Economics of Education Review 63 (2018) 134–153

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Test scores, noncognitive skills and economic growth

Pau Balart[a], Matthijs Oosterveen[*,b], Dinand Webbink[b,c]

[a] Department of Business Economics, Universitat de les Illes Balears, Spain
[b] Department of Economics, Erasmus School of Economics, Erasmus University Rotterdam, Netherlands
[c] Tinbergen Institute, IZA, Bonn, Germany

ARTICLE INFO

Keywords:
Cognitive skills
Noncognitive skills
Long run economic growth

JEL classification:
I25
J24
O40

ABSTRACT

Many studies have found a strong association between economic outcomes of nations and their performance on international cognitive tests. This association is often interpreted as evidence for the importance of cognitive skills for economic growth. However, noncognitive skills also affect performance on cognitive tests. Following Borghans and Schils (2012), we exploit exogenous variation in the ordering of questions asked by the Programme for International Student Assessment (PISA) to decompose student performance into two components: the starting performance and the decline in performance during the test. The latter component is interpreted as a measure of noncognitive skills. Students from different countries exhibit differences in performance at the start of the test and in their rates of deterioration in performance during the test. Both components have a positive and statistically significant association with economic growth, and the estimated effects are quite similar and robust. Our results show that noncognitive skills are also important for the relationship between test scores and economic growth.

1. Introduction

Many studies have found a strong association between the economic outcomes of nations and their performance on international cognitive tests such as the PISA, TIMSS or PIRLS (see, for example, Hanushek & Kimko, 2000; Hanushek & Woessmann, 2008; Hanushek & Woessmann, 2012). This association is interpreted as evidence that cognitive skills are an important determinant of productivity and economic growth. However, the performance on cognitive tests is not only the result of cognitive ability, but is also influenced by noncognitive skills. Pioneers in intelligence testing like Thorndike and Wechsler already recognized that test takers might not exert maximal effort (Wechsler, 1940). Duckworth, Quinn, Lynam, Loeber, and Stouthamer-Loeber (2011) found that under low-stakes testing conditions, such as in the international cognitive tests, some individuals try harder than others. Moreover, scores of low performers can be substantially improved by offering a reward (e.g. Borghans, Meijers, & Ter Weel, 2008; Gneezy & Rustichini, 2000; Segal, 2012). The noncognitive skills that are important for test scores have also been found to be important for productivity and other social outcomes at the individual level (e.g. Heckman, Pinto, & Savelyev, 2013; Heckman & Rubinstein, 2001). This suggests that noncognitive skills might be an important omitted variable in the relationship between cognitive skills and the economic outcomes of nations. These two related issues make it unclear to what extent the strong association between the performance on international cognitive tests and economic growth should be interpreted as evidence on the importance of cognitive versus noncognitive skills. Given the differences in policy interventions required to foster cognitive and noncognitive skills (Cunha, Heckman, & Schennach, 2010), it is important to gain a better understanding of their respective roles in fostering economic growth.[1]

In this paper, we explore the effects of cognitive versus noncognitive skills on economic growth. We decompose the performance on an international test (PISA) into two components: the starting performance and the decline in performance during the test. This decomposition, recently introduced by Borghans and Schils (2012), exploits the random allocation of test booklets to students, which generates exogenous variation in the position of questions in the test. This specific feature of the test allows estimation of the decline in performance during the test that is not confounded by unobserved characteristics of questions, such as the difficulty of the test items. Borghans and Schils (2012) show that differences in the decline in performance during the test are related to noncognitive skills, such as motivation and ambition. They argue that the starting performance of the test score is a measure of cognitive skills that is not confounded by the personality factors that cause the decline in performance.

* Corresponding author.
E-mail addresses: pau.balart@uib.cat (P. Balart), oosterveen@ese.eur.nl (M. Oosterveen), webbink@ese.eur.nl (D. Webbink).

