I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11603

Testing at Length If It Is Cognitive or Non-Cognitive

Giorgio Brunello
Angela Crema
Lorenzo Rocco

JUNE 2018

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11603

Testing at Length If It Is Cognitive or Non-Cognitive

Giorgio Brunello
University of Padova and IZA

Angela Crema
University of Padova

Lorenzo Rocco
University of Padova

JUNE 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Testing at Length If It Is Cognitive or Non-Cognitive*

Using Italian data on standardized test scores, we show that the substantial heterogeneity in how performance changes with the position of questions can alter the rank of individuals and classes as the length of the test increases. We examine whether decomposing test scores into initial performance and performance decline allows to separate the influence of cognitive and non-cognitive skills and find that our measure of cognitive skills – the math grade before the test – not only has a dominant influence on initial performance but also affects substantially performance decline.

JEL Classification: I21

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

Corresponding author:
Giorgio Brunello
Department of Economics and Management "M. Fanno"
University of Padova
via del Santo 33
35123 Padova
Italy
E-mail: giorgio.brunello@unipd.it

* We are grateful to Marco Bertoni, David Kiss and the audiences at talks in Florence and Padova for comments. A special thanks to Patrizia Falzetti for help with the data. The usual disclaimer applies.

NON-TECHNICAL SUMMARY

Using Italian data on standardized test scores, we show that, when the characteristics of pupils vary among classes or schools, altering the length of the test not only reduces measurement error but can also change the ranking of individuals, classes or schools. Given the increasing focus on school accountability and on the relative performance of schools, our results suggest that the implications of test length on relative performance should be carefully considered when designing tests.

We also investigate whether decomposing test scores into initial performance and performance decline can be used to separate the influence of cognitive and non-cognitive skills and find that our measure of cognitive skills – the math grade before the test – not only has a dominant influence on initial performance but also substantially affects performance decline.

While our findings concern primary school students taking a low stake test, they may have broader applicability, as suggested by the fact that our estimate of the mean effect of the position of questions on the test outcome is qualitatively similar to the one obtained in the literature using data from PISA low stake tests taken by students who are about 5 year 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 in this paper.

We speculate, however, 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 are likely to gain from longer tests, while boys instead are likely to lose.

Introduction

When the probability of giving a correct answer to 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 the relationship between the position of questions and the probability of providing a correct answer varies across pupils, and the composition of pupils varies within and between institutions, a longer test not only reduces measurement error (Jacob, 2016) but can also affect the rankings of students, classes and schools.

The observation that students typically perform better on early than on later questions has motivated attempts to disentangle the contributions of cognitive ability and personality traits to test scores. Borghans and Schils, 2012, for instance, exploit the information on the position of questions in PISA and the fact that the booklets containing questions arranged in different order are randomly allocated to students to estimate how the probability of a correct answer varies with the position of the question. They attribute the response to the first question to cognitive ability and the observed decline of performance in later questions to personality traits.

Our contribution to this literature is twofold. First, we show that the relationship between performance in the test and the position of questions varies across students with different observable characteristics. We exploit this heterogeneity to study how the ranking of individuals and classes varies with the length of the test. Second, we verify whether the decomposition of test scores discussed above holds in our data.

We use data on Italian standardized tests. Similarly to PISA, these tests are not high stakes, and are administered every year to the universe of Italian schools by the Italian agency INVALSI. As in PISA, booklets containing the questions in different orders are randomly allocated to students. 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.

We exploit the fact that in our data all measured individual characteristics are by construction independent of the linear trend that captures the position of questions, regardless of whether this position is randomized or not. This feature is useful when we use a multi-level model, which relies on strong independence assumptions. The fact that booklets are randomly assigned to students remains relevant, however, to support another assumption routinely made in this literature, that the effect of the position of questions does not depend on the specific question in that position.

We show that the relationship between test performance and the position of questions in the test is negative on average, as suggested by the literature. There is, however, substantial heterogeneity, and this relationship is positive for 26 percent of pupils. Since the composition of pupils varies between classes, the observed heterogeneity implies that the ranking of classes can change with the length of the test. We show that – conditional on the difficulty of questions - about 80 and more than 90 percent of classes change rank between the first and the 10th question and between the first and the 40th question respectively. After 40 questions, the share of classes with a significant change of rank (at least 10 positions in either direction) is close to 40 percent.

