success (Duckworth et al., 2010; Robertson-Kraft and Duckworth, 2014), we do not provide here an extensive documentation of the effect of grit on such outcomes. Nonetheless, we provide a brief analysis of the effect of our grit measure on later school outcomes and of the way that the mechanism operates mainly through grit rather than class size per se.

In the data, we observe each pupil's score in the Tennessee Comprehensive Assessment Program (TCAP), the standardised achievement test used for comparing Project STAR cohorts for grades four through eight. This information is available for 1,832 (1,693) students in fourth (eighth) grade, corresponding to 84 (77) per cent of our sample. Additionally, for all the 2,188 students of our analytic sample, we know whether they graduated from high school on time and whether they took a college-entrance exam (ACT or SAT).

To investigate the relevance of grit for longer-term school outcomes and possibly understand the mechanism through which test scores are affected, we perform a causal mediation analysis. Imai, Keele, and Tingley (2010) and Imai, Keele, and Yamamoto (2010) provide excellent overviews of the method. Causal mediation analysis help identify the impact of an intermediate variable (in our case grit) that lie on the pathway between the treatment (in our case class size) and school outcomes.

In a potential outcomes framework, let G_i denote the grit level of student i, $S_i = s$ her binary treatment indicator and Y_i the educational school of interest. Since the individual grit is influenced by the treatment received, two potential values exist: $G_i(0)$ and $G_i(1)$. However, only one of these two values is observed for i, that is, $G_i = G_i(S_i)$. Similarly, let us define the multiple potential outcomes as $Y_i[T_i, G_i(S_i)]$. For example, $Y_i(1, 0.5)$ represents the fourth grade test scores that would be observed if student i was assigned to a smaller class during STAR and then has a grit score of 0.5.

Given this notation, Imai, Keele, and Tingley (2010) define the following causal indirect effect of grit as follows:

$$\mu_i(s) = Y_i[s, G_i(1)] - Y_i[s, G_i(0)] \tag{3}$$

for $t = 0, 1$. Therefore, the causal mediation effect μ represents the indirect effect of class size on a given school outcome that operates only through grit. Similarly, we can define the direct effect of small class on the outcome of interest as follows:

$$\nu_i(s) = Y_i[1, G_i(s)] - Y_i[0, G_i(s)] \tag{4}$$

where ν represents the direct effect of small class on student i's school outcomes while keeping her grit constant at the level that would be realized under the such treatment. Finally, the total effect of the treatment is given by the following equation:

$$\tau_i = Y_i[1, G_i(1)] - Y_i[0, G_i(0)] \equiv \frac{1}{2}\sum_{t=0}^{1}[\mu_i(s) + \nu_i(s)] \tag{5}$$

Since we are interested in average effects, we can compute $\bar{\mu}(s)$, $\bar{\nu}(s)$ and $\bar{\tau}$ by averaging over the sample under analysis. In practice, we first estimate a linear system of equations[10] and then calculate the effects of interest $\bar{\mu}(s)$, $\bar{\nu}(s)$ and $\bar{\tau}$. $\bar{\mu}(s)$ represents the average causal mediation effect (ACME) of grit on later school outcomes.

As Imai, Keele, and Yamamoto (2010) underscore, the identification of these direct and indirect effects requires a two-statement conditional independence assumption known as sequential ignorability. Sequential ignorability requires that (i) conditional on some predetermined variables, treatment assignment is random; and that (ii) the mediator is ignorable given the observed treatment and predetermined variables. While the first condition is, in our setting, fulfilled via randomization, the second might be violated if unobserved variables that confound the relation between the outcome and grit exist (even after conditioning on treatment status and predetermined characteristics).

Table 5 presents the results of our causal mediation analysis for the following school outcomes: fourth grade test scores, eighth grade test scores, on-time high school graduation (binary) and college-entry exam-taking (binary). Focusing on columns 1 and two, we see that being assigned to smaller classes during Project STAR has a positive and significant effect on test scores ($\bar{\tau}$). This result is not new: using the same data and a slightly different sample, Schanzenbach (2006) estimates almost identical effects in terms of significance and magnitude. However, we provide novel evidence that almost half of the class size effect operates through grit. The ACME of grit on test scores is significant ($p < 0.01$) and has approximately the same magnitude as the direct effect of small on test scores.

