Brunello, Giorgio; Crema, Angela; Rocco, Lorenzo

**Working Paper**

# Some unpleasant consequences of testing at length

GLO Discussion Paper, No. 286

**Provided in Cooperation with:**
Global Labor Organization (GLO)

*Suggested Citation:* Brunello, Giorgio; Crema, Angela; Rocco, Lorenzo (2018) : Some unpleasant consequences of testing at length, GLO Discussion Paper, No. 286, Global Labor Organization (GLO), Maastricht

This Version is available at:
https://hdl.handle.net/10419/188928

# Some unpleasant consequences of testing at length*

Giorgio Brunello (Padova and IZA)

Angela Crema (New York University)

Lorenzo Rocco (Padova, GLO)[+]

Abstract:

Using Italian data on standardized test scores, we show that the performance decline associated with question position is heterogeneous across students. This fact implies that the rank of individuals and classes depends on the length of the test. Longer tests may also exhibit larger gaps between the variance of test scores and the variance of underlying ability. The performance decline is correlated with both cognitive and non-cognitive abilities and there is also evidence that those with better parental background experience a smaller decline than those with poorer background. Therefore, the gap between the two groups widens in longer tests.

Keywords: low stake tests; position of questions; cognitive and non-cognitive skills; Italy

JEL codes: I21

Acknowledgements: we are grateful to Marco Bertoni, David Kiss, Olivier Marie and the audiences at talks in Bari, Florence, Padova and Seville for comments. A special thanks to Patrizia Falzetti and INVALSI for help with the data. The usual disclaimer applies.

* A previous version of this work circulated under the title "Testing at length if it is cognitive or non-cognitive"

[+] corresponding author: Lorenzo Rocco, Department of Economics and Management, University of Padova – via del Santo, 33 – 35123 Padova (Italy) – email: lorenzo.rocco@unipd.it

*Introduction*

Attained scores in standardized tests taken by students are often used to measure cognitive skills and the quality of schools. When these tests are low stake, however, as in PISA, TIMSS or PIRLS,[1] students may have limited motivation to perform well (Zamarro et al. 2017). Because of insufficient motivation and concentration, their performance typically declines as the test proceeds.

The decline of performance with the position of the question in the test has been recently studied by Borghans and Schils, 2012, who have attributed it to non-cognitive factors, including motivation, conscientiousness and competitive attitude: conditional on cognitive skills, students better endowed with these factors are more likely to perform better as the test proceeds.

When the probability of answering correctly a test question depends on the position of the question, the expected test score – or the expected percentage of correct answers – is a function of the number of questions, or test length. If this probability varies across pupils, and the composition of pupils varies within and between schools, increasing test length not only reduces measurement error (Jacob, 2016) but can also affect the rankings of students, classes and schools.

In the absence of heterogeneous responses to question position, a test long enough to minimize the impact of noise drives to zero the gap between the variance of test scores and the variance of underlying ability. Heterogeneous responses, however, introduce a wedge between the two variances, which could increase with the length of the test. When the gap is an important ingredient in the definition of optimal test length, the unpleasant consequence is that failure to recognize this heterogeneity may lead to longer than optimal tests.

In this paper, we show that the heterogeneous responses of pupils to question position affect both the mean and the distribution of test scores. We use

[1] PISA is the OECD Program for International Student Assessment; TIMSS and PIRLS are for Trends in International Mathematics and Science Studies and Progress in International Reading Literacy Study, both run by Boston College.

administrative data on Italian standardized tests. Similar to PISA, these tests are low stakes, and are run every year on the universe of Italian schools by the Italian agency INVALSI. As in PISA, booklets where questions appear in different orders are randomly allocated to students. We use this variation across booklets to distinguish the effects of question position from those associated to question difficulty. We focus on the math scores of fifth graders in primary school and combine detailed information on the answers to each test question with administrative school data and questionnaires compiled by students and teachers.

Two distinctive features of INVALSI data provide a cleaner setting than PISA for the purposes at hand. First, all INVALSI booklets include the same questions, contrary to PISA booklets, which differ in length and include different (although comparable) portions of the entire test. Second, the change of position across booklets involves single questions in the INVALSI test and clusters of consecutive questions in PISA.

The random allocation of booklets implies that individual characteristics are independent of the position of questions. We exploit this feature to estimate a mixed model with random individual – specific intercepts and slopes.[2] We confirm that, on average, a higher question position reduces test scores. In particular, we estimate that the probability of giving a correct answer to a question is reduced by 0.6 percentage points when that question is moved forward by ten positions in the questionnaire.

We document that, while the average effect of changing the position of a question is moderate, its variance is significant: the decline effect associated to moving forward a question by ten positions is as large as -1.7 at the $10^{th}$ percentile and equal to 0.5 percent at the $90^{th}$ percentile of the distribution of

[2]In the economic literature, mixed models have been used unfrequently because the independence assumption is typically hard to justify

estimated individual effects. Furthermore, there is no decline in performance as the position of questions increases for about one pupil out of four (26 percent).

The negative effect of position on performance is smaller for girls (estimated effect: -0.035) than for boys (-0.083), and so is the range between the 10$^{th}$ and the 90$^{th}$ percentile ([-0.12, +0.06] for females and [-0.22, +0.05] for males). These estimates indicate that performance increases with question position for 32 percent of girls and 24 percent of boys. Therefore, the eventual initial performance gender gap in favor of boys is likely to be partially or entirely compensated by girls in longer tests. Since the composition of pupils varies among classes, the estimated heterogeneity implies that increasing the test length from the current 46 questions to 60 (resp. 70) would change rank by at least five places in either direction for about 22 (resp. 40) percent of classes.

We compare the gap between the variance of test scores and the variance of underlying ability when the responses to question position are either homogeneous and heterogeneous. While this gap declines monotonously in the former case, in the latter case it reaches a minimum when length is 38 questions – close to the 46 questions in the INVALSI test - and increases for longer tests. When length is 67 questions, the gap is as large as at 20 questions, and the gain in terms of a lower contribution of noise is entirely offset by the heterogeneity of responses to question length.

We investigate whether the responsiveness to question position varies with measures of ability (both cognitive and non-cognitive), school and family background variables and find that higher math grades in the semester prior to the test, higher conscientious and self-confidence, a better parental environment, being female or native and enrolment in classes that are more socially inclusive, smaller and where teachers also drill students in the test are positively correlated to experiencing a smaller decline in performance as the test proceeds.

Our findings have two implications. First, the view that the negative relationship between performance and question position depends exclusively on non-

cognitive abilities - as suggested by Borghans and Schils, 2012, – does not hold in our data. Second, another unpleasant feature of longer tests is that they increase the gap in test scores between those with better family conditions and those with a more disadvantaged endowment (immigrants with few books in the house). A lower bound estimate suggests that this gap increases by 0.04 (resp. 0.07) standard deviations when the number of questions in the test rises from the current 46 to 60 (resp. 70).

The paper is organized as follows. Section 1 reviews the literature, Section 2 introduces the data and Section 3 sets up the empirical model. Results are presented in Section 4 and some of our assumptions are discussed in Section 5. Conclusions follow.

*1. Literature Review*

Three areas of economic and psychological literature are relevant for this paper: the first explores the effect of test length on performance; the second tries to disentangle the contribution of cognitive ability and personality traits to test scores, and the last considers the effects of non-cognitive skills on cognitive test scores.

A concept often used to rationalize the (rather intuitive) idea that performance declines as test length increases is ego depletion (Borgonovi and Biecek, 2016): acts of self-control draw from a common, global resource that is limited and vulnerable to become depleted as individuals exercise acts of self-control. Personality traits such as fluid intelligence, anxiety, and attitudes toward learning might work as moderators of ego depletion (see Ackerman and Kanfer, 2009, and Hagger et al. 2010 as references).