If the effect of the position of questions placed after the first on test performance depends exclusively on personality traits, as posited by Borghans and Schils, it should not vary with observed measures of cognition. Yet we show that it varies significantly with the math grade attained in the first quarter of the academic year, well before the test is taken. There is also evidence that the relationship between performance and question position varies with parental background, which typically affects both cognitive and non-cognitive skills.

We also show that the answer to the first question depends mainly on cognitive but also on non-cognitive variables. We conclude that the decomposition of test scores in two components, starting performance and decline in performance, cannot be interpreted as the decomposition of these scores into a cognitive and a non-cognitive component.

The paper is organized as follows. Section 1 reviews the literature, Section 2 sets up the empirical model and Section 3 introduces the data. We discuss a key assumption made in the empirical model and in the related literature in Section 4 and present our results in Section 5. Conclusions follow.

Section 1. Literature Review

This paper refers to three main streams of economic and psychological literature: the first stream explores the effect of test length on performance; the second attempts at disentangling the contribution of cognitive ability and personality traits to test scores, and the third measures the effect of non-cognitive skills on school achievement (and other adult-life outcomes).

A model 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 do 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, Wood, Stiff, and Chatzisarantis, 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 contributors to student performance: cognitive skills and personality traits. Balart, Oosterveen and Webbink, 2018, apply this approach to decompose PISA test scores into a cognitive component, the starting performance, and a non-cognitive component, the decline in performance during the test, and show that both components contribute to economic growth in a sample of countries. In previous work, Balart and Oosterveen, 2017, adopt this strategy using PISA data and show that longer tests decrease the gender gap in math. Their results also suggest that non-cognitive skills are not capable of explaining the lower performance of females compared to males in math and science.[1]

Using PISA data, Borgonovi and Biecek, 2016, 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 goal is to examine whether academic endurance varies across groups of individuals. Their findings suggest that girls and socio-economically advantaged students have higher levels of endurance on average compared to males and pupils with a low socioeconomic background, respectively. Also, they 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 test scores reflect not only ability, knowledge, and intelligence but also personality traits, motivation, grit and self-control[2]. Test takers may not exert maximal effort. When tests are low stakes, like in the OECD PISA project, some individuals try harder than others (see Duckworth, Quinn, Lynam, Loeber and Stouthamer-Loeber, 2011).

[1] 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.

[2] Farrington, Roderick, Allensworth, Nagaoka, Keyes, Johnson, and Beechum, 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.

Scores can also be improved by offering a reward (see Borghans, Duckworth, Heckman and ter Weel, 2008; Segal, 2012). Since test scores reflect differences in individual motivation[3] and not just differences in abilities, ranking countries based on average low-stakes assessments is problematic (see Gneezy, List, Livingston, Sadoff, Qin, and Xu, 2017).

Section 2. The Empirical Model

Consider a standardized test with N questions. The position of each question, from 0 to N-1, varies across booklets, and these booklets are randomly assigned to students. The relationship between the test outcome Y - the answer to each question (correct or wrong[4]) – and the position of the question P is

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

where the indices i, q and p indicate respectively the individual, the question and the position of the question; Q is a question fixed effect; ε is the noise of the test; θ is an indicator of individual skills and traits – including cognitive ability and personality traits; β is the marginal effect of P on Y, which we assume to vary among individuals but not among questions (we discuss this restriction later in the paper). We further assume that $\theta_i \sim N(\alpha, \sigma_\theta)$, $\beta_i \sim N(\delta, \sigma_\beta)$, $\varepsilon_i \sim N(0, \sigma_\varepsilon)$, that θ and β are correlated – with covariance matrix Σ - but independent of ε, and $\sum_1^Q \mu_q Q_q = 0$, a normalization.

Letting $\rho \in \{\beta, \theta\}$, we remark that P_{iq} (as well as each question dummy) in Eq. (1) is independent of ρ, regardless of whether booklets are randomly distributed. The reason is that P_{iq} is a linear trend common to all students, which implies that both $f(P_{iq})$, the marginal distribution of P_{iq}, and

[3] 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).