Table 5: Mediation Analysis of Grit on School Outcomes

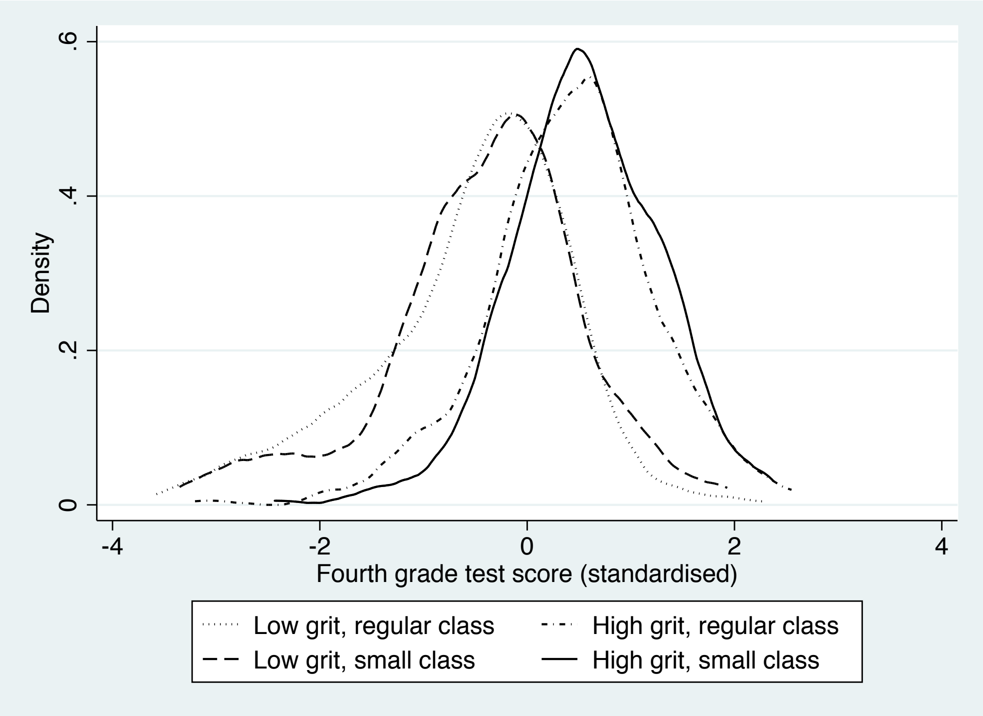
	Test scores grade four (z-score)	Test scores grade eight (z-score)	On-time graduation $YES = 1$	Took ACT/SAT $YES = 1$
	(1)	(2)	(3)	(4)
ACME of grit [$\bar{\mu}(s)$]	0.082***	0.071***	0.011**	0.020***
	(0.026)	(0.023)	(0.004)	(0.008)
Direct effect of small [$\bar{\nu}(s)$]	0.100***	0.094**	–0.004	0.001
	(0.041)	(0.044)	(0.025)	(0.024)
Total effect [$\bar{\tau}$]	0.182***	0.165***	–0.004	0.020
	(0.048)	(0.051)	(0.026)	(0.026)
School-by-entry-wave FE	YES	YES	YES	YES
Student characteristics	YES	YES	YES	YES
Fourth grade characteristics	YES	NO	NO	NO
Observations	1,832	1,693	2,188	2,188

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors clustered are in parentheses (clustered at the classroom level in column 1, clustered at the school-by-entry-wave level in columns 2-3-4). Inference for the ACME is performed vie Delta method. Student characteristics include gender, years of free-lunch eligibility and ethnicity. Fourth grade characteristics include teacher's gender, teacher's ethnicity and the share of non-white classmates. Test scores are from the Tennessee Comprehensive Assessment Program.
Project STAR and Beyond database.

Providing a visual representation of the results of column 1 of Table 5 and further investigating the relation amongst achievement, grit and class size, Figure 2 plots the densities of test scores for

[10] In detail, we estimate three equations as follows: (i) Y on S, X, Z, (ii) G on S, X, Z and (iii) Y on S, G, X, Z.

Figure 2: Density Plot of Test Scores in Grade 4, Divided by Grit Level and Treatment Group



four groups: low-grit pupils in regular classes, low-grit pupils in small classes, high-grit pupils in regular classes and high-grit pupils in small classes. We label pupils as 'low grit' if they have a grit scale below the mean and classify student as 'high grit' if they have an above-average grit score. The figure reveals that the main driver of test score results is grit, not class size per se. The most plausible explanation for these results, as suggested by the patterns in the data, is that class size has a positive impact on grit, which in turn leads to improved test scores. This pathway is also suggested by Cunha and Heckman (2008), who show that whilst non-cognitive skills promote the formation of cognitive skills, the reverse does not hold.

