How Do Shocks to Non-Cognitive Skills Affect Test Scores?

Stefanie Behncke

June 2009 Discussion Paper no. 2009-11

Department of Economics University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
Department of Economics
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email vwaabtass@unisg.ch

Publisher: Department of Economics
University of St. Gallen
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35

Electronic Publication: http://www.vwa.unisg.ch

How Do Shocks to Non-Cognitive Skills Affect Test Scores?[1]

Stefanie Behncke

Author's address:

Stefanie Behncke
SEW-HSG
Varnbüelstr. 14
CH-9000 St. Gallen
Email stefanie.behncke@unisg.ch
Website www.sew.unisg.ch

[1] I am also affiliated with IZA, Bonn, and was visiting the Centre for Health Economics (CHE) in York while a substantial part of this paper was written. I acknowledge financial support from the Swiss National Science Foundation (SNSF). I am grateful to Michael Lechner, David Clingingsmith, Leslie Godfrey and Conny Wunsch for valuable discussions. I also thank Manfred Gärtner for making this research project possible and Peter Gruber for developing the maths test.

Abstract

This paper investigates the extent to which test performance is affected by shocks to non-cognitive skills. 440 students took a low stakes mathematics test. About half of them were exposed to positive affirmation while being given test instructions, whereas the other half served as controls. The students were allocated to 14 tutorials and randomisation was conducted at the tutorial level. Mean comparisons suggest that test scores were raised by the intervention. In particular, students with low maths grades and with self-assessed difficulties in maths gained from the positive affirmation. Results suggest that teachers might increase their students' performance by interventions to their non-cognitive skills.

Inference is obtained by four different methods that take into account that randomisation was clustered at the tutorial group level. These methods are evaluated in a Monte Carlo study for data generating processes which resemble actual data. We find that randomisation inference followed by the wild cluster bootstrap have superior size properties compared to conventional approaches.

Keywords

test scores, non-cognitive skills, cluster randomised trial, wild cluster bootstrap, randomisation inference

JEL Classification

C15, C21, C93, I20

1 Introduction

To a large extent achievement in tests determines schooling and job market decisions. Test scores are considered to signal the ability of the person taking the test. Often ability is equated with *cognitive* skills. However, evidence from the field of psychology suggests that *non-cognitive* skills such as *self-discipline* (Duckworth and Seligman, 2005), *motivation* (see review in Borghans, Duckworth, Heckman, and ter Weel (2008)), *self-confidence* (Bouffard-Bouchard, Parent, and Larivee, 1991), vulnerability to *stereotype threat* (Steele, 1997) and *test anxiety* (Hembree, 1988, 1990) also affect test performance. Personality traits have been shown to affect a range of socio-economic outcomes such as schooling, wages, crime and teenage pregnancy besides their effect on test performance (Heckman, Stixrud, and Urzua, 2006).

If non-cognitive skills matter for achievement, is an intervention able to influence them? Evidence from long term interventions in early childhood suggests that success in life is affected by altering long-term personality factors (Heckman and Masterov, 2007). We examine a short term intervention to non-cognitive skills for young adults. We assess whether it has an impact on their achievement in a cognitive test. The intervention is simple and cheap: students receive positive affirmation before taking a maths test. The outcome is considerable: test scores are significantly raised. In particular, students reporting difficulties in maths and students with low maths grades benefit from the intervention. This suggests that teachers may raise their students' performance by positively affirming them. Furthermore, the finding emphasises the importance of standardised tests when comparing students' cognitive and non-cognitive abilities.

Our work is most closely related to Borghans, Meijers, and Weel (2008). In an experiment, they examine how students with different non-cognitive skills respond to financial rewards when taking a cognitive test. They find that a student's time investment in answering a question depends on his personality traits and economic preferences: students with favourable personality traits such as high performance motivation or conscientiousness invest relatively more time in the absence of rewards and invest less than average time when there is an incentive pay; students with low discount rates and low risk aversion,

though, invest more time when rewards are introduced. Changes in time investment, however, are not accompanied by changes in test scores: financial incentives do not have significantly different effects on test scores for students with different personalities. Our results confirm Borghans, Meijers, and Weel (2008): individual behaviour at cognitive tests depends on non-cognitive skills. They also suggest that other interventions to non-cognitive skills might be more successful in raising test performance compared to financial rewards.

We conducted a field experiment to evaluate the intervention. 440 students took a low stakes maths test. About half of them received positive affirmation before the test was taken. The other half were just given the test instructions. Students were allocated to 14 tutorials taught by 7 tutors. Treatment was randomly assigned at the tutorial level (*cluster randomised trial*). Assignment was blocked with regard to the 7 tutors. In Monte Carlo simulations, we assess different approaches for obtaining standard errors in the presence of intracluster correlation. We add to other Monte Carlo studies on this issue by (i) allowing for *negative* intracluster correlation and by (ii) comparing randomisation inference that takes into account *blocking* with approaches that ignore it. We find that randomisation inference outperforms other approaches for arbitrary intracluster correlation and when randomisation is blocked with respect to covariates that explain the outcome variable. This is at the expense of reduced power. The wild cluster bootstrap achieves comparable size properties when blocking is unnecessary. For mean comparisons we present p-values based on five different approaches to obtain standard errors with clustered data.
This paper proceeds as follows. The next section provides some background on the role of non-cognitive skills on performance. In Section 3 the experimental setting is described. Section 4 discusses methodological issues such as estimation of treatment effects, and inference when observations within a cluster are correlated and provides Monte Carlo simulation. Section 5 describes results. The final section provides a conclusion.

2 Background

Non-cognitive aspects of ability such as motivation and self-confidence have often been neglected in economic models. However, empirical evidence suggests that they matter for success as much as (or even more) than cognitive ability. For instance, Duckworth and Seligman (2005) find that self-discipline outperforms IQ in predicting achievement in tests. Similarly, Heckman, Stixrud, and Urzua (2006) find that a change in non-cognitive ability has a comparable or even greater effect on a range of socio-economic outcomes than an equal change in cognitive skills. Furthermore, evidence suggests that non-cognitive skills are more malleable in later years than cognitive skills, which are fairly settled by the age of eight (Cunha, Heckman, and Lochner, 2006). This raises questions as to whether interventions and investment in non-cognitive skills could be more efficient compared to targeting cognitive skills.

In this section, we review some evidence on the impact of non-cognitive skills on cognitive tests. Before doing so, we want to emphasise that the use of the term "cognitive" in contrast to "non-cognitive" might be misleading (Borghans, Duckworth, Heckman, and ter Weel, 2008). Many aspects of personality are a consequence of cognition; and cognition depends on personality. However, we adapt the common notion in the literature. We subsume raw problem solving ability (intelligence) under "cognitive". We distinguish it from other "non-cognitive" abilities such as perseverance, attention, motivation, and self-confidence.