[4] Skipped questions are considered as wrong answers.

$f(P_{iq}|\rho_i)$, the distribution of P_{iq} conditional on ρ_i, are uniform and defined over the same support. Therefore, $f(P_{iq}|\rho_i) = f(P_{iq})$ and the joint distribution $h(P_{iq}, \rho_i)$ can be written as $h(P_{iq}, \rho_i) = f(P_{iq})g(\rho_i)$. Randomization remains crucial, however, to obtain unbiased estimates when the effect of P_{iq} is allowed to vary across questions, as discussed below.

The independence of θ_i and β_i of P_{iq} satisfies the requirements of multi-level models, a class of models with random intercepts and slopes. In our case, Eq. (1) is a two-level model, with questions representing level 1 and students representing level 2. This model can be estimated using maximum likelihood.

In this paper, we wish to test the hypothesis that test scores can be decomposed into a component determined solely by cognitive skills (the answer to the first question) and a component affected exclusively by personality traits (the decline of performance in later questions). To perform this test, we introduce the K-dimensional vector $X = \{X_c, X_p\}$ of cognitive and non-cognitive characteristics X_c and X_p, assume that $\theta_i = X_{ic}^{'}\pi_c + X_{ip}^{'}\pi_p + Z_i^{'}\pi_z$ and $\beta_i = P_{iq}X_{ic}^{'}\gamma_c + P_{iq}X_{ip}^{'}\gamma_p$ and re-write (1) as

$$Y_{iqp} = X_{ic}'\pi_c + X_{ip}'\pi_p + Z_i'\pi_z + P_{iq}X_{ic}'\gamma_c + P_{iq}X_{ip}'\gamma_p + \sum_{q=1}^{Q}\mu_q Q_q + \varepsilon_{iqp} \quad (2)$$

where Z is a vector of additional controls.

The independence of P_{iq} with respect to X and ε implies that parameters γ_c and γ_p can be consistently estimated using ordinary least squares (see Nizalova and Murtazashvili, 2016) and that we can test the null hypothesis $\gamma_c = 0$. Under the additional assumption that – conditional on Z - X and ε are uncorrelated, we can also test whether $\pi_p = 0$.[5]

[5] We verify whether our 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 π_p 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

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

$$S_i = \theta_i + \beta_i \frac{N-1}{2} + \varepsilon_i \qquad (3)$$

where $S_i = \frac{1}{N}\sum_1^N Y_{iqp}$ and $\varepsilon_i = \frac{1}{N}\sum_1^N \varepsilon_{iqp}$. A unitary increase in test length N changes the expected test score by $\frac{1}{2}\beta_i$. Therefore, if β_i 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

$$V(S_i) = V(\theta_i) + \left(\frac{N-1}{2}\right)^2 V(\beta_i) + (N-1)Cov(\theta_i, \beta_i) + \frac{\sigma_\varepsilon}{N} \qquad (4)$$

When β does not vary across individuals, the variance of the score tends to the variance of skills $V(\theta_i)$ as N increases and the noise of the test goes to zero. When β varies across individuals, however, $Var(S_i)$ and $Var(\theta_i)$ differ even when the noise is negligible. If $Cov(\theta_i, \beta_i)$ is positive, increasing the test length N has two contrasting effects on the gap between $Var(S_i)$ and $Var(\theta_i)$: on the one hand, it reduces $Var(S_i)$-$Var(\theta_i)$ because it attenuates noise; on the other hand, it widens the gap $Var(S_i)$-$Var(\theta_i)$ by magnifying the impact of individual differences in β.

3. 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.

find that this is always the case in the current setup. Detailed results are available from the authors upon request.

The math test consists of 46 questions contained in five booklets. Differently from what happens in PISA, all five INVALSI booklets contain the same questions. Only their order varies across booklets. Of the 46 questions, 18 change position across booklets: 8 take 4 different positions, 8 take 3 alternative positions and the remaining 2 questions take only 2 alternative positions. We focus on this subset of items because only these items allow us to distinguish the effect of the position of questions from question-specific fixed effects.