Columns 3 and 4 of Table 5 focus on high school on-time graduation and whether students took a college-entry exam, respectively. While we find no significant class size effects for these two outcomes, the ACME of grit is positive and significant ($p < 0.05$). This suggests that grit has indeed an indirect effect on both on-time graduation and exam-taking behaviour, but such effect is counterbalanced by other factors not related to grit. Overall, the results on long-term outcomes confirm both theoretical expectations and empirical evidence in the literature on grit (Eskreis-Winkler et al., 2014). These findings make us confident that even if we are not able to elicit grit in the exact same way as Duckworth our measure of grit indeed identifies a personality characteristic that determines test scores and long-term goal achievement and seems to capture the sense of Duckworth's concept.

6 Conclusions

We use a follow-up study of Project STAR and show that smaller classes increase student grit by 0.12 standard deviations. Moreover, when we estimate the effect of small classes on each item of the grit scale, pupils show significant higher persistence in effort both inside (i.e. higher participation, initiative behaviour) and outside the classroom (i.e. more time end effort spent on homework and preparation for tests). Sub-sample analysis reveals that pupils from disadvantaged socio-economic backgrounds and non-white pupils increase their grit significantly more in smaller classes, compared to the average student.

To highlight the importance of grit, we perform additional analyses of grit effects on test scores. This causal mediation analysis reveals that the main driving force of later school outcomes is grit, rather than other effects of reduced class size. These findings indicate that the underlying mechanism is the following: class size positively impacts grit, which in turn leads to higher test scores. Moreover, we examine the consequences of the key ingredient of grit, that is the pursuit of long-term goals and its effect on long-term outcomes. We study the relation between grit and eighth grade test scores, on-time high school graduation, and whether a student took a college-entrance exam. Results show that gritty students score higher on tests, are more likely to graduate from high school on time and are more likely to take a college-entrance exam.

The literature suggest a number of channels through which class size may foster student's grit. First, smaller classes allow a more intense pupil-teacher interaction (Dee, 2007; Finn and Achilles, 1999) and less distraction from peers (Lazear, 2001). Therefore, students engage more actively during instructional time and may learn that more engagement pays. Second, pupils in smaller classes have a higher chance of gaining the teacher's attention through effort (e.g. doing their homework) and interest (e.g. asking questions during class), thereby creating positive feedback loops. As Duckworth (2016) explains, encouraging repetition and refinement helps develop grit. If pupils learn that they can persist through challenges and eventually succeed, they will begin to define themselves by that persistence, not by momentary failures or challenges. Finally, as Heckman and Mosso (2014) suggest, pupils in smaller classes are likely to receive better mentoring and guidance, which in turn might affect their non-cognitive skills.

Despite our robust results showing the effects of attending smaller classes on grit, our grit measure should be seen as an approximation because we were not able to replicate the original grit scale by Duckworth et al. (2007). Moreover, it is difficult for us to draw conclusions of the exact increase in pupil's grit over time without pupil's baseline grit when entering school. Although we provide potential explanations for the underlying mechanism between class size and grit, future research could further investigate the underlying mechanism in detail. Future work could also study the importance of grit for lifetime achievement, investigating the relation amongst grit, educational practices (other than class size) and economic outcomes in later stages of life.

Our study adds to the extensive literature showing both the importance of non-cognitive skills and the power of educational practices. In addition, our study has two important policy implications. First, policy-makers and schools should reconsider class size as a way of promoting non-cognitive skills, particularly grit. Second, we extend the literature on class size effects on (non-)cognitive skills, showing that the class size effect operates for a large part through grit. Therefore, both policy-makers and researchers should consider class size as an educational practice for changing not only cognitive skills and test scores but also non-cognitive skills such as grit. Note, however, that these policy implications are valid if we assume that hypothetical reductions in class size would hold all other educational inputs constant. This ceteris paribus interpretation might not be always valid, due to the complexity of class size reductions and parents' potential reactions to class size reductions.