Zigler and Butterfield (1968) provide evidence that *motivation* can substantially increase performance in IQ tests. They find that children from deprived backgrounds achieved higher test scores in tests that maximised their motivation to perform well compared to standard IQ tests. They also find that increases in children's standard IQ test results after attending a nursery school were due to motivational factors rather than changes in the rate of intellectual development. Borghans, Duckworth, Heckman, and ter Weel (2008) summarise several studies that show that *extrinsic incentives* can increase cognitive test performance. Positive effects of cash and candy incentives are mainly found

for children with a low IQ or socio-economic background. For high school students with a high IQ no incentive effects can be established. An interesting experiment in this respect is Borghans, Meijers, and Weel (2008), who find that financial incentives have heterogeneous effects on time spent on cognitive tests depending on an individual's personality traits. Students with high performance motivation, an internal locus of control, curiosity, and other favourable personality traits invest less time in answering a question when there is a financial reward compared to those who score lower on these traits. On the other hand, students with lower discount rates and lower risk aversion invest more time when there is an incentive pay reward. While financial rewards change time investment, they do not result in significantly different effects on test scores for students with different personalities. Increased time investment is generally not increasing test scores. Overall this suggests that extrinsic incentives can increase test performance, but not necessarily for *all* individuals. Furthermore, individuals appear to have quite diverse responses to rewards depending on their personality traits. It should also be kept in mind that extrinsic incentives can have detrimental effects on intrinsic motivation (Deci and Moller, 2005).

Self-confidence has also been found to affect achievement. Bandura (1993) argues that perceived self-efficacy affects cognitive and motivational processes: students' beliefs in their efficacy to regulate their own learning and to master academic activities determine their aspirations, their level of motivation, and, consequently, their academic accomplishments. Bouffard-Bouchard, Parent, and Larivee (1991) find that children with the same level of skill development in mathematics differed significantly in their maths problem solving process depending on the strength of their beliefs about self-efficacy. In a related literature, it has been shown that self-confidence can be negatively affected by *stereotype threats*. Steele (1997) provides evidence that stereotypes (for instance that "women are not good at mathematics") impair test performance when a member of a stereotyped group is tested. Self-affirmation has been found to improve the performance of individuals under stereotype threat (Martens, Johns, Greenberg, and Schimel, 2006).

Finally, Hembree (1988) and Hembree (1990) provide evidence that *test anxiety* can seriously impair test performance. Test anxiety relates to students' self-esteem and to their

fears of negative evaluation. Its impact depends on the perceived degree of difficulty of the respective test.

3 The experiment

The population of the experiment consists of students participating in the first tutorial of undergraduate macroeconomics in the autumn term of 2007 at the University of St. Gallen in Switzerland. Tutorials accompany lectures: material from the lecture is reviewed and problem sets are solved. Students have an incentive to participate in tutorials because they serve as exam preparation. Their participation is not obligatory. Of around 1000 students who took the final exam, around 450 students attended the first tutorial. There were 14 tutorials held by 7 tutors and they took place four days after the lecture. Each tutor was teaching one class from 12 to 2 pm and one from 2 to 4 pm. In order to achieve equal class sizes, the University of St. Gallen has implemented a so-called bidding system, where students can bid a share of a fixed amount of points to be assigned to a certain tutorial according to their preferences. As a consequence, each student is assigned to one of the 14 tutorials. However, whether students actually attend the assigned classes or not is not controlled. In practice, it might be that students switch between tutorials according to their time, tutor or peer preferences.

A low stakes mathematics test was conducted at the beginning of the first tutorial. This was to show students and their tutors the extent to which students have difficulties in certain mathematical methods that are important for undergraduate macroeconomics. Questions consisted of solving linear equation systems, drawing linear functions, deriving the total differential and applying the rules for logarithms and exponents. The test was anonymous. In a questionnaire attached to the front of the test, students were asked about their age, gender, and field of study. They were also asked whether they had difficulties in economics due to lack of mathematical skills, how long they had revised or prepared for the first lecture, and what their grade was in the mathematics exam in their first year (assessment level). Students had 15 minutes to solve ten questions, and then the tests were collected. They were marked twice by different people who did not know

which of the 14 tutorials the student belonged to.
The treatment was positive affirmation whilst giving the test instructions. Before the test was taken each tutor read an instruction text from a printed paper. The original German text is shown in Table A.1 in the Appendix. The English translation is:

"Before we discuss the problem set, we will conduct a self-test in mathematics. The test is anonymous. It will not count towards your grade. The topics of the test are precisely those mathematical methods that are important for this macroeconomics class. If you encounter difficulties in solving these problems, we recommend reviewing the respective topics. The test serves two purposes. First, it will allow you to determine your own weak spots. Second, it will show the tutors which questions cause problems. You have 15 minutes to answer the questions. We will then collect and correct it. You will receive printed solutions to assess your own level."

In the treatment group the following sentences were added:
"I am sure that you will solve the given problems very well. You have already taken tests in the past with success; otherwise you would not be here."

The treatment was meant to positively affirm students. The first sentence was intended to signal to the students that the tutor believed that their mathematical skills were more than sufficient for this test. This sentence could affect a student's non-cognitive skills through various channels. First, a teacher's expectations might have some self-fulfilling components (Pygmalion effects). Second, performance impairing test anxiety might be reduced when students do not need to worry that they will not perform well. Third, intrinsic motivation to achieve good results might be increased if this goal is described as achievable. The second sentence was supposed to remind them of their past successes when taking tests. From the field of psychology it is known that reminding individuals of their past achievements boosts self-confidence. This will reinforce the non-cognitive skills that have been activated by the first sentence. Arguably, this sentence might also provoke

the opposite effect if said to individuals with a history of under-achievements. However, these were all students who have been admitted to the University and passed a difficult examination in their first University year, so that it was considered to be most likely that the sentence caused the intended effect.

Randomisation was implemented at the student group level: 7 out of 14 student groups were randomised to receive the treatment, while the other half served as the control group. A cluster randomised trial can have the drawback of reduced estimation precision, compared to randomisation at the individual level, as individuals within a cluster cannot be considered as independent observations. However, in our context, students are clustered in student groups anyway. They self-select into the 14 tutorials, experience common shocks within a class room and interact with their peer students. Consequently, it was a natural choice to randomise student groups instead of students. The other option would have been to print the positive affirmation at the beginning of the maths test. However, this would probably have had a weaker treatment effect. The positive affirmation would have been less personal, less noticeable and less credible. Moreover, there would have been the risk of spill-over effects if the neighbouring student belonged to a different group.
Furthermore, student groups were blocked before randomisation according to their tutor. When there are a few clusters, blocking with respect to important covariates is recommended to reduce imbalance (Cox and Reid, 2002; Bloom, 2004; Donner and Klar, 2004). In this setting, each tutor was assigned one group under treatment and another without, in order to avoid a bad draw in which many tutors had only one treatment status. If there are tutor fixed effects (for instance, because better students select a certain tutor) a trial without blocking could have resulted in a situation in which students assigned to the treatment group would be different from students in the control group. Thus, blocking with respect to the tutor ensures that treatment and control group each represent each block in the same proportion. This increases the precision of the treatment estimator by reducing standard errors, provided that tutors explain variations in test scores.
The internal validity of the experiment might be harmed if the following problems oc-

cur: (i) randomisation changes the process of selection into the treatment (randomisation bias); (ii) students in the control group receive close substitutes for the positive affirmation (substitution bias); (iii) students assigned to one group do not (or only partly) participate (drop-out bias), (iv) students behave differently because they know that an experiment is conducted (Hawthorne effects) (see Heckman and Smith (1995)). As students were not informed before the tutorial that an experiment would be conducted, randomisation bias and Hawthorne effects can be ruled out. While it may be the case that students in the second tutorial were informed that a maths test would be conducted, they were unlikely to know about the two different instruction texts. As students did not leave the classroom during the test, substitution bias can be ruled out as well. Drop-out bias might occur if students assigned to treatment did not complete the tests. However, all returned tests did contain responses. No students are known to have not returned the test. Thus, it is unlikely that drop-out bias pollutes the randomisation design. Therefore, the interval validity of the randomisation evaluation appears to be warranted.