Given the evidence of extensive cheating in Italian standardized tests (see for instance Bertoni, Brunello, and Rocco, 2013; Angrist, Battistin, and Vuri, 2017), we only consider the classes (about 2000) where the tests were supervised by an external examiner. In these classes, the cheating algorithm developed by INVALSI indicates that no cheating is to be expected. Our final sample consists of 19, 656 pupils.

Table 1 shows the summary statistics for the variables used in the paper. The outcome variable Y assumes 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 in the empirical economics literature 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 (anxiety 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) and one of poor social relations at school (being bullied). We measure intrinsic and extrinsic motivation with school behaviour driven by internal and external rewards and poor social relations with being the target of threats, intimidation and physical violence. As described more in detail in the Appendix, each indicator is obtained using principal components analysis.

We measure cognitive skills with the math grade in the quarter before the test. Math grades range from 4 (bottom) to 10 (top), with grades under 6 being considered below the passing line. We define the dummy "High math grade" as taking the value 1 for grades 9 and 10 and 0 otherwise. In our sample, close to 33 percent of fifth grades have a high grade (see Table 1).

Finally, we verify that the allocation of booklets to pupils is random by means of balancing tests. We regress in turn all measures in Table 1 that vary across students on booklet dummies and test whether the coefficients associated to each booklet are equal. Table 2 presents the results, which support randomization.

4. Results

We organize this section in two parts. In the first sub-section, we illustrate the heterogeneous response of outcome Y to changes in position P. In the second sub-section, we test the hypothesis that, while the decline of performance as the test proceeds depends exclusively on non-cognitive skills, the answer to the first question relies only on cognitive skills.

4.1 Heterogeneous responses and the effects of test length

Table 3 reports the estimates of the two-level model (1) for the full sample and separately by gender. We find that, on average, performance declines with the position. The average marginal estimated effect, $\delta,$ is equal to -0.060 in the full sample, and to -0.083 and -0.035 for males and females respectively. The variance of β_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 θ and β 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 θ and β in our sample by plotting in Figure 1 their best linear unbiased predictions, after normalizing θ within the unit range. While δ is negative, individual β turns out to be positive for close to 26 percent of the sample. Table 4 shows how average values of θ and β 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 θ and a higher average absolute value of β.

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 indicates that individuals with higher values of cognitive and non-cognitive skills θ experiment either a lower decline of performance as the position of questions increases or even an increase in performance. This could be due either to the positive correlation between cognitive and non-cognitive skills or to the fact that both the answer to the first question and the decline in performance vary with cognitive as well as with non-cognitive skills.

The substantial 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 θ and with the associated value of β.[6] 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).[7]

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. As shown in Figure 2, we find that about 1 percent of the 1117 classes in our final sample changes rank by at least 10 positions after the 10th question. This percentage increases to 12.9 percent after 20 questions, to 25.8 percent after 30 questions and to 37.1 percent after 40 questions. The classes gaining at least 10 positions in the rank after 40 questions 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.

Using the estimates in Table 3, we compute the gap $Var(S_i)$-$Var(\theta_i)$ and show that it declines rapidly as N increases, reaches a minimum just before N=40 and increases again afterwards (see Figure 3). We conclude that the actual test length (N=46) is just above the value of N that minimizes the gap.

4.2 Testing the decomposition of test scores

Our estimates of Eq.(2) are shown in Table 6. The first column shows the results of a parsimonious model that includes only P and individual fixed effects. The second column shows the OLS estimates when P is interacted with both measures of cognitive skills (the math grade) and with personality traits. The third column adds further interactions of P with individual and class characteristics. In the less parsimonious specifications, we control for class fixed effects and for individual characteristics, including gender, immigrant

[6] In practice, we consider the average value of β for individuals with values of θ within a small interval of average gender specific θ.

[7] As already pointed out by Balart and Oosterveen, 2017, the average gender gap in test scores declines with the length of the test.

status, parental background, attendance of kindergarten and of childcare facilities, and age at enrolment in primary school.[8]

We find that the interaction of position P with conscientiousness is positive and statistically significant. There is also evidence that the interaction of P with the index of poor social relations attracts a negative and statistically significant coefficient. The remaining interactions between P and indicators of personality traits turn out to be individually not statistically significant in the conventional sense. However, when we test whether all the interactions of personality traits with P are statistically different from zero, we reject the null of no joint statistical significance.