In the economic literature, the negative correlation between the likelihood of getting an answer correct and the position of the question has been exploited to distinguish between two factors affecting student performance: cognitive skills and personality traits (see Borghans and Schils, 2012). Balart et al., 2018, apply this approach to decompose PISA test scores into a cognitive component, the

starting performance, and a non-cognitive component, the decline effect during the test, and show that both components contribute to economic growth in a sample of countries. Balart and Oosterveen, 2018, adopt a similar strategy to show that longer tests decrease the gender math gap,[3] and Battaglia and Hidalgo, 2018, evaluate for Spain the impact of a remedial policy on performance decline, which they treat by as a measure of non-cognitive skills.

Borgonovi and Biecek, 2016, also use PISA and interpret the decline in student performance over the test as a measure of academic endurance, defined as the ability to maintain the baseline rate of successful test completion throughout the assessment. Their findings suggest that girls and socio-economically advantaged students have on average higher levels of endurance than males and pupils with a low socioeconomic background, respectively. They also observe that endurance tends to be positively associated with initial performance: "…will and skill are not orthogonal but are positively associated because high-achieving students tend to spend less self-regulatory capacities to maintain concentration and focus; they have higher task value and expected performance because of greater self-beliefs" (p. 135).

There is a growing awareness that cognitive test scores reflect not only ability, knowledge, and intelligence but also personality traits, motivation, grit and self-control[4]. Test takers may not exert maximal effort. When tests are low stakes, as in the OECD PISA project, some individuals try harder than others (see Zamarro et al. 2017, Duckworth et al, 2011). Scores can also be improved by offering a reward (see Borghans et al. 2008; Segal, 2012). Since test scores

[3]Rodríguez-Planas and Nollenberger, 2018, provide evidence that second-generation girls whose parents come from more gender-equal countries outperform their male counterparts in reading, science, and math. Using the method first suggested by Borghans and Schils, they show that this finding is driven by cognitive – rather than non-cognitive – skills.

[4]Farrington et al., 2012, rationalize the existing literature on non-cognitive skills and school performance by identifying five general categories of non-cognitive factors: academic behaviours, academic perseverance, academic mind-sets, learning strategies, and social skills.

reflect differences in individual motivation[5] and not just differences in abilities, ranking countries based on average low-stakes assessments is problematic (see Gneezy et al., 2017).

*2. The data*

Our data are drawn from the administrative records of INVALSI, the Italian agency in charge of standardized tests in schools. INVALSI kindly provided the necessary information on the question order faced by each student, which is not available in the public data files. We focus on the 2015 math test taken by primary school fifth graders. Compared to second graders, who are also tested, fifth graders after the test are administered a questionnaire, which collects information on parental background, school environment and non-cognitive skills.

The test consists of 46 questions listed in five booklets. Differently from PISA, booklets in the INVALSI tests contain the same questions, but in different order. In particular, 18 questions out of 46 change positon across booklets.[6] In our empirical analysis, we consider only these questions. By so doing, we are able to distinguish the effect of position from question-specific fixed effects.

The distribution of booklets to students taking the test is designed to avoid that adjacent students receive the same booklet. To reduce the risk that booklets are not distributed as prescribed, we only consider the sub-sample of schools and classes randomly selected by INVALSI to have an external examiner supervise

[5] There is a distinction in the economic and psychometric literature between extrinsic and intrinsic motivation, defined as contingent rewards and the desire to perform a task for its own sake, respectively (Bénabou and Tirole, 2003). Motivation scales such as the Academic Motivation Scale (AMS) identify a spectrum of motivation types increasing in internationalization and quality, ranging from a-motivation to intrinsic motivation, passing through different forms of extrinsic motivation (Utvær and Haugan, 2016).

[6] More in detail, 8 questions take 4 different positions, 8 take 3 different positions and 2 take 2 different positions.

the test and prevent the extensive cheating documented, among others, by Bertoni et al., 2013, and Angrist et al., 2017.[7]

The final sample consists of 19,656 pupils distributed in about 1100 classes. Table 1 presents the summary statistics of the variables used in our empirical analysis. The outcome variable Y takes the value 0 if the answer is wrong (or skipped) and 100 if the answer is correct. Its sample average is 52.81. Females are 48.8 percent of the sample, and average age (in months) is somewhat above 10 years; the average share of immigrants is 10 percent and more than 35 percent of pupils have less than 26 books at home; about 35 percent are in classes with less than 20 pupils and more than 48 percent are regularly drilled by teachers using tests similar to those administered by INVALSI.

A broadly accepted taxonomy of personality traits is the Five – Factor Model (FF). According to the definition by Nyhus and Pons, 2005, this model includes the following factors: agreeableness, conscientiousness, emotional stability, extraversion and autonomy. We use the questionnaire administered to pupils at the end of the test to generate two indicators of emotional stability (neuroticism and confidence) and two variables capturing agreeableness and conscientiousness. In our data, neuroticism measures worry and anxiety before and during the test, confidence captures self-esteem with respect to math skills, agreeableness refers to the ability to interact with and help classmates, and conscientiousness measures the ability to concentrate and complete assigned tasks.

We add to these measures two indicators of motivation (intrinsic and extrinsic), using information on whether school behaviour is driven by internal and external rewards. Finally, we proxy the quality of social relations in class with an indicator of bully victimization, based on self-reported information on being the target of threats, intimidation and physical violence. As described more in

[7] INVALSI has developed an algorithm showing no cheating (by students or teachers) in the schools / classes with the external examiner.

detail in the Appendix, each indicator is obtained using principal components analysis.

A natural candidate to capture cognitive skills is the math grade attained in the semester before the test.[8] This grade, however, is likely to reflect also non-cognitive skills. As shown by Cornwell et al., 2013, teachers are influenced by the non-cognitive skills of students when assigning grades. Cunha and Heckman, 2008, 2010, and Borghans et al., 2008, have also documented that non-cognitive abilities concur to shape the development of cognitive abilities.

In order to purge the influence of non-cognitive factors from individual math grades, we regress them on our indicators of non-cognitive skills, age, gender and class dummies, take the residuals and define the dummy "High math grade" (HG) as equal to 1 when residuals are above the median and to 0 otherwise.

The randomization of booklets to students implies independence of individual characteristics. To verify whether independence holds, we run balancing tests by regressing the individual variables in Table 1 on booklet dummies. As shown in Table 2, we never reject the hypothesis that the coefficients associated to each booklet are equal, which supports randomization.

*3. The Empirical Model*

Consider a standardized test with N questions. The position P of each question, from 0 to N-1, varies across booklets, and these booklets are randomly assigned to students. The relationship between the answer to each question Y (correct or wrong) – and its position P is

$$Y_{iq} = \theta_i + \beta_i P_{iq} + \sum_{q=1}^{Q} \mu_q Q_q + \varepsilon_{iq} \qquad (1)$$

where the indices $i$ and $q$ indicate the student and the question respectively; Q is a question fixed effect; ε is the noise of the test; θ is the individual - specific intercept and β is the individual - specific slope parameter associated to the

[8] Math grades range from 4 (bottom) to 10 (top), with grades under 6 being considered below the passing line. In our sample, the average grade is just below 8 (see Table 1).

position of a question. We further assume that θ and β are jointly normally distributed, with means α and δ and variance-covariance matrix Σ. The random noise ε follows a zero-mean normal distribution independent of θ and β. Finally we posit that $\sum_1^Q \mu_q Q_q = 0$, a normalization.

Since we have set *P=0* for the first question and we control for question difficulty, parameter $\theta_i$ captures the answer to the first question and can be interpreted as an indicator of individual ability. Parameter $\beta_i$ is instead the marginal effect of P on Y, which can be positive or negative. The fact that booklets and question positions are randomly allocated implies that these parameters are independent of $P_{iq}$. Notice that parameter $\beta_i$ varies across individuals but not across questions. We shall discuss this assumption, commonly adopted in this literature, as well as the normality of $\theta_i$ and $\beta_i$ in Section 5.