References

Alan, Sule, Teodora Boneva, and Seda Ertac. 2015. “Ever failed, try again, succeed better: Results from a randomized educational intervention on grit.” Working paper no. 2015-009, Human Capital and Economic Opportunity Global Working Group.

Antecol, Heather, Ozkan Eren, and Serkan Ozbeklik. 2016. “Peer effects in disadvantaged primary schools: Evidence from a randomized experiment.” *Journal of Human Resources* 51 (1):95–132.

Bertrand, Marianne, Jessica Pan, and Emir Kamenica. 2013. “Gender identity and relative income within households.” NBER Working Papers 19023, National Bureau of Economic Research, Inc.

Blatchford, Peter, Paul Bassett, and Penelope Brown. 2011. “Examining the effect of class size on classroom engagement and teacher-pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools.” *Learning and Instruction* 21 (6):715–730.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas Weel. 2008. “The economics and psychology of personality traits.” *Journal of Human Resources* 43 (4):972–1059.

Chamorro-Premuzic, Tomas and Adrian Furnham. 2003. “Personality predicts academic performance: Evidence from two longitudinal university samples.” *Journal of Research in Personality* 37 (4):319–338.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. “How does your kindergarten classroom affect your earnings? Evidence from Project STAR.” *The Quarterly Journal of Economics* 126 (4):1593–1660.

Culin, Katherine R. Von, Eli Tsukayama, and Angela Lee Duckworth. 2014. “Unpacking grit: Motivational correlates of perseverance and passion for long-Term goals.” *Journal of Positive Psychology* 9 (4):306–312.

Cunha, Flavio and James J. Heckman. 2008. “Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation.” *Journal of Human Resources* 43 (4):738–782.

Dee, Thomas and Martin West. 2008. “The Non-Cognitive Returns to Class Size.” Tech. rep., National Bureau of Economic Research.

Dee, Thomas S. 2007. “Teachers and the gender gaps in student achievement.” *Journal of Human Resources* 42 (3):528–554.

Dee, Thomas S. and Martin R. West. 2011. “The non-cognitive returns to class size.” *Educational Evaluation and Policy Analysis* 33 (1):23–46.

Doyle, Orla, Colm P. Harmon, James J. Heckman, and Richard E. Tremblay. 2009. “Investing in early human development: Timing and economic efficiency.” *Economics & Human Biology* 7 (1):1–6.

Duckworth, Angela. 2016. *Grit: The Power of Passion and Perseverance*. Simon and Schuster, New York (USA).

Duckworth, Angela L, Christopher Peterson, Michael D Matthews, and Dennis R Kelly. 2007. “Grit: Perseverance and passion for long-term goals.” *Journal of Personality and Social Psychology* 92 (6):1087–1101.

Duckworth, Angela Lee and James J. Gross. 2014. “Self-control and grit: Related but separable determinants of success.” *Current Directions in Psychological Science* 23 (5):319–325.

Duckworth, Angela Lee, Teri A. Kirby, Eli Tsukayama, Heather Berstein, and K. Anders Ericsson. 2010. “Deliberate practice spells success: Why grittier competitors triumph at the national spelling bee.” *Social Psychological and Personality Science* 2 (2):174–181.

Duckworth, Angela Lee and Patrick D. Quinn. 2009. “Development and validation of the short grit score (grit-s).” *The Journal of Personality Assessment* 91 (2):166–174.

Duckworth, Angela Lee and Martin E. P. Seligman. 2006. “Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores.” *The Journal of Educational Psychology* 98 (1):198–208.

Elder, Todd E. 2010. “The importance of relative standards in ADHD diagnoses: evidence based on exact birth dates.” *Journal of Health Economics* 29 (5):641–656.

Eskreis-Winkler, Lauren, Angela Lee Duckworth, Elizabeth P Shulman, and Scott Beal. 2014. “The grit effect: Predicting retention in the military, the workplace, school and marriage.” *Frontiers in Psychology* 5 (Feb):1–12.

Finn, Jeremy D and Charles M Achilles. 1990. “Answers and questions about class size: A statewide experiment.” *American Educational Research Journal* 27 (3):557–577.

Finn, Jeremy D. and Charles M. Achilles. 1999. “Tennessee’s class size study: Findings, implications, misconceptions.” *Educational Evaluation and Policy Analysis* 21 (2):97–109.

Finn, Jeremy D., DeWayne Fulton, Jayne Zaharias, and Barbara A. Nye. 1989. “Carry-over effects of small classes.” *The Peabody Journal of Education* 67 (1):75–84.