We also find that the interaction between P and our measure of cognitive skills – the math grade before the test – is positive, statistically significant and sizeable: we estimate that switching from a low to a high grade reduces the marginal effect of P by 57 percent (0.041/0.072). This effect is much larger than the one associated to changes in personality traits.

These estimates suggest that both personality traits and cognitive skills affect how test performance varies with the position of questions, with the latter skills having a larger quantitative impact than the former. Adding further interactions of P with individual and class characteristics (Column (3) in the table) does not change this qualitative result. The size of the effect of the math grade, however, is almost halved. Interestingly, we find that the decline in performance as P increases is significantly smaller for pupils in small classes who have been drilled by the teacher using material similar to the test.

While our tests indicate that both personality traits and cognitive skills affect the answer to the first question, cognitive skills – measured by the math grade – play a prominent role. We estimate that switching from a low to a high grade increases the probability that the first question is correctly answered by 29.4%

[8] We deal with missing values by adding to the regressions missing value dummies.

(15.53/52.81) with respect to the mean. In contrast, increasing confidence by 100% only increases initial performance by 0.3 percent (3.55*0.047/52.81).

In summary, our estimates support only partially the hypothesis that test scores can be decomposed into a cognitive component – the answer to the first question – and a non-cognitive component – the decline of performance after the first question. On the one hand, cognitive skills have a dominant role in the answer to the first question, in line with the proposed decomposition. On the other hand, these skills play a relevant role also in the relationship between performance and question position, which is affected also by measured personality traits.

Balart, Oosterveen and Webbink, 2018, find that both starting performance and performance decline during the test are associated with economic growth, and that the estimated effect of performance decline is approximately equal to the estimated effect of starting performance. They interpret this as indication that cognitive and non-cognitive skills have similar importance for growth. By showing that the performance decline reflects not only personality traits but also cognitive skills, our results cast some doubts on this interpretation.

5. Does the Effect of Position P on Outcome Y Vary with the Questions Asked?

We have assumed in Eq. (1) that the estimated marginal effect $\frac{\partial Y_{iqp}}{\partial P_{iq}}$ does not vary with the question q for any given individual. This may be restrictive if, for instance, easy questions are correctly answered regardless of the position they take, while difficult questions are more likely to be answered if located at the beginning of the questionnaire.

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

$$Y_{iqp} = \theta_i + \sum_{q=1}^{Q} \rho_q P_{iq} Q_q + \sum_{q=1}^{Q} \mu_q Q_q + \varepsilon_{iqp} \qquad (5)$$

By estimating equation 5 we test whether the null hypothesis of constant effects H_0: $\rho_q = \rho$ holds for all questions or for a subset of questions.

Notice that in Eq. (5) the individual effect θ_i and the interactions $P_{iq}Q_q$ are not mechanically uncorrelated as it was the case for Eq. (1), because the position taken by question q depends on the booklet that student i has received. However, in Eq. (5) independence is guaranteed by the random distribution of the booklets to students.

To maximise efficiency and increase the power of the test, we estimate (5) using random effects. Our results indicate that the null hypothesis holds for 11 of the 18 available questions. When we restrict our sample to these questions (the poolable questions), we find that the marginal effect of P on Y is somewhat lower in absolute value but statistically not different from that obtained from the much larger sample that includes 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, its violation introduces only a minor bias to our estimates.

An additional potential complication is that while the marginal effect of P_{iq} can be computed only because different booklets are assigned to different students, this same feature also implies that several questions swap *simultaneously* their position. If the marginal effect of P_{iq} depended on the order of all questions in the questionnaire, our estimates would be biased by the presence of a differential frame effect experienced by students assigned to different booklets.[9]

To investigate this possibility, we turn to the questions that do not change their position across booklets, which have been excluded from the analysis so far. We test whether the probability of a correct answer to these questions depends on the booklet. If yes, there would be evidence of a frame effect, because the

[9] This would happen, for instance, if βi in Eq.(1) varied with the difficulty of the first questions.

position of other questions would influence the probability of responding correctly to questions with a fixed position. 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 thus exclude the presence of frame effects.