The independence of $\theta_i$ and $\beta_i$ of $P_{iq}$ satisfies the requirements of mixed models, a class of models which accommodates random intercepts and slopes and includes multi-level models as a special case. Eq. (1) is a two-level model, with questions representing the first and students representing the second level. We can re-write it as $Y_{iqp} = \alpha + \delta P_{iq} + \sum_{q=1}^{Q} \mu_q Q_q + (u_{0i} + u_{1i} P_{iq} + \varepsilon_{iqp})$ , where $\theta_i = \alpha + u_{0i}$ and $\beta_i = \delta + u_{1i}$, and the random variables $u_{0i}$ and $u_{1i}$ have a zero mean normal distribution with variance-covariance matrix Σ. Both variables are independent of $\varepsilon_{iq}$.

On the one hand, the random component of (1) - $u_{0i} + u_{1i} P_{iq} + \varepsilon_{iqp}$ - follows a normal distribution with parameters that can be estimated using maximum likelihood. On the other hand, the fixed component - $\alpha + \delta P_{iq} + \sum_{q=1}^{Q} \mu_q Q_q -$ can be estimated using generalised least squares (see Cameron and Trivedi, 2006 ch. 22.8). The estimated parameters can be used to compute for each student the best linear unbiased predictions of the intercept $\theta_i$ and the slope $\beta_i$.

Eq. (1) implies that the individual test score S (defined as the proportion of correct answers) in a test of length N is given by

$$S_i = \theta_i + \beta_i \frac{N-1}{2} + \varepsilon_i \quad (2)$$

where $S_i = \frac{1}{N}\sum_1^N Y_{iq}$ and $\varepsilon_i = \frac{1}{N}\sum_1^N \varepsilon_{iq}$. A unitary increase in test length N changes the expected test score by $\frac{1}{2}\beta_i$. Therefore, if $\beta$ varies among individuals and the composition of individuals varies across schools, changes in test length can affect the ranking of individuals and schools.

The variance of the score – in a class, grade or school – is given by

$$V(S_i) = V(\theta_i) + \left(\frac{N-1}{2}\right)^2 V(\beta_i) + (N-1)Cov(\theta_i, \beta_i) + \frac{\sigma_\varepsilon}{N} \quad (3)$$

and can be computed using the estimates of the mixed model. When $\beta$ does not vary across individuals, this variance tends to $V(\theta_i)$ as N increases and the noise of the test goes to zero, and can be interpreted as a measure of the dispersion of ability across students, classes and schools.

However, when $\beta$ is individual-specific, the gap between $Var(S_i)$ and $Var(\theta_i)$ does not go to zero even when the noise is negligible. Increasing $N$ has two contrasting effects on $Var(S_i)$: on the one hand, it attenuates noise; on the other hand, it magnifies the impact of the variance of $\beta$. If $Cov(\theta_i, \beta_i)$ is positive, $Var(S_i)$ systematically overestimates the true dispersion of ability.

We investigate the determinants of the probability of giving a correct answer to the first question and of the decline in performance as the question position increases by estimating a parametric model where $\theta$ and $\beta$ are modeled as linear combinations of cognitive traits $X_c$, non-cognitive traits $X_{nc}$, and class and family characteristics E. We assume that $\theta_i = X_{ic}^{'}\pi_c + X_{inc}^{'}\pi_{nc} + E_i^{'}\pi_e$ and $\beta_i = X_{ic}^{'}\gamma_c + X_{inc}^{'}\gamma_{nc} + E'\gamma_e$ and we re-write (1) as

$$Y_{iqp} = X_{ic}'\pi_c + X_{inc}'\pi_{nc} + Z_i'\pi_e + P_{iq}X_{ic}'\gamma_c + P_{iq}X_{inc}'\gamma_{nc} + P_{iq}E_i'\gamma_e + \sum_{q=1}^{Q}\mu_q Q_q + \varepsilon_{iqp} \quad (4)$$

The independence of $P_{iq}$ with respect to *X, E,* and ε implies that parameters $\gamma$ can be consistently estimated using ordinary least squares (see Nizalova and Murtazashvili, 2016). Under the additional assumption that – conditional on E - X and ε are uncorrelated, the estimates of $\pi$ are unbiased too. [9]

*4. Results*

Table 3 reports the estimates of the two-level model (1) for the full sample and separately by gender. We confirm that, on average, performance declines with the position of questions. The average marginal estimated effect, $E(\beta_i)$, is equal to -0.060 in the full sample, and to -0.083 and -0.035 for males and females respectively. The variance of $\beta_i$ is statistically different from zero and equal to 0.035 in the full sample, to 0.041 and 0.029 for males and females. The covariance between the two random effects $\theta$ and $\beta$ is also positive and statistically significant, and the implied correlation is equal to 0.20 in the full sample and to 0.27 and 0.16 for males and females.

We illustrate the heterogeneity of $\theta$ and $\beta$ in our data by plotting in Figure 1 their best linear unbiased predictions. While $E(\beta_i)$ is negative, individual $\beta$ turns out to be positive for close to 26 percent of the sample.[10] Table 4 shows how the average values of $\theta$ and $\beta$ vary across individuals with different background – measured either by the number of books at home or by immigrant status. Typically, a less privileged background is associated to a lower average value of $\theta$ and to a higher average absolute value of $\beta$.

[9] We verify whether our OLS estimates are sensitive to the omission of un-observables using the tests proposed by Oster, 2017. The test establishes bounds to the true value of the parameters under two polar cases. In the first case, there are no un-observables and parameters are consistently estimated. In the second case, there are un-observables, but observables and un-observables are equally related to the treatment. If zero can be excluded from the bounding set, then accounting for un-observables would not change the direction of our estimates. We find that this is always the case in the current setup. Detailed results are available from the authors upon request.

[10] The assumption that β is normally distributed requires that some β are positive but is silent on the share of positive β. We return to this assumption in Section 5 of the paper.

The correlation between random intercepts and slopes in our data is about five times as large as the one found by Borghans and Schils, 2012 (0.043). A positive covariance can be interpreted as suggesting that individuals with higher values of ability/skills $\theta$ are more likely to experiment either a lower decline of performance as the position of questions increases or even an increase in performance. As the size of the test $N$ increases, a positive covariance amplifies the differences in test scores across students.

The observed heterogeneity in the relationship between performance and the position of questions implies that individual differences in the expected test score $S$ vary with test length $N$. Consider for instance two hypothetical pupils, a male and a female, with initial performance equal to the average gender–specific value of $\theta$ and with the associated value of $\beta$.[11] As shown in Table 5, while the female pupil starts with a lower score (84.1 versus 86.2), she overtakes the male pupil by N=40 (82.3 versus 81.8).[12]

With homogeneous responses to the position of questions, the gap between the variance of test scores and the variance of $\theta$ declines monotonously as the length of the test increases and the variance of noise falls. Heterogeneous responses, however, introduces additional elements to the gap, which may partially or entirely compensate the effect of length on the variance of noise. Using the estimates in Table 3, we show in Figure 3 the gap $Var(S_i)$-$Var(\theta_i)$ both with and without heterogeneous responses to the position of questions. While the latter declines monotonously, the former declines to reach a minimum just before $N$=40 and increases again afterwards, indicating that the actual test length of INVALSI test ($N$=46) is just above the value which minimizes the gap. When the gap $Var(S_i)$-$Var(\theta_i)$ is an ingredient in the definition of optimal test length, these results suggest that – everything else equal - failure to consider the

[11] We define the associated value of β as the average value for individuals having θ within a small interval of average gender specific θ.