[1] Noncognitive skills have many different names in the literature. Soft skills, personality traits, character skills, noncognitive ability and socioemotional skills are often used.

https://doi.org/10.1016/j.econedurev.2017.12.004
Received 26 October 2017; Received in revised form 19 December 2017; Accepted 19 December 2017
Available online 20 December 2017
0272-7757/

Countries differ in both the starting performance and the decline in performance during the test and these differences are stable over time. We use the results of this decomposition to estimate the association between the two components and economic growth, and compare these findings with the estimated effect of test scores before the decomposition, which is the standard approach in the previous literature. For the analysis, we stay as close as possible to the seminal paper by Hanushek and Woessmann (2012), hereafter HW (2012). This study documented a strong association between test scores and economic growth, and provided convincing evidence that supports a causal interpretation of this relationship.

The exact interpretation of our findings critically depends upon which type of skills are measured by the two components. We argue that, at a minimum, our study answers the question whether noncognitive skills are partly responsible for the well-studied relationship between test scores and economic growth. To this end, we need a measure of noncognitive skills that is not confounded with cognitive skills. Most of our effort is therefore devoted to discussing the performance decline, which lends itself to two different types of interpretations. Let us call interpretation A that "*the performance decline captures noncognitive skills and does not capture cognitive skills*" while interpretation B admits the possibility that "*the performance decline captures both cognitive and noncognitive skills*". As recognized by Borghans, Duckworth, Heckman, and Ter Weel (2008), it is not only empirically, but also conceptually difficult to separate cognitive ability from noncognitive skills.[2] In fact, many aspects of personality and cognition are closely related.[3] Evidence presented in Section 3.1, however, supports interpretation A–that the performance decline measures only noncognitive skills.

The results from our cross-country growth regressions indicate that both components of test scores have a positive and statistically significant association with economic growth. In fact, their estimated effects are very similar in terms of magnitude. Moreover, we find that the effect of cognitive skills is approximately 40 percent smaller when we control for noncognitive skills, suggesting that noncognitive skills are important for explaining the relationship between test scores and economic growth.

Reverse causality is an obvious concern if one is interested in putting a causal interpretation on the results of macroeconomic growth regressions. In our study, we address the issue of reverse causality by applying the decomposition method to an international test administered in 1991. We find that our results are consistent with an effect of skills on growth and not vice versa.

To our knowledge, there are only two recent studies investigating the relationship between personality traits (in the form of average measures of patience per country) and productivity at the macroeconomic level (Dohmen, Enke, Falk, Huffman, & Sunde, 2016; Hübner & Vannoorenberghe, 2015). Our main contribution is to test whether the well-studied relationship between economic growth and test scores is mediated by noncognitive skills and to provide estimates of the effect of noncognitive skills. The lack of works studying the relationship between noncognitive skills and economic prosperity at the aggregate level contrasts with the abundance of studies at the individual level. One reason for this might be the lack of international comparable measures for noncognitive skills. Most studies on personality traits rely on self-reports of individuals, which complicates international comparisons. Performance based measures have the advantage that they do not suffer from the typical measurement issues related to self-reports, such as reference bias (e.g. Kautz, Heckman, Diris, ter Weel, & Borghans, 2014; Paulhus, 1984).

This study is organized as follows. Section 2 discusses the previous literature on the effect of cognitive skills on economic growth and the recent literature on the importance of noncognitive skills. Sections 3 and 4 explain the PISA-decomposition and the estimation of the cross-country growth regressions. The data used in the analyses are described in Section 5. Section 6 shows the main estimation results. Section 7 investigates the robustness of the results to using stricter measures of the performance decline and addresses concerns of reverse causality. Section 8 concludes.