Conclusions

Using Italian data on standardized test scores, we have shown that, when the characteristics of pupils vary among classes or schools, altering the length of the test not only reduces measurement error but can also change the ranking of individuals, classes or schools. Given the increasing focus on school accountability and on the relative performance of schools, our results suggest that the implications of test length on relative performance should be carefully considered when designing tests.

We have also investigated whether decomposing test scores into initial performance and performance decline can be used to separate the influence of cognitive and non-cognitive skills and found that our measure of cognitive skills – the math grade before the test – not only has a dominant influence on initial performance but also substantially affects performance decline. Our results cast doubts on recent interpretations of this decomposition.

While our findings concern primary school students taking a low stake test, they may have broader applicability, as suggested by the fact that our estimate of the mean effect of the position of questions on the test outcome is qualitatively similar to the one obtained by Borghans and Schils, 2012, and Borgonovi and Biecek, 2016, using data from PISA low stake tests taken by students who are about 5 year 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 in this paper.

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 are likely to gain from longer tests, while boys instead are likely to 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. (2017). Wait and See: Gender Gaps throughout Cognitive Tests. *JOLE Working Paper* 17679.

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

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.

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.

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.

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* (No. w24004). National Bureau of Economic Research.

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.

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. Available online at https://doi.org/10.1080/07350015.2016.1227711.

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.

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.

Tables and figures.

Table 1. Summary statistics

	Mean	St.Dev.
Y	52.81	49.92
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 (dummy)	0.327	0.469
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 are statistically equal.

Table 3. Estimates of the two-level model

	All	Males	Females
$E(\theta) = \alpha$	85.18***	86.21***	84.10**
$E(\beta) = \delta$	-0.060***	-0.083***	-0.035***
$Var(\theta)$	315.77***	304.74***	302.67***
$Var(\beta)$	0.035***	0.041***	0.029***
$Cov(\theta, \beta)$	0.65**	0.94***	0.49***
ρ (θ,β)	0.20	0.27	0.16
σ_{ε}	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 (α) and average slope parameters (δ), by number of books at home and immigrant status

	α Males	α Females	δ Males	δ 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 effect of question position P on Y. With and without interactions between P and cognitive and non-cognitive variables.

	(1)	(2)	(3)
P	-0.060***	-0.072***	-0.119***
	(0.006)	(0.008)	(0.014)
P * Conscientiousness		0.011**	0.011**
		(0.005)	(0.005)
P * Neuroticism		-0.004	-0.005
		(0.004)	(0.004)
P * Confidence		0.003	0.006
		(0.004)	(0.004)
P * Agreeableness		-0.001	-0.003
		(0.004)	(0.004)
P * Bullied		-0.009**	-0.008**
		(0.004)	(0.004)
P * Intrinsic motivation		-0.003	-0.004
		(0.003)	(0.003)
P * Extrinsic motivation		-0.006*	-0.003
		(0.003)	(0.003)
P * Math grade before the test		0.041***	0.033***
		(0.013)	(0.013)
Conscientiousness		-1.200***	-1.206***
		(0.164)	(0.164)
Neuroticism		-1.111***	-1.081***
		(0.122)	(0.123)
Confidence		3.549***	3.476***
		(0.144)	(0.146)
Agreeableness		0.153	0.203
		(0.138)	(0.138)
Bullied		0.012	-0.017
		(0.125)	(0.125)
Intrinsic motivation		-0.076	-0.039
		(0.099)	(0.099)
Extrinsic motivation		-0.756***	-0.830***
		(0.095)	(0.095)
Math grade		15.530***	15.736***
		(0.434)	(0.434)
P * Books			0.034**
			(0.012)
P * Female			0.047***

			(0.010)
P * Immigrant			-0.066**
			(0.019)
P * Small class			0.023*
			(0.012)
P * Trained to the test			0.027**
			(0.013)
Constant	85.145***	71.993***	79.675***
	(0.332)	(4.246)	(1.174)
Number of observations	353,142	353,142	353,142
Test that interactions of P with non-cognitive variables are jointly significant (p-value)		0.005	0.011
Test that non-cognitive variables are jointly significant (p-value)		0.000	0.000

Note: the regression in the first column includes question dummies and individual fixed effects. The regression in the second column includes question and class dummies, dummies for missing values of the relevant variables and the following additional controls: number of books in the house, gender, immigrant status, 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 and in the sub-sample of poolable questions.