[12] This finding confirms for a different environment (primary schools) and dataset (INVALSI for Italy rather than PISA) the results reported by Balart and Oosterveen, 2018, who show that the average gender gap in test scores declines with the length of the test.

heterogeneous response of students to the position of questions is likely to produce longer than optimal tests.

We investigate the determinants of the variability of $\beta$ and $\theta$ by estimating Eq. (4), using the dummy HG to proxy cognitive abilities, the personality traits discussed in Section 2 as measures of non-cognitive abilities, gender, the number of books at home and immigration status as measures of parental background, class size, exposure to training to the INVALSI test and an index of bully victimization as characteristics of the educational environment. Our results are shown in Table 6. The first column reports the effects of the covariates on the intercept θ and the second column the effects on the slope β. The regression includes class and question fixed effects, age, attendance of kindergarten and of childcare facilities, and age at enrolment in primary school.[13]

We find that the dummy HG has a strong positive effect on both the intercept and the slope. We estimate that switching from HG=0 to HG=1 increases the probability that the first question is correctly answered by 25.5 percent (13.46/52.81) with respect to the mean and reduces the decline in performance as the test proceeds by 18.5 percent (0.022/0.119). While all personality traits but intrinsic motivation have a significant effect on $\theta$, only conscientiousness and self-confidence influence $\beta$ in a statistically significant way, and reduce the estimated decline.[14] To illustrate, switching from the 25$^{th}$ percentile (-0.750) to the 75$^{th}$ percentile (1.168) of the confidence index increases initial performance by 18 percent (4.92*(1.168+0.750)/52.81). Performance decline for students with a conscientiousness index equal to the 75 percentile (+0.91) is 24 percent smaller than for students at the 25 percentile (-0.90).

[13] We deal with missing values by adding to the regressions missing value dummies.
[14] Notice, however, that when we test whether the effects of all personality traits on $\beta$ are jointly significant, we reject the null of no joint significance.

Furthermore, there is evidence that the number of books at home – our key indicator of parental background - affects significantly both $\theta$ and $\beta$. In particular, the decline of performance with question position is lower by about one quarter among pupils with more than 26 books at home. Conversely, immigrant status negatively affects $\theta$ and aggravates performance decline. This decline is significantly smaller for pupils in small classes who have been drilled by the teacher using material similar to the test.[15] There is also evidence that the index of poor social relations (bully victimization) negatively effects both $\theta$ and $\beta$.[16]

Overall, our results suggest that both the answer to the first question and the decline of performance after the first question depend on many factors, in contrast to the view – proposed by Borghans and Schils (2012) – that the former is a measure of cognitive skills and the latter an indicator of non-cognitive skills.

To further illustrate the effect of test length on the distribution of test scores, we focus on the factors which are significantly correlated with $\beta$ and define two groups of students, the "highly-endowed" and the "poorly-endowed". The former (latter) are students with HG=1 (0), whose conscientiousness and confidence is above (below) the median, have more (less) than 26 books at home, are natives (immigrants), have (not) been trained to the INVALSI test and have experienced low (high) levels of bully victimization. For these two groups, we predict individual $\theta$ and $\beta$ and use eq. (2) to simulate the effect on test scores of being administered tests of different length.

The gap between highly and poorly endowed students is monotonous in test length, corresponds to 1.57 standard deviations when the test has 46 questions (the current INVALSI test length), and increases by about 0.07 and 0.12 standard deviations when test length $N$ is equal to 60 and 70 questions respectively. One might argue, however, that the measured gap reflects both

[15] The inclusion of class fixed effects in the regressions prevents the identification of the effects of class size and being drilled on the intercept.
[16] Bully victimization is defined at the individual level – see Section 2 for details.

differences in the received background – which matter for equality of opportunity (see Dworkin, 1981, Roemer, 1998, Fleurbaey, 2008) – and differences in individual behaviour (e.i. school effort). To purge the simulated gap from the latter, we estimate a restricted version of (4), where $\theta$ and $\beta$ depend only on variables that are mainly outside individual control, such as the number of books at home and immigrant status, in addition to age, gender and question dummies. The restricted version considers both the direct contribution of family background and immigration status on test scores, and the indirect one operating via cognitive and non-cognitive abilities and school choice. We simulate the test scores of highly and poorly endowed students as the test length increases from 46 to 60 (resp. 70) questions and confirm that the gap between the two groups increases, albeit somewhat less than in the unrestricted version (by 0.04 (resp. 0.07) standard deviations).

Since the composition of pupils varies among classes (and schools), the relative ranking of classes in terms of their average expected test scores also varies with *N*. Using the estimated values of individual $\theta$ and $\beta$, we find that a reduction of the length of the INVALSI test from 46 to 30 (resp. 20) questions would trigger a significant change in the ranking, with about 29 (resp. 47) percent of all classes changing their position by at least 5 places in either direction. The classes gaining at least five positions have a higher percentage of female pupils and of pupils with a more privileged background (measured by the number of books in the house) than the rest of the sample.

*5. Discussion of some assumptions*

We have assumed in Eq. (1) that the estimated marginal effect $\frac{\partial Y_{iq}}{\partial P_{iq}}$ is common across questions for any given individual. This may be restrictive if, for instance, the decline of individual performance with question position of questions depends on the difficulty of the question.

When the marginal effect of P on Y varies with the question being asked rather than with the individual taking the test, Eq. (1) can be rewritten as

$$Y_{iq} = \theta_i + \sum_{q=1}^{Q} \rho_q P_{iq} Q_q + \sum_{q=1}^{Q} \mu_q Q_q + \varepsilon_{iq} \quad (6)$$

This specification allows us to test whether the null hypothesis of constant effects $H_0: \rho_q = \rho$ holds for all questions or for a subset of questions. To maximise efficiency and increase the power of the test, we estimate Eq. (6) using random effects. Our results indicate that the null hypothesis $H_0$ holds for 11 of the 18 available questions. However, when we restrict the sample used to estimate Eq. (1) to these questions, we find that the marginal effect of P on Y is somewhat lower in absolute value but not statistically different from the one estimated using the much larger sample with all 18 questions – see Table 7. We conclude that, although the assumption of constant (across-question) effects implied by Eq. (1) is only partially supported in our data, removing it would alter our baseline estimates only marginally.

An additional issue not considered when estimating Eq. (1) is that the decline of performance with the position of questions depends on the difficulty of the first question(s). On the one hand, difficult questions at the start of the test might require more time to be answered, increasing anxiety and pressure to complete the test on time, with negative effects on the relationship between P and Y. On the other hand, difficult initial questions might boost attention not only at the beginning but for the entire length of the test, reducing the marginal negative effect of P on Y.

The heterogeneity induced by factors that vary at the position-by-question level is difficult to study with our data. One way to address this issue is to construct an index of question difficulty using the proportion of correct answers by question, and augment Eq. (1) with the interaction of question position P with either the difficulty of the first question or the average difficulty of the first three questions. The results reported in Table 8 indicate that these interactions are generally positive but never statistically significant.

In our data, the estimation of the marginal effect of $P_{iq}$ on the test score is possible because different booklets are assigned to different students. However, the random assignment of booklets also implies that several questions swap their position *simultaneously*. If the marginal effect of $P_{iq}$ depend on the order of all questions in the questionnaire, our estimates are biased by the presence of a differential frame effect experienced by students assigned to different booklets.[17]

To investigate this issue, we consider the questions that do not change their position across booklets, which have been excluded from the analysis so far, and test whether the probability of a correct answer to these questions depends on the booklet. A positive answer would be evidence of a frame effect, because the position of other questions would influence the probability of responding correctly to questions having a fixed position. Since our estimates do not reject the hypothesis that all booklets equally affect the probability of a correct answer to the questions with a fixed position, we exclude the presence of frame effects.[18]

We have assumed that the parameters $\theta_i$ and $\beta_i$ in Eq. (1) are normally distributed, a necessary condition to implement the mixed model. While this assumption may seem restrictive, we argue that it is not. First, parameters $\theta$ and $\beta$ can be modelled as functions of many factors, including cognitive and non-cognitive abilities, parental background and school quality. Assuming that the relationship between factors and parameters is linear, a large number of factors guarantees normality because of the central limit theorem. Second, we use the estimates in Table 6, which do not rely on the normality of θ and β and include relatively few factors, to predict θ and β for each student. The resulting empirical distributions – shown in Figures 3A and 3B – turn out to be approximately normal.