Finn, Jeremy D., Gina M. Pannozzo, and Charles M. Achilles. 2003. “The “why’s” of class size: Student behavior in small classes.” *Review of Educational Research* 73 (3):321–368.

Finn, Jeremy D. and Donald A. Rock. 1997. “Academic success among students at risk for school failure.” *The Journal of Applied Psychology* 82 (2):221–234.

Folger, John and Carolyn Breda. 1989. “Evidence from Project STAR about class size and student achievement.” *Peabody Journal of Education* 67 (1):17–33.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2016. “Parental responses to public investments in children: Evidence from a maximum class size rule.” *Journal of Human Resources* 51 (4):832–868.

Gerhards, Leonie and Christina Gravert. 2015. “Grit trumps talent? An experimental approach.” Economics Working Papers 2015-18, Department of Economics and Business Economics, Aarhus University (Denmark).

Hanushek, Eric A. 1999. “Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects.” *Educational Evaluation and Policy Analysis* 21 (2):143–163.

Heckman, James J. and Tim Kautz. 2012. “Hard evidence on soft skills.” *Labour Economics* 19 (4):451–464.

Heckman, James J. and Stefano Mosso. 2014. “The Economics of Human Development and Social Mobility.” IZA Discussion Papers 8000, Institute for the Study of Labor (IZA).

Heckman, James J., Rodrigo Pinto, and Peter Savelyev. 2013. “Understanding the mechanisms through which an influential early childhood program boosted adult outcomes.” *American Economic Review* 103 (6):2052–2086.

Heckman, James J. and Yona Rubinstein. 2001. “The importance of noncognitive skills: Lessons from the GED testing program.” *American Economic Review* 91 (2):145–149.

Heckman, James J., Jora Stixrund, and Sergio Urzua. 2006. “The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior.” *The Journal of Labor Economics* 24 (3):411–482.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. “A general approach to causal mediation analysis.” *Psychological Methods* 15 (4):309–334.

Imai, Kosuke, Luke Keele, and Teppei Yamamoto. 2010. “Identification, inference and sensitivity analysis for causal mediation effects.” *Statistical Science* 25 (1):51–71.

Jepsen, Christopher and Steven Rivkin. 2009. “Class size reduction and student achievement: The potential tradeoff between teacher quality and class size.” *Journal of Human Resources* 44 (1):223–250.

Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. 2014. “Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success.” NBER Working Papers 20749, National Bureau of Economic Research, Inc.

Konstantopoulos, Spyros. 2008. “Do small classes reduce the achievement gap between low and high achievers? Evidence from Project STAR.” *The Elementary School Journal* 108 (4):275–291.

Krueger, Alan B. 1999. “Experimental estimates of education production functions.” *The Quarterly Journal of Economics* 114 (2):497–532.

Krueger, Alan B and Diane M Whitmore. 2001. “The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR.” *The Economic Journal* 111 (468):1–28.

Lazear, Edward P. 2001. “Educational production.” *The Quarterly Journal of Economics* 116 (3):777–803.

Lee, David S. 2009. “Training, wages, and sample selection: Estimating sharp bounds on treatment effects.” *The Review of Economic Studies* 76 (3):1071–1102.

Maddi, Salvatore R. 2006. “Hardiness: The courage to grow from stresses.” *Journal of Positive Psychology* 1 (3):160–168.

McCrae, Robert R. and Oliver P. John. 1992. “An introduction to the five-factor model and its applications.” *Journal of Personality* 60 (2):175–215.

McDonald, Jennifer Dodorico. 2008. “Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments.” *Enquire* 1 (1):1–18.

Paulhus, Delroy L. 1991. “Measurement and control of response bias.” In *Measures of Personality and Social Psychological Attitudes*, edited by John P Robinson, Phillip R Shaver, and Wrightsman Lawrence S, chap. 2. Academic Press, San Diego (USA), 17–59.

Paulhus, Delroy L and Simine Vazire. 2007. “The self-report method.” *Handbook of Research Methods in Personality Psychology* 1:224–239.

Robertson-Kraft, Claire and Angela Lee Duckworth. 2014. “True grit: Trait-level perseverance and passion for long-term goals predicts effectiveness and retention among novice teachers.” *Teachers College Record* 166 (3):1–27.