We now turn to check whether the randomisation did balance important covariates. Table A.2 in the Appendix provides some descriptive statistics on students' characteristics, which were obtained from the questionnaire. The first two columns in the upper panel show mean characteristics for the overall treatment and control group. The other columns show mean characteristics for each of the 14 student groups. The table also provides a t-test for equality of means, where standard errors were clustered according to the cluster-robust variance matrix (described in the next section).

Overall, 233 students were exposed to the positive affirmation and 207 students served as controls. Around 30 percent of them are female. On average, they were 22 years old. More than half of the students were business students, the others were economics, international affairs and law students. Their average grade in mathematics in their first year was around 4.75, where 6 is the highest grade and a student failed with a grade lower than 4. Around one third of the students considered themselves to have difficulties in economics because they lacked maths skills. On average, they prepared or revised for

the first macroeconomics lecture for less than an hour. Equality of means tests do not suggest that there are systematic differences between the two groups when aggregated. The rows below show average test outcomes for the overall test. On average, students in the treatment group achieved 6.27 out of 10 points, while students in the control group achieved 6.01 points. The differences are not significant. However, as will be discussed in the next section, inference in the presence of a small number of clusters is not straightforward. Results with more appropriate inference are therefore provided in Section 5.

When considering the student groups separately, we notice some heterogeneity between groups. In the largest two tutorials, 45 students participated, whilst there were only 14 students in the smallest tutorial. There are usually more students in tutorials from 12 to 2 pm compared to the later one. The share of students by gender, field of study and self-assessed difficulties in economics differs by groups. The test scores achieved range from an average of 5.19 to 6.73. These differences could suggest a potential self-selection into groups. Furthermore, we notice that the average test scores of students taught by the same tutor are more alike, which could be a sign for tutor effects.

Figure A.1 in the Appendix shows the distribution of the test scores in treatment and control groups. The test scores are discrete, as one point was given for any correct answer and otherwise zero was given. Each student could answer at least a part of the test. The minimum test score (1 point) and the maximum test score (10 points) were each achieved by seven students. The median test score in both groups was 7. The variance in the control group is higher than in the treated group. Both distributions are negatively skewed.

4 Methodology

4.1 Estimation of treatment effects

The outcome of interest is the achieved test score, which takes discrete values between 0 and 10. In the following it is assumed that it is cardinal. In other words, the distance

between achieving $y = x$ and $y = x + 1$ point is the same for all x. We argue that in our case this is a valid assumption, because all ten questions are considered to measure the same degree of difficulty. Under this assumption, the difference in average test scores is meaningful.

Since assignment to treatment was at random, unbiased treatment effects can be obtained by comparing the means in the control and treatment groups. This is illustrated in the potential outcome framework introduced by Rubin (1974). Let Y_i denote the test score of student i. Let D_i denote the treatment state of this student, i.e. $D_i = 1$ if he was exposed to positive affirmation and $D_i = 0$ if he was not. We would like to compare test outcomes for the same student if he was exposed to treatment (as denoted by Y_i^1) and if he was not exposed to treatment (as denoted by Y_i^0). But for every student only one of these potential outcomes is observed. Therefore, we cannot obtain the treatment effect for a given student. However, we can obtain an estimate of the average impact of the intervention. The average treatment effect (ATE) is defined as the difference in test scores under treatment and no treatment for an individual randomly drawn from the population:

$$E[Y^1 - Y^0].$$

In non-experimental frameworks, individuals who receive treatment are different from individuals without treatment not only with regard to their treatment status, but also in other covariates that affect outcomes. Then, a simple comparison of outcomes between treated and non-treated individuals would result in selection bias. In an experiment, however, the selection bias is removed, because assignment is random. As assignment is uncorrelated with the attributes of the individual, on average individuals in the $D = 1$ group are similar to individuals in the $D = 0$ group. In other words, random assignment ensures that treated and control groups have the same distribution of characteristics. This implies that potential outcomes in treatment and control groups equal outcomes in the

population:

$$E[Y^1|D=1] = E[Y^1|D=0] = E[Y^1],$$
$$E[Y^0|D=1] = E[Y^0|D=0] = E[Y^0].$$

Thus, we can consistently estimate the ATE by the difference in means between treatment and control groups. Furthermore, because treated individuals are randomly drawn from the population of interest, the average treatment effect on the treated (ATET) equals the ATE. Besides estimating the average treatment effect, we also estimate quantile effects to assess how the intervention affects the distribution of outcomes. But we confine ourselves to ATEs when conducting inference.

4.2 Inference in the presence of clusters

While obtaining consistent estimates for the ATE is not affected by the presence of clusters, inference is. If individuals within the same group are subject to common shocks, their outcomes might be correlated. Moreover, because all individuals in the same group have the same treatment status, the correlation in their outcomes might mistakenly be interpreted as the treatment effect. To illustrate this, suppose there was a positive correlation within clusters. Then each individual would contribute less to statistical efficiency compared to a case where there were independent observations. Standard errors which were not adjusted for clustering would underestimate the true variance of the treatment effect. This would result in over-rejections of the null hypothesis of no treatment effect. Bertrand, Duflo, and Mullainathan (2004) illustrate the severity of this issue: they find 5% significant effects of a non-existing intervention in up to 45% of the placebo intervention.

In the following, we discuss this issue more formally. Let us consider the linear regression model of y_i on a constant and the treatment indicator D_i, where $i = 1, ...N$ denotes the ith sample observation. The coefficient of the treatment indicator gives the ATE. There are $c = 1, ...C$ clusters in the overall sample. Suppose that the ith individual in the sample

is the jth individual in the cth cluster. Then the model for clustered data is:

$$y_{jc} = \alpha + \beta D_{jc} + u_{jc}, \qquad j = 1, ..N_c, \quad c = 1, ...C, \tag{1}$$

where $Cov[u_{jc}, u_{kc}] \neq 0$ and $Cov[u_{jc}, u_{kd}] = 0$ for $c \neq d$. Thus, errors of individuals belonging to the same cluster may be correlated, while errors of individuals belonging to different clusters are uncorrelated. Note that this model allows for arbitrary individual effect heterogeneity, i.e. β is not restricted to be the same for all individuals and could be different in different clusters and different within clusters. Stacking observations within a cluster yields:

$$\boldsymbol{y}_c = \alpha + \beta \boldsymbol{D}_c + \boldsymbol{u}_c, \tag{2}$$

where $\boldsymbol{y}_c$, $\boldsymbol{D}_c$ and $\boldsymbol{u}_c$ are $N_c \times 1$ vectors. The ATE estimator is

$$\widehat{\beta} = \overline{Y}^1 - \overline{Y}^0, \tag{3}$$

where $\overline{Y}^1$ is the mean in the treatment group and $\overline{Y}^0$ is the mean in the control group. Let $\boldsymbol{X_c}$ denote a $N_c \times 2$ matrix with consisting of a unit vector and $\boldsymbol{D}_c$. Then, the central limit theorem yields

$$\sqrt{N}(\widehat{\beta} - \beta) \rightarrow \mathcal{N}[\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}], \tag{4}$$

where

$$\boldsymbol{A} = plim N^{-1} \sum_{c=1}^{C} \boldsymbol{X}_c' \boldsymbol{X}_c,$$

$$\boldsymbol{B} = plim N^{-1} \sum_{c=1}^{C} \boldsymbol{X}_c' \boldsymbol{u}_c \boldsymbol{u}_c' \boldsymbol{X}_c.$$