2. Previous studies

2.1. The relationship between cognitive test scores and economic growth.

A large empirical literature has studied the impact of human capital on economic growth. One of the major challenges is to find a good proxy for human capital. As a consequence, many studies have used average educational attainment as their measure (see, for example, Barro, 1991; Cohen & Soto, 2007; Doménech & De la Fuente, 2006; Krueger & Lindahl, 2001; Sala-i Martín, Doppelhofer, & Miller, 2004; Sunde & Vischer, 2015). However, this proxy seems quite imperfect as it assumes that a year spent in school produces the same amount of human capital across all countries. Therefore, Lee and Lee (1995) and Hanushek and Kimko (2000) introduced a new approach that uses the performance on international cognitive tests as a proxy for human capital. The main advantage of this approach is that cognitive test scores can be considered as an output measure that captures what students have learned inside and outside of school. The basic cross-country growth specification in Hanushek and Kimko (2000) regresses the average economic growth of country c (G_c) for a specific period on their measure of human capital (H_c), GDP per capita at the beginning of the period (GDP_{0c}) and control variables (Z_{nc}) such as years of schooling and population growth:

$$G_c = \beta_0 + \beta_1 H_c + \beta_2 GDP_{0c} + \sum_n \delta_n Z_{nc} + \epsilon_c \quad (1)$$

This approach has been extended in a series of studies, which estimate Eq. (1) and have very similar results and interpretation (see Barro, 2001; Bosworth & Collins, 2003; Hanushek, 2013; Hanushek & Woessmann, 2008; Hanushek & Woessmann, 2011b; Hanushek & Woessmann, 2011c; Hanushek & Woessmann, 2012; Jamison, Jamison, & Hanushek, 2007). Eq. (1) is consistent with the endogenous growth models of Romer (1990) and Nelson and Phelps (1966). In these models, growth is attributed to the stock of human capital, which generates innovations or facilitates the adoption and imitation of new technologies. We focus on the most recent paper, HW (2012), which uses data on economic growth from 50 countries for the period 1960–2000 and cognitive test scores for the period 1964–2003. The authors find consistent evidence that cognitive test scores are strongly associated with economic growth and interpret this as indicating the importance of cognitive skills. The estimated effects of cognitive skills are large: a one standard deviation increase in test scores is associated with 1.25 to 2 percentage points higher average annual growth rate in GDP per capita across 40 years.

An important question is to what extent the association between test scores and economic growth reflects a causal effect of cognitive skills on economic performance. This is a difficult question because it is very hard to address typical identification issues like omitted variables, reverse causality and measurement error. However, HW (2012) show that the estimated effects of cognitive test scores on economic growth are robust to alternative estimation approaches, such as instrumental variables, differences-in-differences and longitudinal analysis of changes in cognitive test scores and in growth rates. Moreover, they

[2] They define cognitive skills as the "ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought". Noncognitive skills are referred to as patterns of thoughts, feelings and behaviors.

[3] Phelps (2006) argues that the mechanisms of emotion and cognition are intertwined from early perception to reasoning and Cunha and Heckman (2008) and Cunha et al. (2010) show noncognitive skills are important for the development of cognitive skills, but not vice versa.

argue that international test scores are not driven by differences in resources across countries and note that their estimation relies upon the assumption that the average scores for a country tend to be relatively stable over time. This leads them to conclude that differences in cognitive skills lead to economically significant differences in economic growth.

Although several studies find a consistent positive relationship between cognitive test scores and economic growth, a growing literature highlights the impact of noncognitive skills on test performance, making it difficult to know how these results should be interpreted.

2.2. Noncognitive skills, long-term individual outcomes and cognitive test scores

Many studies in psychology and a more recent literature in economics have established the importance of noncognitive skills for individual socioeconomic outcomes. Noncognitive skills are defined as relatively enduring patterns of thoughts, feelings and behaviors that reflect the tendency to respond in certain ways under certain circumstances (Roberts, 2009). These studies often use the Big Five inventory as measures of noncognitive skills (Costa & McCrae, 1992; John & Srivastava, 1999) and find that these measures are as predictive as cognitive measures for important outcomes such as schooling, wages, crime, teenage pregnancy, and longevity, even after controlling for family background and cognition (see for example Almlund, Duckworth, Heckman, & Kautz, 2011; Heckman, Stixrud, & Urzua, 2006; Heckman, 2008; Mueller & Plug, 2006). Intervention studies, like the Perry Pre School Program, provide evidence for a causal effect of changes in noncognitive skills on economic and social outcomes (Heckman et al., 2013). Further evidence on the importance of noncognitive skills for individual economic success can be found in Heckman and Rubinstein (2001), Borghans, Duckworth et al. (2008), Heckman and Kautz (2012) and Kautz et al. (2014).