Schanzenbach, Diane Whitmore. 2006. “What have researchers learned from Project STAR?” *Brookings Papers on Education Policy* 9 (2006-2007):205–228.

Segal, Carmit. 2008. “Classroom behavior.” *Journal of Human Resources* 43 (4):783–814.

———. 2013. “Misbehavior, education, and labor market outcomes.” *Journal of the European Economic Association* 11 (4):743–779.

Word, Elizabeth, John Johnston, Helen Pate Bain, BD Fulton, Jayne Boyd Zaharias, Charles M Achilles, Martha Nannette Lintz, John Folger, and Carolyn Breda. 1990. “The state of Tennessees Student/Teacher Achievement Ratio (STAR) Project: Final summary report 1985-1990.” *Nashville: Tennessee State Department of Education (USA)* .

A Appendix: Supplementary Material

Table A.1: The Five-Item Grit Scale Compared to Duckworth's Eight-Item Grit Scale

Eight-item Grit Scale For Children Duckworth et al. (2007)	*Five-item Grit Scale Adapted from SPQ*
1. New ideas and projects sometimes distract me from previous ones.	—
2. Setbacks (delays and obstacles) don't discourage me. I bounce back from disappointments faster than most people.	Student gets discouraged and stops trying when encounters an obstacle.
3. I have been obsessed with a certain idea or project for a short time but later lost interest.	Student doesn't take initiative, must be helped to get started and kept going on work.
4. I am a hard worker.	Student is persistent when confronted with difficult problems.
5. I often set a goal but later choose to pursue (follow) a different one.	—
6. I have difficulty maintaining (keeping) my focus on projects that take more than a few months to complete.	—
7. I finish whatever I begin.	Student tries to finish difficult assignments.
8. I am diligent (hard working and careful).	Student does more than just the assigned work.

Table A.2: The Student Participation Questionnaire

This student –	Never		Sometimes		Always
1. pays attention in class.	1	2	3	4	5
2. completes homework in time.	1	2	3	4	5
3. works well with others.	1	2	3	4	5
4. loses, forgets, or misplaces materials.	1	2	3	4	5
5. comes late to class.	1	2	3	4	5
6. attempts to do his/her work thoroughly and well, rather than just trying to get by.	1	2	3	4	5
7. acts restless, is often unable to sit still.	1	2	3	4	5
8. participates actively in discussions.	1	2	3	4	5
9. completes assigned seat work.	1	2	3	4	5
10. thinks that school is important.	1	2	3	4	5
11. needs to be reprimanded.	1	2	3	4	5
12. annoys or interferes with peers' work.	1	2	3	4	5
13. is persistent when confronted with difficult problems.	1	2	3	4	5
14. doesn't seem to know what is going on in class.	1	2	3	4	5
15. does more than just the assigned work.	1	2	3	4	5
16. is withdrawn, uncommunicative.	1	2	3	4	5
17. approaches new assignments with sincere effort.	1	2	3	4	5
18. is critical of peers who do well in school.	1	2	3	4	5
19. asks questions to get more information.	1	2	3	4	5
20. talks with classmates too much.	1	2	3	4	5
21. doesn't take independent initiative, must be helped to get started and kept going on work.	1	2	3	4	5
22. prefers to do easy problems rather than hard ones.	1	2	3	4	5
23. criticises the importance of the subject matter.	1	2	3	4	5
24. tries to finish assignments even when they are difficult.	1	2	3	4	5
25. raises his/her hand to answer a question or volunteer information.	1	2	3	4	5
26. goes to dictionary, encyclopaedia, or other reference on his/her own to seek information.	1	2	3	4	5
27. gets discouraged and stops trying when encounters an obstacle in school work, is easily frustrated.	1	2	3	4	5
28. engages teacher in conversation about subject matter before or after school, or outside of class.	1	2	3	4	5
29. attends other school activities such as athletic contests, carnivals, and fund-raising events.	1	2	3	4	5
	Above Average		Average		Below Average
30. The student's overall academic performance is	1		2		3
			No		Yes
31. Does this student attend special education			1		2