Different approaches for the estimation of B have been suggested. Moulton (1986) imposes assumptions about the structure of $\boldsymbol{u_c u_c'}$. In his model, the error term is decomposed into a random cluster specific constant α_c and a homoskedastic individual-specific component

ϵ_{jc}, i.e.:

$$y_{jc} = \alpha + \beta D_{jc} + \alpha_c + \epsilon_{jc} \tag{5}$$
$$\epsilon_{jc} \sim [0, \sigma_\epsilon^2]$$
$$\alpha_c \sim [0, \sigma_\alpha^2]$$

It has a positive and constant intraclass correlation coefficient which is defined as

$$\rho = Cor[\alpha_c + \epsilon_{jc}, \alpha_c + \epsilon_{kc}] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}. \tag{6}$$

For simplicity assume equal cluster size $N_c = m$ for every cluster. Let P denote the proportion of those treated. Then, the variance for the ATE estimator is:

$$Var[\widehat{\beta}]_{Moulton} = \frac{1}{P(1-P)} \frac{m\sigma_\alpha^2 + \sigma_\epsilon^2}{mC} \tag{7}$$

Due to the positive intracluster correlation, the variance is always higher compared to a situation where randomisation had been conducted at the level of the individual. An obvious drawback of this model is the strong assumption of homoskedasticity and positive and constant intracluster correlation.

A less parametrically restrictive approach is to use a cluster-robust variance estimator in the spirit of White (1980). If there are many clusters B can be consistently estimated by replacing $\boldsymbol{u}_c$ by $\widehat{\boldsymbol{u}_c} = \boldsymbol{y}_c - \widehat{\alpha} - \widehat{\beta}\boldsymbol{D}_c$. It follows that the ATE estimator is asymptotically normally distributed with variance:

$$Var[\widehat{\beta}]_{cluster-robust} = \frac{(N-1)}{(N-2)} \cdot \frac{C}{(C-1)} \left(\sum_{c=1}^{C} \boldsymbol{X}_c' \boldsymbol{X}_c \right)^{-1} \sum_{c=1}^{C} \boldsymbol{X}_c' \widehat{\boldsymbol{u}_c} \widehat{\boldsymbol{u}_c}' \boldsymbol{X}_c \left(\sum_{c=1}^{C} \boldsymbol{X}_c' \boldsymbol{X}_c \right)^{-1} \tag{8}$$

This formula places no restriction on heteroskedasticity and correlation within a cluster. But, it does assume that N_c is small and $C \to \infty$. Bertrand, Duflo, and Mullainathan (2004), Kézdi (2004) and Cameron, Gelbach, and Miller (2007) study properties of this variance estimator when the number of clusters is small ($C <= 30$). They find that stan-

dard errors are underestimated, leading to over-rejections in the usual Wald tests.

Cameron, Gelbach, and Miller (2007) suggest using bootstrap procedures when there are only a few clusters. In Monte Carlo studies they find that bootstrap leads to considerable improvement. From a theoretical point of view bootstrapping Wald statistics are preferred over bootstrapping standard errors, because the former allow for asymptotic refinement as they are asymptotically pivotal, i.e. their asymptotic distribution does not depend on unknown parameters. Indeed, they are found to have rejection rates closer to the theoretical value compared to bootstrapping procedures without asymptotic refinement. When comparing different bootstrap procedures with asymptotic refinement, Cameron, Gelbach, and Miller (2007) find that the *wild cluster bootstrap* procedure does especially well. It is a cluster generalisation of the wild bootstrap for heteroskedastic models (Wu, 1986). Its implementation is described in Table A.3 in the Appendix.

Finally, Duflo, Glennerster, and Kremer (2006) suggest the use of *randomisation inference* when the number of clusters is small or the covariance structure is unknown. Non-parametric randomisation inference was originally developed by Fisher (1935) and later extended by Rosenbaum (2002). It takes advantage of knowing the randomisation process. In particular, it is the only approach that is able to take into account that tutors have been blocked before randomisation. All other approaches are ignorant with respect to the actual randomisation process. Randomisation inference involves generating a set of all possible placebo random assignments $\{R_i\}$. With 7 tutors, each having one treatment and one control group, there are $2^7 = 128$ possible random assignments. (If there was no blocking and 7 out of 14 tutorials had been assigned to treatment, this set would be considerably larger with 3432 possible assignments.) For each possible random assignment the average treatment effect is estimated, resulting in a set of 128 $\{\hat{\beta}_R\}$. Since R_i is a placebo random assignment, $E[\beta_R] = 0$. Let $\hat{F}(\hat{\beta}_R)$ be the empirical c.d.f of $\hat{\beta}_R$ for all elements in $\{R_i\}$. We can now test whether the average treatment effect is significantly different from zero by checking whether it lies in the tails of the distribution of placebo treatments. We can reject $H_0 : \beta = 0$ with a confidence level of $1 - \alpha$ if $\widehat{\beta} \leq \hat{F}^{-1}(\frac{\alpha}{2})$ or $\widehat{\beta} \geq \hat{F}^{-1}(1 - \frac{\alpha}{2})$.

4.3 Monte Carlo simulations

We conduct Monte Carlo simulations to examine finite sample properties of five tests which use different approaches to address intracluster correlation:

1. a robust White estimator (ignoring clustering),
2. a Moulton estimator, as in equation 7 (assuming a positive and constant intracluster correlation),
3. a cluster-robust estimator, as in equation 8 (allowing for arbitrary covariance, but relying on observing many clusters),
4. a wild-cluster bootstrap implemented according to Table A.3 (allowing for a few clusters with arbitrary covariance)
5. and randomisation inference (allowing for a few clusters with arbitrary covariance and taking blocking into account).