[17] This would happen, for instance, if βi in Eq. (1) varied with the difficulty of the first questions.
[18] Results available from the authors upon request.

Finally, we have implicitly assumed that students answer questions in the same order as they appear in the questionnaire. Since the test designed by INVALSI is paper based, we cannot exclude that some students have followed different orders, skipping for instance difficult questions and returning to them later on. This possibility suggests that we interpret our estimates as intention to treat (ITT) rather than treatment (TE) effects, with the former being a lower bound to the latter. Clearly, if the order actually followed by students (their take up) were uncorrelated with that of the questionnaire, we should have found no ITT effect, contrary to our results.

*Conclusions*

Using Italian data on standardized test scores, we have shown that there is important heterogeneity in the relationship between student performance and the position of questions in the test. An unpleasant consequence of this heterogeneity is that increasing the length of the test in order to reduce the contribution of noise is likely to change the ranking of individuals, classes or schools and can also raise test score inequality between "highly" and "poorly" endowed students. Given the increasing focus on school accountability and on the relative performance of schools, our results suggest that the implications of test length on the distribution of scores should be carefully considered when designing tests.

Another unpleasant consequence of heterogeneous responses is that the gap between the variance of test scores and the variance of underlying skills does not go to zero but can even widen as the length of the test increases. When this gap is an ingredient in the definition of optimal test length, ignoring heterogeneous responses produces longer than optimal tests.

We have studied the determinants of both the performance in the first question and the decline in test scores as the position of questions increases. Our results suggest that cognitive skills, non-cognitive skills, family and class characteristics influence both the initial performance and performance decline,

casting some doubts on the decomposition proposed by Borghans and Schils, 2012.

While our findings concern primary school students taking a low stake test, they may have broader applicability, as suggested by the fact that our estimates of the mean effect of question position on test scores are qualitatively similar to the one obtained by Borghans and Schils, 2012, and Borgonovi and Biecek, 2016, for students who are about five years older than those in our sample.

To what extent our results apply also to high stake tests, such as the SAT, the GRE, and the admission tests organized by many universities - including the most prestigious academic institutions - is an open question that we cannot answer. On the one hand, we speculate that the heterogeneity in the relationship between performance and the position of questions may be lower in high than in low stake tests, because candidates take the former more seriously. If this is the case, the probability that ranking changes as the number of questions increases may also be lower in high stake tests.

On the other hand, we believe that individuals sitting high stake tests are under heavier pressure than those taking low stake tests. Since the ability to endure pressure and stress varies across individuals, the relationship between performance and position could be more heterogeneous in high than in low stake tests, and both the final ranking and the probability of being admitted to top academic institutions could depend in a non-negligible way on test length, for any given distribution of ability across candidates.

An implication of our research is that educational institutions can vary the composition of the pool of admitted students by altering the test length. In particular, our results suggest that, when the relative performance on math tests is one of the requirements for admission to elite schools, girls and natives are likely to gain from longer tests, while boys and immigrants may lose.

*References*

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163-181.

Angrist, J. D., Battistin, E., & Vuri, D. (2017). In a small moment: Class size and moral hazard in the Italian Mezzogiorno. *American Economic Journal: Applied Economics*, *9*(4), 216-49.

Balart, P., & Oosterveen, M. (2018). Wait and See: Gender Gaps throughout Cognitive Tests. mimeo.

Balart, P., Oosterveen, M., & Webbink, D. (2018). Test scores, noncognitive skills and economic growth. *Economics of Education Review, 63,* 134-153.

Battaglia, M. and Hidalgo Hidalgo, M. (2018) "Ability to Sustain Test Performance and Remedial Education: Good news for girls", mimeo

Benabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The review of Economic Studies*, *70*(3), 489-520.

Bertoni, M., Brunello, G., & Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics*, *104*, 65-77.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, *43*(4), 972-1059.

Borghans, L., & Schils, T. (2012). The leaning tower of PISA. *Unpublished Manuscript*.

Borgonovi, F., & Biecek, P. (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and Individual Differences*, *49*, 128-137.

Cameron, D., & Trivedi, P.K. (2006). *Microeconometrics*. Cambridge University Press, Cambridge, MA

Cornwell, C., Mustard, D.B. and Van Parys, J. (2013). Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments. Evidence from Primary School. *The Journal of Human Resources*. 48(1):236-264

Cunha, F. and Heckman, J.J. (2008) Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *The Journal of Human Resources*, 43(4):738-782

Cunha, F. and Heckman, J.J. (2010) Investing in our Young People. NBER Working Paper 16201

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*(19), 7716-7720.

Duncan, T.G. & McKeachie, W.J. (2005) The Making of the Motivated Strategies for Learning Questionnaire, *Educational Psychologist*, 40(2):117-128

Dworkin, R. (1981). What is Equality? Part 1: Equality of Welfare. *Philosophy & Public Affairs* 10(3): 185-246

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance--A Critical Literature Review*. Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637.

Fleurbaey, M. (2008). *Fairness, Responsability and Welfare*. Orxford University Press

Gneezy, U., List, J. A., Livingston, J. A., Sadoff, S., Qin, X., & Xu, Y. (2017). Measuring success in education: the role of effort on the test itself National Bureau of Economic Research Working Paper n.24004.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological bulletin*, *136*(4), 495-525.

Jacob, B. A. (2016). Student Test Scores: How the Sausage Is Made and Why You Should Care. Evidence Speaks Reports, Vol 1,# 25. *Center on Children and Families at Brookings*.

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PloS One*, *8*(8), e70270.

John, O.P. and Srivastava, S. (1999). The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. L. Pervin and John O.P. (Eds.), *Handbook of personality: Theory and research (2nd ed.)*. New York: Guilford.

Nizalova, O.Y. and Murtazashvili, I., 2016. Exogenous treatment and endogenous factors: Vanishing of omitted variable bias on the interaction term. *Journal of Econometric Methods*, *5*(1), pp.71-77.

Nyhus, E. K., & Pons, E. (2005). The effects of personality on earnings. *Journal of Economic Psychology*, *26*(3), 363-384.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, forthcoming.

Pope, D. G., & Fillmore, I. (2015). The impact of time between cognitive tasks on performance: Evidence from advanced placement exams. *Economics of Education Review*, *48*, 30-40.

Rodríguez-Planas, N., & Nollenberger, N. (2018). Let the girls learn! It is not only about math… it's about gender social norms. *Economics of Education Review*, *62*, 230-253.

Roemer, J. (1998). *Equality of Opportunity*. Harvard University Press, Cambridge

Segal, C., 2012. Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, *58*(8), pp.1438-1457.

Utvær, B. K. S., & Haugan, G. (2016). The academic motivation scale: dimensionality, reliability, and construct validity among vocational students. *Nordic Journal of Vocational Education and Training*, *6*(2), 17-45.