	All questions	Subset of poolable questions
P	-0.060***	-0.054***
	(0.006)	(0.009)
P * Subset		
Cross-equation test of equality (p.val.)	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.

Figure 1. Distribution of random effects β and θ

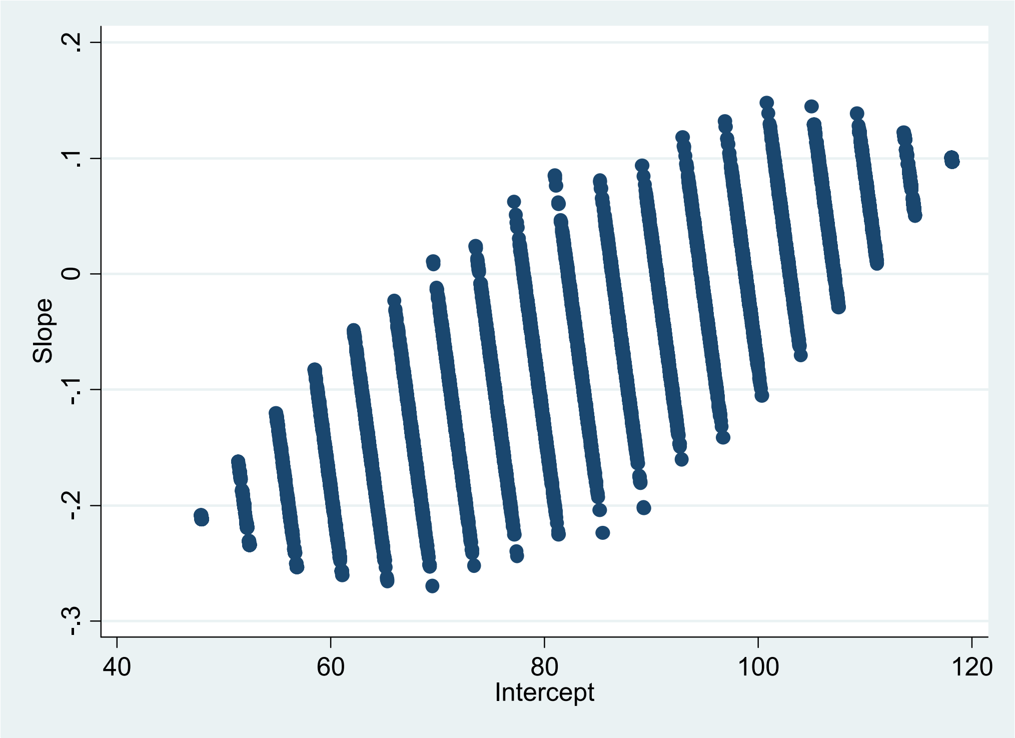


Figure 2. Percent of classes changing rank by at least 10 positions as N increases.