The properties of the first four estimators have already been studied by Bertrand, Duflo, and Mullainathan (2004), Kézdi (2004) and Cameron, Gelbach, and Miller (2007). We add to this literature (i) by allowing for negative correlation within a cluster and (ii) by comparing randomisation inference, which takes blocking in the randomisation into account with approaches that ignore blocking. Furthermore, we aim to study properties for data generating processes which resemble the actual observations in our experiment. This will allow a more precise evaluation of variance estimators for our data. Therefore, we generate discrete data which take values that have been actually observed as test scores. We use as many clusters ($C = 14$), and as many observations within each cluster, as in our experiment. Each data generating process is generated as

$$y_{jc} = \begin{cases} 1 & \text{if} \quad error_{jc} <= q_1 \\ 2 & \text{if} \quad q_1 < error_{jc} <= q_2 \\ 3 & \text{if} \quad q_2 < error_{jc} <= q_3 \\ \vdots & \\ 10 & \text{if} \quad q_9 < error_{jc} \end{cases} \tag{9}$$

where $error_{jc}$ is a random component and q_i with $i = 1, ...9$ are the quantiles of $error_{jc}$ which are chosen to reflect the distribution of actual test scores. For instance, q_1 is the 0.016 quantile as there are 7 out of 440 individuals with a test score of 1. The intracluster correlation is modelled by assuming different forms of the error term $error_{jc}$. We impose $H_0 : \beta = 0$ for every data generating process. We then use the actual treatment indicator to estimate the ATE. For a given data generating process, we perform R replications, where each replication yields a newly estimated ATE and either rejection or non-rejection of H_0. We estimate the actual rejection rate a of the particular test by $\widehat{a}$, the fraction of replications for which H_0 is rejected. This is an estimate of the true size of the test. With a finite number of replications, the simulation standard error is: $s_{\widehat{a}} = \sqrt{\widehat{a}(1-\widehat{a})/(R-1)}$. Ideally, we would like to find an approach for which the estimated size is close to the actual size.

Table A.4 in the Appendix presents the results from a Monte Carlo simulation with 1000 replications for different data generating processes. For the wild cluster bootstrap we use 500 replications. This lower value is justified, as the bootstrap simulation error cancels out across the Monte Carlo simulation. In the first row, the error term is assumed to be homoskedastic without any intracluster correlation. In the second to fourth rows, we assume positive and constant intracluster correlation coefficients of 0.5, 0.2 and 0.01 respectively. In the fifth to tenth rows, we assume an AR(1), i.e. autoregressive process of order 1, for the error term. This implies that the intracluster correlation decreases with an increasing distance between observations within a cluster. In rows five, six and nine it is always positive. In the other rows it alternates between negative and positive. In rows five to eight the correlation parameter in the AR(1) process is the same for every cluster, while it is different in rows nine and ten. Rows 11 to 20 add tutor fixed effects to the respective error process.

The White robust estimator only achieves a rejection rate of 0.05 if the data are not clustered (in rows 1 and 11). Otherwise it greatly over-rejects the true null hypothesis if there is a positive intracluster correlation and under-rejects it for processes with positive

and negative correlations between clustered observations. Its performance worsens when the degree of correlation increases. The Moulton estimator only achieves rejection rates close to the nominal size for positive intracluster correlation. Its rejection rates increase with higher correlation. This results in over-rejection when correlation is high (not shown here) and in under-rejection when there is no, or only a small, or negative correlation. Rejection rates decline in the presence of tutor fixed effects. The cluster-robust estimator over-rejects the null hypothesis of no treatment effect if there are no tutor fixed effects. This replicates results for a small number of clusters in Cameron, Gelbach, and Miller (2007). In the presence of tutor fixed effects, the cluster-robust estimator over-rejects when there is a high and positive correlation and under-rejects when there is a small or negative correlation. The wild cluster bootstrap achieves rejection rates close to the nominal size in the absence of tutor fixed effects. However, with tutor fixed effects its rejection rates decline: in particular, it severely under-rejects if correlation is small or alternating in sign. Finally, randomisation inference yields rejection rates as desired. Regardless of the form of correlation within clusters and regardless of tutor fixed effects, estimated size is never significantly different from nominal size. Thus, randomisation inference outperforms all other approaches in its size properties.

In order to study the power properties of these approaches, we add a treatment effect of $\beta = 0.25$ to the error term in the data generating process. This value is chosen because it is the point estimate in our results. We then estimate how often the tests reject the false hypothesis of a zero treatment effect. Table A.5 shows rejection rates for a significance level $\alpha = 0.05$. When there is no intracluster correlation and no tutor fixed effects, applying bootstrap and randomisation inference results in only small power losses. When there is intracluster correlation, approaches that take this into account are associated with lower power compared to the White estimator. Their power increases as the degree of correlation decreases. Power properties are better if there is positive and negative intracluster correlation. The presence of tutor fixed effects makes it difficult to detect the true positive effect for the Moulton, cluster-robust and bootstrap methods, while rejection rates for randomisation inference are not much affected.

In total, these results suggest that randomisation inference outperforms other approaches when observations within clusters are correlated and the randomisation is blocked with respect to covariates that explain outcomes. It achieves good size properties for all data generating processes considered here. This is arguably at the expense of reduced power, but is the only way to avoid over-rejections. The wild cluster bootstrap has comparable performance when there are no tutor fixed effects, but is otherwise inferior.

5 Results

Mean differences in test scores for treatment and control groups are shown in the second column of Table 1. On average, students exposed to the intervention achieved 0.25 point higher test scores. With a maximum score of 10 points, this corresponds to an increase of 2.5 percentage points. The next five columns show p-values for five different approaches (described above) to test whether this difference is equal to zero. The null hypothesis is rejected for all but one approach: when randomisation inference is applied, the average treatment effect is significant at the 5% level. Since our Monte Carlo simulations have suggested that randomisation inference is superior to the other approaches, we consider this as evidence that positive affirmation has significantly increased test performance.

When looking at subgroups, we notice that estimated ATEs are usually positive. Furthermore, they increase and become significant in some subgroups despite smaller sample sizes. This suggests considerable effect heterogeneity. On average, students who had a low mathematics grade in their first year achieved a test score 0.57 points higher when treated. Students with self-reported difficulties in mathematics achieved test score 0.8 points higher than their controls. Treatment effects are significant at the 10% level according to all five approaches considered. With regard to the preferred randomisation inference, they are even significant at the 1% level. Consistent with this finding, quantile treatment effects (not reported here) are positive for the lower part of the distribution and are usually zero for the upper part. This suggests that the intervention to non-cognitive skills has its strongest impact for students with below average maths skills compared to

Table 1: Estimated ATE and p-values for $H_0 : ATE = 0$

	N	ATE	p-values White	Moulton	Cluster-robust	BS	RI
all	440	0.25	0.17	0.28	0.27	0.32	0.05
female	129	0.33	0.38	0.41	0.40	0.42	0.35
male	305	0.14	0.49	0.56	0.57	0.60	0.45
$\leq$ 22 years of age	286	0.16	0.45	0.46	0.32	0.35	0.27
$>$ 22 years of age	151	0.55	0.12	0.19	0.16	0.22	0.16
business	261	0.19	0.42	0.54	0.52	0.54	0.14
low grade	174	0.57	0.05	0.09	0.06	0.09	0.00
high grade	205	-0.03	0.90	0.91	0.91	1.00	1.00
difficulties	149	0.80	0.02	0.06	0.03	0.10	0.00
no difficulties	266	0.06	0.75	0.76	0.74	0.76	0.76
prepared/revised	210	0.26	0.34	0.35	0.22	0.29	0.08
not prepared/revised	223	0.15	0.56	0.67	0.65	0.67	0.50

Note: ATE is the mean difference between treatment and control group. BS is an abbreviation for bootstrap; RI stands for randomisation inference. P-values are given for a robust White estimator (without clustering), a Moulton estimator (equation 7), a cluster-robust estimator (equation 8), a wild cluster bootstrap with 10000 replications (see Table A.3) and randomisation inference.

their peers. It has even stronger effects when students perceive a lack of maths skills. Of course, these two groups are overlapping. A decomposition in further subgroups limits inference as the sample is small. But mean comparisons suggest that the 97 students with self-reported difficulties and low maths grades achieve the highest gains from the intervention with an average treatment effect of 0.97. The 164 students with good maths grades and no difficulties are hardly affected by the intervention: their test score is only increased by 0.01 points. The 77 students with a low maths grade and no self-reported difficulties achieve test scores 0.21 points higher and the 34 students with high maths grade and self-reported difficulties achieve test scores 0.18 points higher when treated. There is also some evidence that the intervention had a significant effect for students who prepared or revised for the lecture. This could suggest that an intervention to non-cognitive skills might be more successful when knowledge has been acquired.