Zamarro G., Hitt, C., and Mendez, I. (2017) “When Students Don’t Care: Reexamining International Differences in Achievement and Non-Cognitive Skills”, EDRE Working Paper 2016-18

**Tables and figures**.

Table 1. Summary statistics

| | Mean | St.Dev. |
|---|---|---|
| Y | 52.81 | 21.80 |
| Confidence | 0.046 | 1.396 |
| Conscientiousness | 0.004 | 1.308 |
| Neuroticism | -0.022 | 1.548 |
| Bullied | 0.011 | 1.520 |
| Agreeableness | 0.031 | 1.448 |
| Intrinsic motivation | -0.070 | 2.230 |
| Extrinsic motivation | -0.036 | 1.942 |
| Math grade in the last semester | 7.915 | 1.098 |
| Female | 0.488 | 0.500 |
| Age (in months) | 129.94 | 4.791 |
| Immigrant status | 0.10 | 0.300 |
| Less than 26 books in the house | 0.350 | 0.477 |
| Small class (dummy) | 0.346 | 0.476 |
| Trained to test (dummy) | 0.483 | 0.500 |

Table 2. Balancing tests

| | Female | Age | Math grade | Confidence | Extrinsic motivation | Intrinsic motivation | Neuroticism | Bullied | Consci.ness | Agree.ness | Books in the house | Born abroad |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| booklet 1 | 0.45*** | 129.56*** | 7.74*** | 0.26 | 0.18* | 0.00 | 0.08 | 0.80*** | -0.08 | -0.42*** | 3.05*** | 0.04*** |
| | (0.03) | (0.11) | (0.54) | (0.26) | (0.10) | (0.07) | (0.16) | (0.02) | (0.06) | (0.14) | (0.21) | (0.01) |
| booklet 2 | 0.45*** | 129.38*** | 7.74*** | 0.22 | 0.19* | -0.00 | 0.07 | 0.87*** | -0.06 | -0.45*** | 3.07*** | 0.04*** |
| | (0.03) | (0.10) | (0.54) | (0.26) | (0.10) | (0.07) | (0.16) | (0.02) | (0.06) | (0.14) | (0.21) | (0.01) |
| booklet 3 | 0.45*** | 129.52*** | 7.74*** | 0.21 | 0.19* | -0.07 | 0.08 | 0.80*** | -0.12** | -0.47*** | 3.07*** | 0.03*** |
| | (0.03) | (0.11) | (0.54) | (0.26) | (0.10) | (0.07) | (0.16) | (0.02) | (0.06) | (0.14) | (0.21) | (0.01) |
| booklet 4 | 0.44*** | 129.40*** | 7.76*** | 0.21 | 0.22** | 0.03 | 0.08 | 0.79*** | -0.08 | -0.49*** | 3.06*** | 0.03*** |
| | (0.03) | (0.11) | (0.54) | (0.26) | (0.10) | (0.07) | (0.16) | (0.02) | (0.06) | (0.14) | (0.21) | (0.01) |
| booklet 5 | 0.46*** | 129.42*** | 7.73*** | 0.23 | 0.21** | 0.02 | 0.09 | 0.83*** | -0.07 | -0.45*** | 3.08*** | 0.03*** |
| | (0.03) | (0.11) | (0.54) | (0.26) | (0.10) | (0.07) | (0.16) | (0.03) | (0.06) | (0.14) | (0.21) | (0.01) |
| Observations | 19,656 | 19,655 | 18,907 | 19,401 | 18,881 | 18,603 | 19,253 | 19,259 | 19,295 | 19,349 | 19,231 | 19,656 |
| R-squared | 0.51 | 1.00 | 0.99 | 0.10 | 0.14 | 0.15 | 0.12 | 0.12 | 0.10 | 0.13 | 0.89 | 0.15 |
| Test | 0.614 | 0.376 | 0.911 | 0.462 | 0.898 | 0.408 | 0.978 | 0.172 | 0.385 | 0.321 | 0.82 | 0.111 |

Notes: each regression includes a constant and 1116 class dummies. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence. Consci.ness is for conscientiousness and Agree.ness is for agreeableness. In the last row, we report the p-value of the joint test that the coefficients associated to the booklet dummies are statistically equal.

Table 3. Estimates of the two-level model

| | All | Males | Females |
|---|---|---|---|
| $\mathbf{E(\theta) = \alpha}$ | 85.18*** | 86.21*** | 84.10** |
| $\mathbf{E(\beta) = \delta}$ | -0.060*** | -0.083*** | -0.035*** |
| $\mathbf{Var(\theta)}$ | 315.77*** | 304.74*** | 302.67*** |
| $\mathbf{Var(\beta)}$ | 0.035*** | 0.041*** | 0.029*** |
| $\mathbf{Cov(\theta, \beta)}$ | 0.65** | 0.94*** | 0.49*** |
| $\rho$ ($\theta$,$\beta$) | 0.20 | 0.27 | 0.16 |
| $\boldsymbol{\sigma_\varepsilon}$ | 1846.17*** | 1815.91*** | 1865.13*** |

Note: maximum likelihood estimates. Number of observations in the full sample: 353,808; in the sample of females: 173,016; in the sample of males: 180,792. The standard errors are clustered at the class level.

Table 4. Average intercepts and average decline effect, by number of books at home and immigrant status

| | E(θ) Males | E(θ) Females | E(β) Males | E(β) Females |
|---|---|---|---|---|
| 0-10 books at home | 79.82 | 77.89 | -0.121 | -0.062 |
| 11-25 books at home | 83.82 | 81.06 | -0.099 | -0.047 |
| 26-99 books at home | 87.77 | 85.12 | -0.073 | -0.031 |
| 100-199 books at home | 89.15 | 87.22 | -0.063 | -0.023 |
| 200 or more books at home | 90.05 | 88.49 | -0.059 | -0.017 |
| | | | | |
| Natives | 86.81 | 84.62 | -0.079 | -0.033 |
| Immigrants | 80.81 | 79.48 | -0.119 | -0.054 |

Table 5. Simulated change in the test score for hypothetical males and females with average initial performance

| N | Male | Female |
|---|---|---|
| 0 | 86.21 | 84.10 |
| 10 | 85.79 | 83.93 |
| 20 | 84.92 | 83.58 |
| 30 | 83.58 | 83.04 |
| 40 | 81.79 | 82.32 |
| 50 | 79.54 | 81.41 |

Note: N is the test length.

Table 6. The determinants of θ and β.

| | Intercept θ | Slope β |
|---|---|---|
| Position | | -0.119*** |
| | | (0.014) |
| High Grade (HG) | 13.650*** | 0.022** |
| | (0.367) | (0.011) |
| Conscientiousness | -1.079*** | 0.011** |
| | (0.164) | (0.005) |
| Neuroticism | -1.334*** | -0.006 |
| | (0.121) | (0.004) |
| Confidence | 4.921*** | 0.009** |
| | (0.138) | (0.004) |
| Agreeableness | 0.319** | -0.002 |
| | (0.135) | (0.004) |
| Intrinsic motivation | -0.057 | -0.004 |
| | (0.098) | (0.003) |
| Extrinsic motivation | -1.413*** | -0.004 |
| | (0.092) | (0.003) |
| Books | 1.890*** | 0.035*** |
| | (0.405) | (0.012) |
| Immigrant | -0.522 | -0.066*** |
| | (0.630) | (0.019) |
| Female | -4.038*** | 0.048*** |
| | (0.255) | (0.011) |
| Small class | | 0.022* |
| | | (0.012) |
| Trained to the test | | 0.027** |
| | | (0.013) |
| Bullism | -0.231* | -0.008** |
| | (0.125) | (0.004) |
| Constant | 78.081*** | |
| | (1.149) | |
| # Observations | 353,142 | |
| Question Fixed Effects | Y | |
| Class Fixed Effects | Y | |
| Additional controls | Y | |
| Test that the effects on the slope of non-cognitive traits are jointly significant (p-value) | | 0.003 |
| Test that the effects on the intercepts of non-cognitive traits are jointly significant (p-value) | 0.000 | |