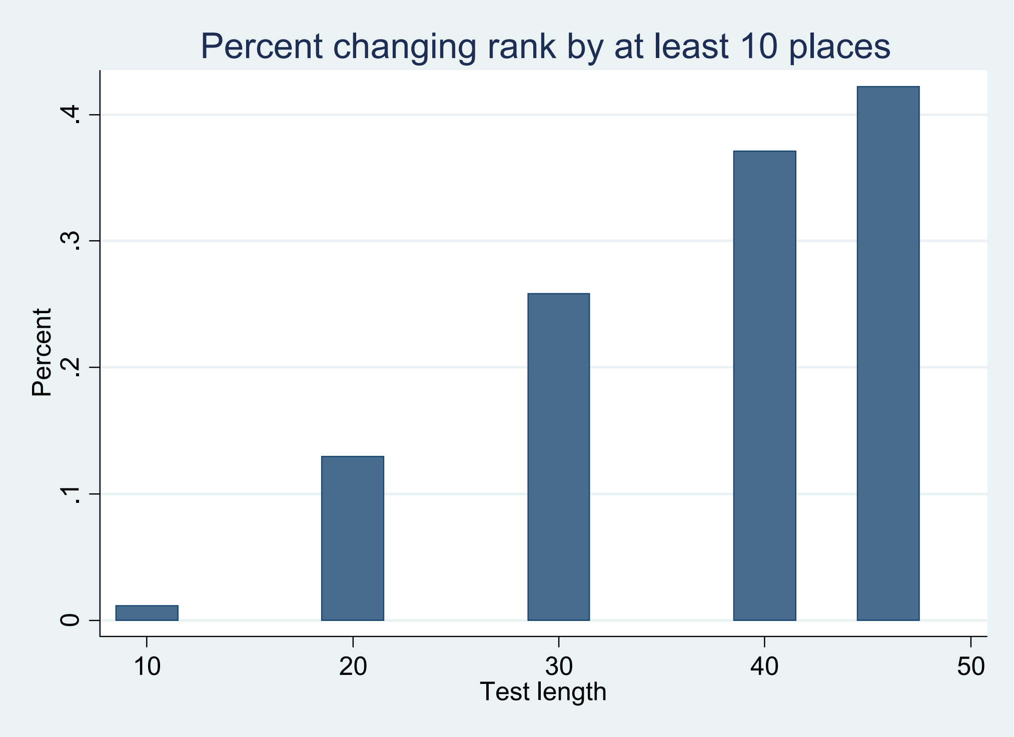
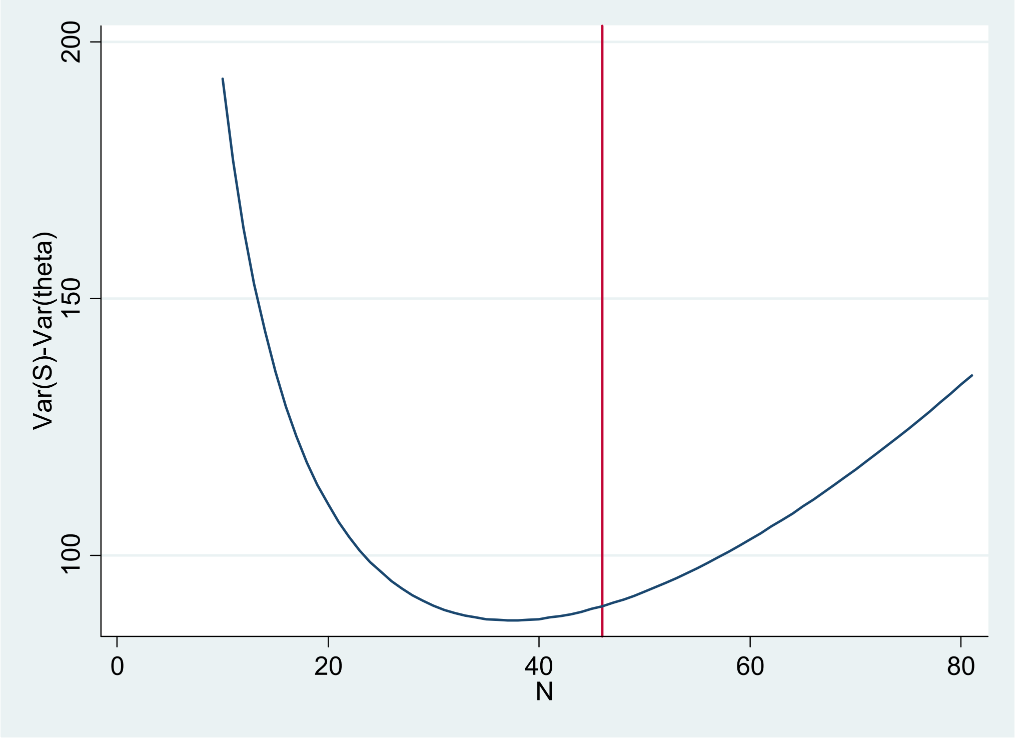


Figure 3. Estimated variance of the test score



Appendix

The information required to construct our indicators of personality traits 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.

For each relevant question, we use principal component analysis to extract the factor associated to the highest eigenvalue. We obtain the following indicators:

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).
7. *Poor social relations at school (bullied)*. The relevant question is: 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).