In order to raise efficiency, we also control for some covariates that are expected to explain the outcome variable, but which are not affected by the intervention. Table A.6

shows average treatment effects when gender, maths grade, and difficulties in maths are added in a linear regression. In other specifications, we also added tutor fixed effects and a dummy for the time of the tutorial: we obtained comparable effects. Due to controlling for covariates the estimated ATEs increase in every case. The ATE for all students is not only significant according to randomisation inference, but also according to the other approaches. We consider this as confirmation of our evidence, especially as our most preferred and second-choice approaches indicate a significant effect. As before, students with low maths grades and self-reported difficulties in maths are the subgroups with the highest and most significant treatment effects. For students who prepared or revised for the lecture, the wild cluster bootstrap suggests a significant effect at the 10% level, while randomisation inference no longer suggests a significant effect. Therefore, we are rather careful and argue that there is no strong evidence that this particular subgroup has significantly gained from the intervention. In contrast to the above results, we also find significant effects for female students when applying randomisation inference. But again we only consider this as weak evidence, since it is not found when not controlling for covariates. As a sensitivity check, we also estimate ATE in a semi-parametric matching approach. In contrast to the linear regression, it does not impose any functional form assumptions and allows for individual effect heterogeneity. Estimated effects from matching are different to the linear regression results as a different weighting scheme is applied. We do not report them here, but they are usually higher than ATEs in linear regression. This reassures us that potential misspecification in linear regression does not yield upward biased results.

6 Conclusion

In a field experiment, we examined whether a student's performance could be raised by a shock to their non-cognitive skills. The shock consisted of a positive affirmation intended to raise a student's motivation and self-confidence and to reduce test anxiety.

Students who were exposed to the intervention achieved higher average test scores than their controls. This was mainly due to positive quantile effects in the lower part of the

distribution. This is also reflected in effect heterogeneity with regard to different subgroups: students with a low maths grade and low self-assessed maths skills especially benefited from the intervention. These students are likely to be the students with the lowest self-confidence and highest test anxiety. No subgroup was harmed by the intervention. This suggests that teachers who aim to raise performance of their students can repeat this intervention without risk of harm to particular students. However, other research is necessary to assess whether results also hold for other populations, for other tests and in other contexts. Currently, evidence from other studies suggests that results generalise to pupils (Bouffard-Bouchard, Parent, and Larivee, 1991) and women under stereotype threat (Martens, Johns, Greenberg, and Schimel, 2006).

We also note that this specific intervention to non-cognitive skills (positive affirmation) might be more promising than introducing financial incentives when aiming to raise a student's test performance: extrinsic incentives were often found to be non-effective, are likely to have diverse (even negative) effects for different individuals and are more expensive to implement.

7 Appendix

Table A.1: German test instruction

for all:	*Bevor wir mit der Besprechung der Aufgaben beginnen, wird ein Mathematik-Selbsttest durchgeführt. Dieser Test ist anonym. Er beeinflusst Ihre Note nicht. Gegenstand des Testes sind genau jene mathematischen Methoden, die für Makro 2 wichtig sind. Falls Sie mit diesen Aufgaben Schwierigkeiten haben sollten, empfehlen wir den betreffenden Stoff selbständig zu wiederholen. Der Test hat zwei Ziele. Erstens dient er Ihnen dazu festzustellen, wo bei Ihnen ein Nachholbedarf besteht. Zweitens zeigt er uns Übungsleitern, bei welchen mathematischen Aufgaben Sie Schwierigkeiten haben. Sie haben 15 Minuten für die Beantwortung der Fragen Zeit. Danach wird der Test eingesammelt und von uns korrigiert. Sie erhalten Musterlösungen, um Ihren Kenntnisstand selbst festzustellen.*
for treated:	***Ich bin mir sicher, dass Sie die Aufgaben sehr gut lösen können. Sie haben ja auch in der Vergangenheit erfolgreiche Testergebnisse erzielt, sonst wären Sie ja nicht hier.***

Table A.2: Means of characteristics and test outcomes by treatment group

tutor	all	all	1	1	2	2	3	3
treatment	1	0	1	0	1	0	1	0
time 12am-2pm			1	0	1	0	0	1
observations	233	207	40	32	45	23	40	21
female	0.29	0.32	0.30	0.31	0.22	0.30	0.30	0.19
age	22.18	22.22	22.65	22.81	22.98	21.65	21.45	23.05
business	0.62	0.58	0.43	0.59	0.75	0.57	0.64	0.71
economics	0.17	0.21	0.13	0.16	0.11*	0.30	0.13	0.29
international affairs	0.12	0.12	0.28	0.16	0.09	0.04	0.15*	0.00
law	0.07	0.07	0.13	0.09	0.05	0.09	0.08	0.00
grade	4.75	4.81	4.70	4.65	4.84	4.73	4.84	4.93
difficulties	0.38	0.33	0.44	0.50	0.41	0.30	0.40	0.19
preparation in hours	0.77	0.74	0.93	0.87	0.73	0.43	0.64**	1.24
test score	6.27	6.01	5.85	5.19	5.96	5.91	6.50	6.10

tutor	4	4	5	5	6	6	7	7
treatment	1	0	1	0	1	0	1	0
time 12am-2pm	11	1	0	1	0	1	0	1
observations	45	38	18	38	28	41	16	14
female	0.31	0.42	0.47	0.26	0.29	0.22	0.25***	0.75
age	22.07	21.87	20.72**	22.21	22.41	22.24	22.13	21.38
business	0.64	0.47	0.47	0.58	0.75	0.63	0.63	0.50
economics	0.27	0.16	0.53***	0.18	0.07	0.20	0.13	0.40
international affairs	0.07	0.11	0.00*	0.18	0.11	0.17	0.00	0.10
law	0.02***	0.26	0.00	0.00	0.04	0.00	0.25*	0.00
grade	4.67	4.82	4.75	4.92	4.70	4.76	4.70	5.09
difficulties	0.41	0.38	0.18	0.32	0.31	0.32	0.40*	0.09
preparation in hours	0.69	0.92	0.72	0.66	0.90**	0.37	0.81	1.00
test score	6.73	6.47	6.72	6.29	6.03	6.07	6.19	5.79

Equality of means between groups is tested using t-tests, where standard errors are clustered according to cluster-robust variance matrix. Significant t-stats are indicated by *, **, *** for 10, 5 and 1% level.