Note: The model includes question and class dummies, dummies for missing values of the relevant variables and the following additional controls: age, dummies for childcare and kindergarten and a dummy for enrolment in primary schools at age 6. Standard errors are clustered at the class level. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 7. The relationship between P and Y in the full sample using 18 questions and in the sub-sample of 11 questions.

| | All questions | Only 11 questions |
|---|---|---|
| P | -0.060*** | -0.054*** |
| | (0.006) | (0.009) |
| Test of equality across equations (p-value) | 0.225 | |
| Observations | 353,808 | 216,216 |
| # of questions | 18 | 11 |

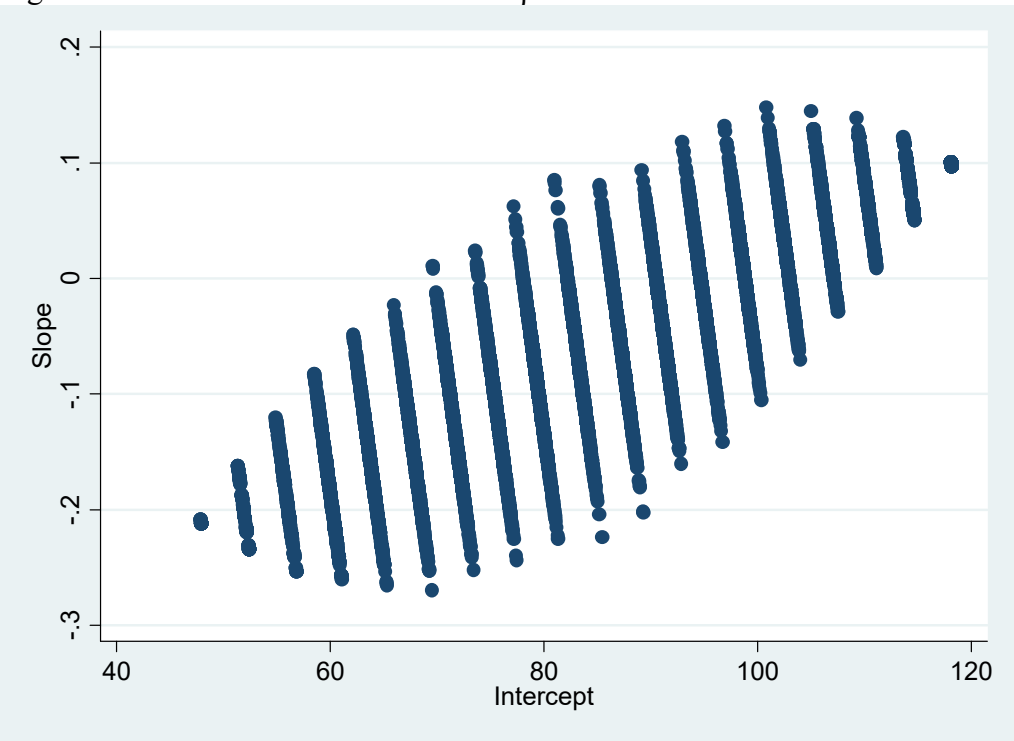
Note: random effects estimates. Each regression includes a constant and question dummies. The variable subset is a dummy equal to 1 for the subset of 11 questions and to 0 otherwise. The estimates in the third column include also the interactions of question dummies with a dummy subset. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Table 8. Heterogeneity of the decline effect by difficult of the initial questions of each student questionnaire.

| | All | Males | Females | All | Males | Females |
|---|---|---|---|---|---|---|
| P | -0.088*** | -0.100*** | -0.075** | -0.095*** | -0.080* | -0.111** |
| | (0.025) | (0.034) | (0.037) | (0.032) | (0.043) | (0.046) |
| P * difficulty first question | 0.034 | 0.020 | 0.048 | | | |
| | (0.030) | (0.041) | (0.044) | | | |
| P * avg diff. first three qs | | | | 0.064 | -0.007 | 0.140 |
| | | | | (0.058) | (0.080) | (0.085) |
| Observations | 353,142 | 180,468 | 172,674 | 353,142 | 180,468 | 172,674 |

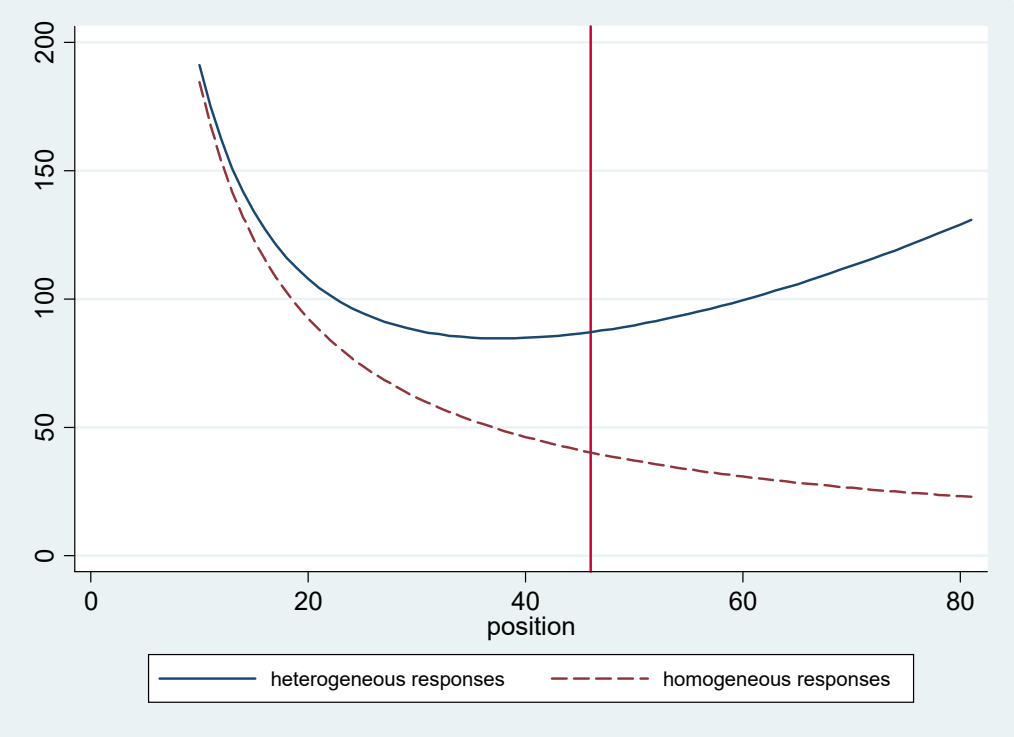
Note: fixed effects estimates. Each regression includes question dummies. Standard errors are clustered at the class level. One, two and three stars for statistical significance at the 10, 5 and 1 percent level of confidence.

Figure 1. Distribution of random effects β and θ




Note: predicted individual values using the mixed model.

Figure 2. Estimated variance of the test score minus estimated variance of θ




Note: based on estimates of the mixed model.