Table A.3: Effect of Class Size on Dis-aggregated Grit Scale, OLS and TSLS

	Does not take initiative (reversed) (1)	Does more than the assigned work (2)	Gets discouraged easily (reversed) (3)	Is persistent when confronts problems (4)	Tries to finish difficult assignments (5)
A. *OLS*					
Assigned to small	0.101**	0.112**	0.061	0.144***	0.079*
	(0.051)	(0.046)	(0.045)	(0.046)	(0.046)
B. *TSLS*					
Average class size in STAR	–0.014**	–0.015**	–0.008	–0.020***	–0.011*
	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)
School-by-entry-wave FE	YES	YES	YES	YES	YES
Student characteristics	YES	YES	YES	YES	YES
Fourth grade characteristics	YES	YES	YES	YES	YES
Observations	2,188	2,188	2,188	2,188	2,188

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Robust standard errors clustered at the classroom level are in parentheses. Student characteristics include gender, free-lunch eligibility and ethnicity. Fourth grade characteristics include teacher's gender, teacher's ethnicity and the share of non-white classmates. All dependent variables are standardised.

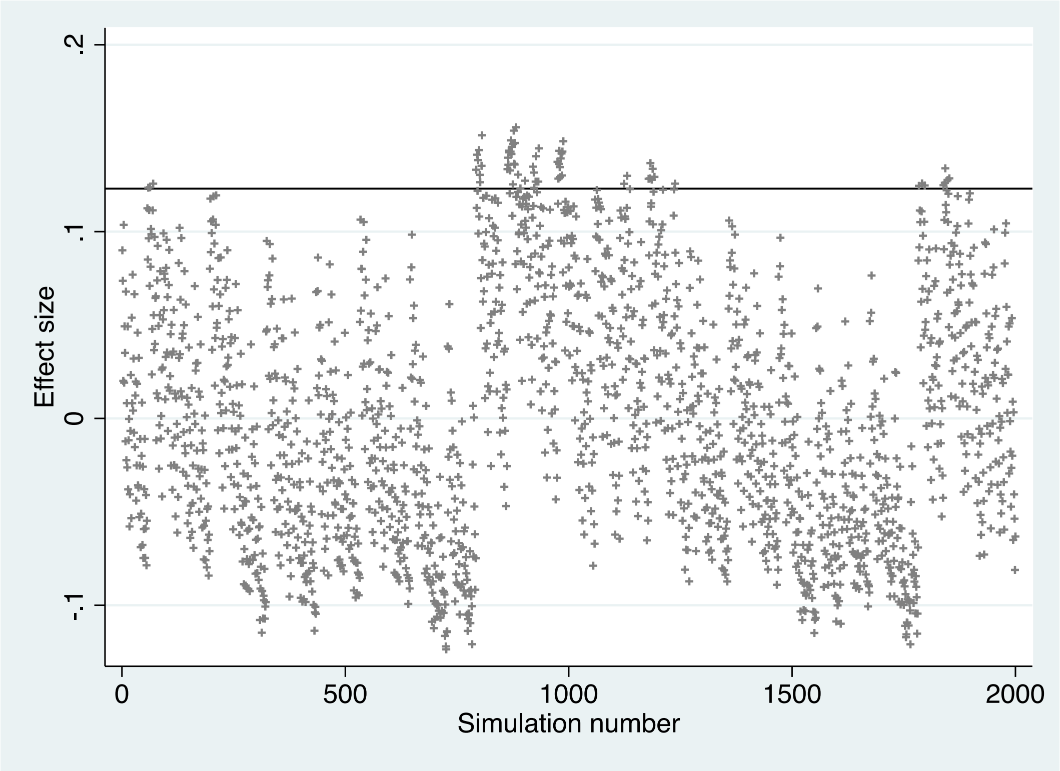
Project STAR and Beyond database.

Table A.4: Attrition Analysis

	Fourth grade data			Survey participation		
	Small	Regular	Difference	Small	Regular	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Attrition rate	0.558	0.542	0.02	0.418	0.319	0.99***
Non-random attrition						
Boy	0.495	0.530	–0.04*	0.307	0.320	–0.01
Free-lunch eligible	0.538	0.513	0.03	0.287	0.303	–0.02
Non-white	0.491	0.491	0.00	0.354	0.321	0.03

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Ordinary least squares models with robust standard errors clustered at the school-by-entry-wave level. Sample size for columns (1), (2) and (3) ranges from 11,467 to 11,601 and for columns (4), (5) and (6) ranges from 6,333 to 6,339.
Project STAR and Beyond database.

Figure A.1: Permutation Test: Effect Sizes on Randomly Selected Indices


Effect size
Simulation number