Figure A.1: Distribution of the test scores

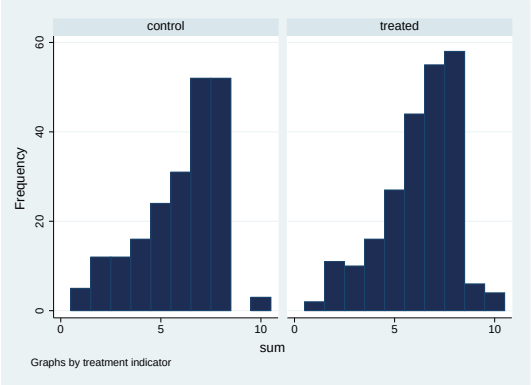

Table A.3: Implementation of the wild cluster bootstrap with H_0 imposed

step 1	In the original sample, estimate average treatment effect for the model $\boldsymbol{y}_c = \alpha + \beta \boldsymbol{D}_c + \boldsymbol{u}_c$ and obtain $\widehat{\beta}$. Form Wald statistic for $H_0 : \beta = 0$: $$w = \frac{\widehat{\beta}}{s_{\widehat{\beta}}},$$ where $s_{\widehat{\beta}}$ is the standard error using the cluster-robust variance estimator from equation 8.
step 2	In the original sample, estimate restricted model $\boldsymbol{y}_c = \alpha + \boldsymbol{u}_c$ that imposes $H_0 : \beta = 0$. Obtain restricted mean estimator $\widehat{\alpha}^R$ and the associated restricted residuals $\boldsymbol{u}_1^R, ..., \boldsymbol{u}_C^R$.
step 3	Do B iterations of this step. On the bth iteration:
step 3a	Form a pseudo sample of C clusters $(\widehat{\boldsymbol{y}}_1^*, \boldsymbol{D}_1), ..., (\widehat{\boldsymbol{y}}_C^*, \boldsymbol{D}_C)$ by the following method. For each cluster $c = 1, ..., C$, form either $\boldsymbol{u}_c^{R*} = \boldsymbol{u}_c^R$ with probability 0.5 or $\boldsymbol{u}_c^{R*} = -\boldsymbol{u}_c^R$ with probability 0.5. Then form $\widehat{\boldsymbol{y}}_c^* = \widehat{\alpha}^R + \widehat{\boldsymbol{u}}_c^*$. (Multiplication of residuals with 1 and -1 with probability 0.5 are so called Rademacher weights, which lead to asymptotic refinement if errors are symmetric distributed (Cameron, Gelbach, and Miller, 2007)).
step 3b	Calculate the Wald test statistic $$w_b^* = \frac{\widehat{\beta}_b^*}{s_{\widehat{\beta}_b^*}},$$ where $\widehat{\beta}_b^*$ and its standard error $s_{\widehat{\beta}_b^*}$ are obtained from the unrestricted model $\widehat{\boldsymbol{y}}_c^* = \alpha + \beta \boldsymbol{D}_c + \boldsymbol{u}_c$ in the bth pseudo sample, with $s_{\widehat{\beta}_b^*}$ computed using the cluster-robust variance matrix from equation 8.
step 4	Reject $H_0 : \beta = 0$ at level α if $w < w_{[\alpha/2]}^*$ or $w > w_{[1-\alpha/2]}^*$, where $w_{[q]}^*$ denotes the qth quantile of $w_1^*, ..., w_B^*$.

This implementation is suggested by Cameron, Gelbach, and Miller (2007).

Table A.4: Rejection rates under $H_0 : \beta = 0$ for $\alpha = 0.05$, 1000 simulations

$error_{jc}$	White	Moulton	Cluster-robust	Bootstrap	Randomisation
ϵ_{jc}	0.05	0.03	0.09	0.05	0.04
	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + \alpha_c$	0.64	0.05	0.08	0.05	0.05
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + 0.5 * \alpha_c$	0.48	0.04	0.10	0.06	0.04
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + 0.1 * \alpha_c$	0.10	0.04	0.09	0.06	0.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = 0.9 * \nu_{c(j-1)} + \epsilon_{jc}$	0.59	0.05	0.08	0.05	0.05
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = 0.5 * \nu_{c(j-1)} + \epsilon_{jc}$	0.24	0.03	0.08	0.05	0.06
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = -0.9 * \nu_{c(j-1)} + \epsilon_{jc}$	0.00	0.00	0.09	0.05	0.05
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = -0.5 * \nu_{c(j-1)} + \epsilon_{jc}$	0.00	0.00	0.09	0.05	0.05
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = \rho_c * \nu_{c(j-1)} + \epsilon_{jc}$	0.46	0.03	0.07	0.05	0.05
	(0.02)	(0.00)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = -\rho_c * \nu_{c(j-1)} + \epsilon_{jc}$	0.02	0.01	0.10	0.06	0.06
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
tutor fixed effects					
$\epsilon_{jc} + t$	0.05	0.00	0.00	0.00	0.04
	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
$\epsilon_{jc} + \alpha_c + t$	0.64	0.03	0.08	0.04	0.05
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + 0.5 * \alpha_c + t$	0.47	0.02	0.05	0.03	0.05
	(0.02)	(0.00)	(0.01)	(0.00)	(0.01)
$\epsilon_{jc} + 0.1 * \alpha_c + t$	0.09	0.00	0.00	0.00	0.05
	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
$\nu_{cj} = 0.9 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.59	0.05	0.08	0.05	0.05
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = 0.5 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.24	0.01	0.03	0.02	0.05
	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
$\nu_{cj} = -0.9 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.00	0.00	0.02	0.01	0.04
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
$\nu_{cj} = -0.5 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.00	0.00	0.00	0.00	0.05
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
$\nu_{cj} = \rho_c * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.46	0.02	0.04	0.02	0.05
	(0.02)	(0.00)	(0.01)	(0.00)	(0.01)
$\nu_{cj} = -\rho_c * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.02	0.00	0.05	0.03	0.06
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)

Standard errors in parentheses. Data generating process is given in equation 9.
$\epsilon_{jc} \sim N(0,1)$, $\alpha_c \sim N(0,1)$, $\nu_{c1} \sim N(0,1)$, $\rho_c \sim U(0,1)$
tutor fixed effects: $t = t_1 + t_2 + t_3 + t_4 + t_5 + t_6 + t_7$ where $t_i \sim U(0,1)$

Table A.5: Rejection rates assuming $\beta = 0.25$ for $\alpha = 0.05$, 1000 simulations