Figure 3A – Empirical distribution of θ derived from the estimates of Eq. (5)

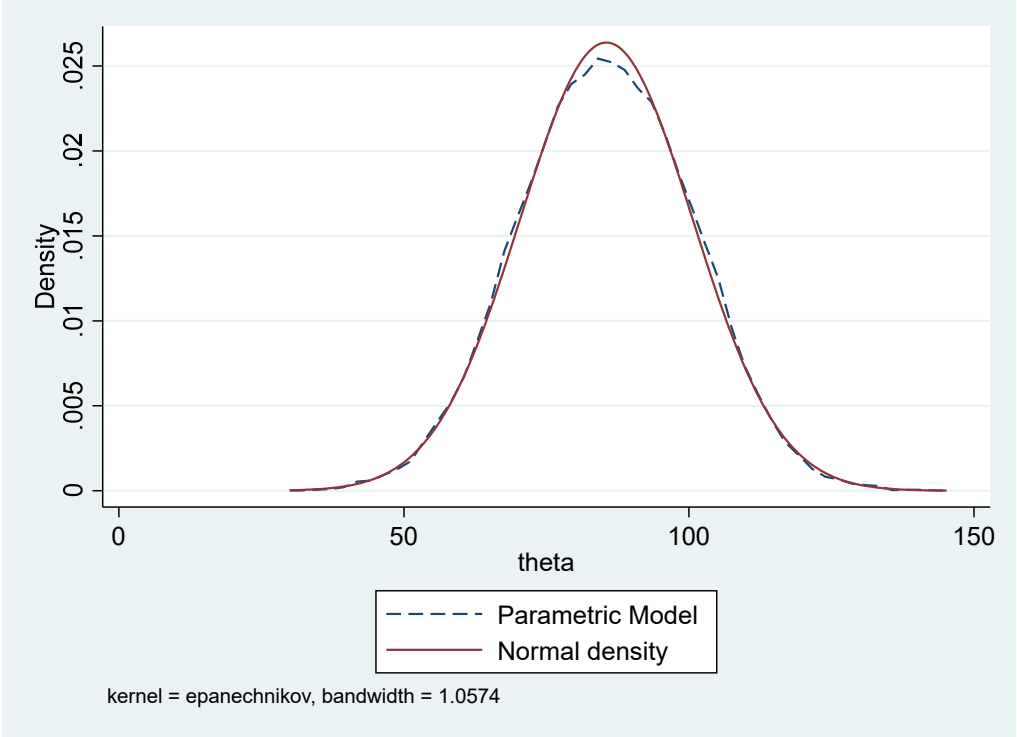
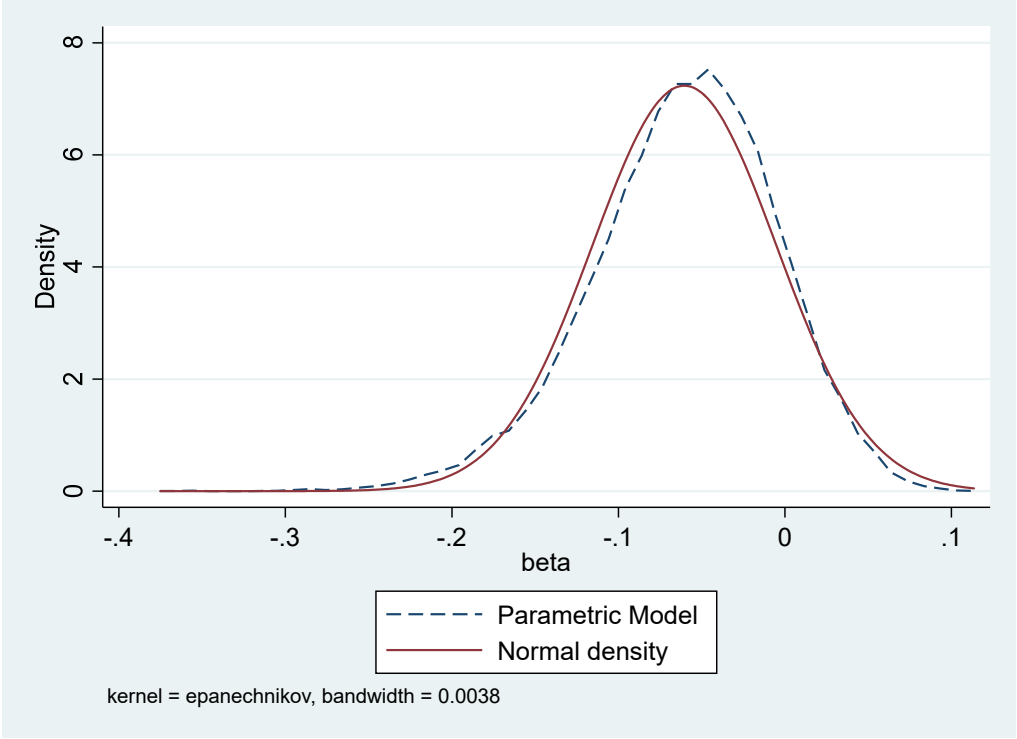


Figure 3B - Empirical distribution of β derived from the estimates of Eq. (5)



**Appendix**

The information required to construct our indicators of non-cognitive abilities and bully victimization originates from a student questionnaire that was administered to test takers after the conclusion of the test. The format of the relevant questions consists of a number of items.

Adapting to this context the taxonomies developed by John and Strivastava (1999) for the Big Five and by Duncan and McKeachie (2005) for motivation, we divide questions in groups and we use a principal component analysis to extract the factor associated to the highest eigenvalue. The latter correspond to our indicators of personality traits and motivation:

1. *Conscientiousness*. The relevant question is: can you manage to a) complete your homework in time; b) focus on study when there are other interesting things to do; c) concentrate on your study without distractions; d) remember what the teacher has explained in class. For each item, the pupil could choose between four answers: never (coded 1); to some extent (coded 2); often (coded 3) and very often (code 4).
2. *Agreeableness*. The relevant question concerns the interaction with classmates. There are four sub-questions: a) how many classmates talk to you? b) how many classmates can you consider as your friends? c) how many classmates would you help? d) how many classmates have good relationships with you? For each item, the pupil could choose between four answers: none (coded 1); few (coded 2); some (coded 3); many (coded 4) and all (code 5). For each sub-question, the pupil could choose between four answers: none (coded 1); few (coded 2); some (coded 3); many (coded 4) and all (code 5).
3. *Confidence*. The relevant question is: do you agree with the following statements? a) I usually do well in Math; b) I learn Math easily; c) Math is more difficult for me than for my classmates. For each item, the pupil could choose between four answers: not at all (coded 1); somehow (coded 2); enough (coded 3); very much (coded 4).
4. *Neuroticism.* The relevant question is: do you agree with the following statements? a) I was worried about the test before starting it; b) I was so nervous I could not answer; c) during the test I felt I was not going well; d) during the

test I felt OK. For each item, the pupil could choose between four answers: not at all (coded 1); somehow (coded 2); enough (coded 3); very much (coded 4).

5. *Intrinsic motivation.* There are three relevant questions: why should you perform well? Why should you do your homework? What do you think about studying? For the first question, we use the following items: a) I feel bad if I do not perform well; b) I like to perform well; c) I feel ashamed if I do not perform well; d) doing well at school is fun. For the second question, we use the items: a) I feel guilty if I do not do my homework; b) doing my homework is good for me; c) I am ashamed if I do not do my homework; d) I like to do my homework. For the last question, we use the following items: a) I think that learning new things is important; b) I think that learning as much as possible is important; c) it is important to understand well what I study; d) it is important to improve during the year. For each item, the pupil could choose between four answers: not at all (coded 1); somehow (coded 2); enough (coded 3); very much (coded 4).
6. *Extrinsic motivation.* There are three relevant questions: why should you perform well? Why should you do your homework? What do you think about studying? For the first question, we use the following items: a) if I do well I could get an award; b) if I do well they let me do what I want; c) if I do not perform well I could be punished. For the second question, we use the item: a) I will be punished if I do not do my homework. For the last question, we use the following items: a) it is important for me to show others that I am good; b) it is important to appear to be cleverer than my classmates; c) it is important for me to show that I do well on tests. For each item, the pupil could choose between four answers: not at all (coded 1); somehow (coded 2); enough (coded 3); very much (coded 4).

The *bully victimization* index is the principal component obtained from the following questions: during this year, how often did you experience: a) to be insulted by other students; b) to be beaten up by other students; c) to be excluded by other students. For each item, the pupil could choose between four answers: never (coded 1); to some extent (coded 2); often (coded 3) and very often (code 4).