$error_{jc}$	White	Moulton	Cluster-robust	Bootstrap	Randomisation
ϵ_{jc}	0.71	0.54	0.76	0.66	0.60
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
$\epsilon_{jc} + \alpha_c$	0.69	0.07	0.12	0.07	0.06
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + 0.5 * \alpha_c$	0.64	0.11	0.19	0.13	0.11
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + 0.1 * \alpha_c$	0.68	0.44	0.65	0.54	0.49
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
$\nu_{cj} = 0.9 * \nu_{c(j-1)} + \epsilon_{jc}$	0.61	0.05	0.10	0.06	0.06
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = 0.5 * \nu_{c(j-1)} + \epsilon_{jc}$	0.56	0.16	0.34	0.23	0.23
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = -0.9 * \nu_{c(j-1)} + \epsilon_{jc}$	0.10	0.09	0.48	0.36	0.32
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)
$\nu_{cj} = -0.5 * \nu_{c(j-1)} + \epsilon_{jc}$	0.65	0.60	0.91	0.86	0.82
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = \rho_c * \nu_{c(j-1)} + \epsilon_{jc}$	0.55	0.08	0.16	0.13	0.13
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = -\rho_c * \nu_{c(j-1)} + \epsilon_{jc}$	0.44	0.39	0.66	0.57	0.54
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)
tutor fixed effects					
$\epsilon_{jc} + t$	0.70	0.06	0.18	0.10	0.60
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$\epsilon_{jc} + \alpha_c + t$	0.70	0.06	0.12	0.07	0.07
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + 0.5 * \alpha_c + t$	0.65	0.08	0.15	0.10	0.14
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\epsilon_{jc} + 0.1 * \alpha_c + t$	0.79	0.05	0.18	0.10	0.53
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$\nu_{cj} = 0.9 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.61	0.04	0.09	0.06	0.07
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = 0.5 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.62	0.10	0.23	0.14	0.27
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = -0.9 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.15	0.09	0.27	0.17	0.38
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$\nu_{cj} = -0.5 * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.75	0.14	0.38	0.19	0.84
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
$\nu_{cj} = \rho_c * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.60	0.04	0.12	0.06	0.17
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
$\nu_{cj} = -\rho_c * \nu_{c(j-1)} + \epsilon_{jc} + t$	0.35	0.20	0.44	0.34	0.46
	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)

Standard errors in parentheses. Data generating process is given in equation 9.
$\epsilon_{jc} \sim N(0,1)$, $\alpha_c \sim N(0,1)$, $\nu_{c1} \sim N(0,1)$, $\rho_c \sim U(0,1)$
tutor fixed effects: $t = t_1 + t_2 + t_3 + t_4 + t_5 + t_6 + t_7$ where $t_i \sim U(0,1)$

Table A.6: Estimated ATE controlling for covariates and p-values for $H_0 : ATE = 0$

	N	ATE	p-values				
			White	Moulton	Cluster-robust	BS	RI
all	440	0.34	0.03	0.10	0.07	0.10	0.02
female	129	0.57	0.09	0.12	0.07	0.11	0.03
male	305	0.22	0.22	0.30	0.25	0.30	0.28
$\leq$ 22 years of age	286	0.19	0.30	0.33	0.20	0.24	0.10
$>$ 22 years of age	151	0.71	0.02	0.07	0.06	0.11	0.16
business	261	0.24	0.27	0.41	0.42	0.46	0.23
low grade	174	0.63	0.02	0.04	0.02	0.05	0.02
high grade	205	0.00	0.99	0.99	0.99	0.99	0.99
difficulties	149	0.94	0.00	0.02	0.01	0.05	0.00
no difficulties	266	0.06	0.72	0.73	0.70	0.72	0.74
prepared/revised	210	0.38	0.09	0.11	0.03	0.07	0.13
not prepared/revised	223	0.20	0.38	0.52	0.50	0.54	0.52

ATE is the coefficient for the treatment indicator in a linear regression with gender, maths grades and difficulties in maths as additional covariates. BS is an abbreviation for bootstrap; RI stands for randomisation inference. P-values are given for a robust White estimator (without clustering), a Moulton estimator (equation 7), a cluster-robust estimator (equation 8), a wild cluster bootstrap with 10000 replications (see Table A.3) and randomisation inference.

References

Bandura, A. (1993): "Perceived Self-Efficacy in Cognitive Development and Functioning," *Educational Psychologist*, 28(117-148).

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 119, 249–275.

Bloom, H. (2004): "Randomizing Groups to Evaluate Place-Based Programs," in *Learning More From Social Experiments: Evolving Analytic Approaches*. Russell Sage Foundation.

Borghans, L., A. Duckworth, J. Heckman, and B. ter Weel (2008): "The Economics and Psychology of Personality Traits," *Journal of Human Resources*, 43, forthcoming.

Borghans, L., H. Meijers, and B. T. Weel (2008): "The Role Of Noncognitive Skills In Explaining Cognitive Test Scores," *Economic Inquiry*, 46(1), 2–12.

Bouffard-Bouchard, T., S. Parent, and S. Larivee (1991): "Influence of Self-Efficacy on Self-Regulation and Performance among Junior and Senior High-School Age Students," *International Journal of Behavioral Development*, 14, 153–164.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2007): "Bootstrap-Based Improvements for Inference with Clustered Errors," NBER Technical Working Papers 0344, National Bureau of Economic Research, Inc.

Cox, D., and N. Reid (2002): *Theory of the Design of Experiments.* London: Chapman and Hall.

Cunha, F., J. J. Heckman, and L. Lochner (2006): *Interpreting the Evidence on Life Cycle Skill Formation* vol. 1 of *Handbook of the Economics of Education*, chap. 12, pp. 697–812. Elsevier.

Deci, E., and A. Moller (2005): *The Concept of Competence: A Starting Place for Understanding Intrinsic Motivation and Self-Determined Extrinsic Motivation* pp. 579–597. Handbook of Competence and Motivation, Guilford Press.

Donner, A., and N. Klar (2004): "Pitfalls of and Controversies in Cluster Randomization Trials," *American Journal of Public Health*, 94, 416–422.

Duckworth, A., and M. Seligman (2005): "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents," *Psychological Science*, 16, 939–44.

Duflo, E., R. Glennerster, and M. Kremer (2006): "Using Randomization in Development Economics Research: A Toolkit," NBER Technical Working Papers 0333, National Bureau of Economic Research, Inc.

Fisher, R. (1935): *Design of Experiments.* Oliver and Boyd:Edinburgh.

Heckman, J., and D. Masterov (2007): "The Productivity Argument for Investing in Young Children," *Review of Agricultural Economics*, 29, 446–93.

Heckman, J., J. Stixrud, and S. Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 23(3), 411–82.

Heckman, J. J., and J. A. Smith (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9(2), 85–110.

Hembree, R. (1988): "Correlates, Causes, Effects, and Treatment of Test Anxiety," *Review of Educational Research*, 58, 47–77.

——— (1990): "The Nature, Effects, and Relief of Mathematics Anxiety," *Journal for Research in Mathematics Education*, 21, 33–46.

Kézdi, G. (2004): "Robust Standard Error Estimation in Fixed-Effects Models," *Hungarian Statistical Review*, 9, 95–116.

Martens, A., M. Johns, J. Greenberg, and J. Schimel (2006): "Combating stereotype threat: The effect of self-affirmation on women's intellectual performance," *Journal of Experimental Social Psychology*, 42, 236–243.

Moulton, B. (1986): "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 32, 385–397.

Rosenbaum, P. (2002): "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*, 17, 286–304.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in randomized and nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Steele, C. (1997): "A Threat in the Air - How Stereotypes Shape Intellectual Identity and Performance," *American Psychologist*, 52, 613–629.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

Wu, C. (1986): "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1261–1295.

Zigler, E., and E. Butterfield (1968): "Motivational Aspects of Changes in IQ Test Performance of Culturally Deprived Nursery School Children," *Child Development*, 39, 1–14.