Noncognitive skills have also been related to the performance of students on cognitive tests. The possibility that test takers might not exert maximal effort has been largely recognized by researchers on intelligence testing. For instance, Wechsler (1940) noted that intelligence tests not only measure intelligence and pointed out that the tendency to try hard on low stakes intelligence tests might derive from non-intellective traits, such as competitiveness and compliance with authority. More recently, Duckworth et al. (2011) provide evidence for the role of test motivation in intelligence testing. Observer ratings of test motivation, based on the behavior of adolescent boys completing intelligence tests, explains IQ-scores and reduces the predictive validity of IQ-scores for life outcomes, particularly for nonacademic outcomes. Their findings show that under low-stakes testing conditions some individuals try harder than others. Economists have also recognized that engaging in complex thinking is effortful and therefore motivation to exert effort affects the performance on achievement tests. For example, in Borghans, Meijers et al. (2008) subjects were given questionnaires to determine psychological traits and were asked to make trade-offs to determine relevant economic preference parameters. They found that preferences have a direct impact on cognitive test scores.[4] Moreover, various studies have found that offering a material reward can substantially improve scores on cognitive tests (Gneezy & Rustichini, 2000; Segal, 2012).

These findings have motivated using answering patterns to obtain measures of noncognitive skills that do not rely on self-reports. In addition to Borghans and Schils (2012), Hernández and Hershaff (2015) propose using skipped items in a non-penalized test to measure noncognitive skills; Hitt (2016), Zamarro, Cheng, Shakeel, and Hitt (2016) and Zamarro, Nichols, Duckworth, and D'Mello (2017) explore careless answering patterns in survey questionnaires; and Hitt, Trivitt, and Cheng (2016) find that skipped questions at six nationally-representative, longitudinal surveys of American youth are a significant predictor of later-life educational attainment net of cognitive ability. The use of self-reports to measure noncognitive skills has been challenged (Duckworth et al., 2011; Duckworth & Yeager, 2015; Paulhus, 1984; West et al., 2016). By relying on Borghans and Schils (2012), we can avoid the problems associated with self-reports and simultaneously measure cognitive and noncognitive skills.

3. The test score decomposition

Borghans and Schils (2012) developed an approach to decompose test scores into two elements: the starting performance and the decline in performance during the test. They observed that students perform worse on questions that appear later in the test. Because knowledge should be the same at the beginning and end of the test, they attribute the decline in performance during the test to motivation, which can be thought of as a noncognitive skill. One concern with this interpretation is that the performance decline might be related to unobservable characteristics, such as the difficulty of the test items.[5] If this were the case, the performance decline would be a consequence of cognitive skills rather than noncognitive skills.

To address this important issue, Borghans and Schils (2012) exploit the variation in the question ordering of the PISA test. As shown in Table 1, PISA 2006 has 13 different versions of the test (booklets), all of them containing four clusters of questions (test items). A booklet contains approximately 60 test items. Each cluster of questions represents 30 minutes of test time, which means each student undertakes two hours of testing. Students are allowed a small break after one hour, typically shorter than 5 minutes, where they are allowed to stand up and stretch. There are 13 clusters of test items (7 science, 2 reading and 4 math) and they are distributed over the 13 different booklets according to a rotation scheme. Each cluster appears in each of the four possible positions within a booklet once (OECD, 2009). This means that one specific test item appears in four different positions of four different booklets. For instance, cluster Science 1 is included in booklets 1, 9, 12 and 10 as respectively the first, second, third and fourth cluster. This rotation scheme generates exogenous variation in the question number (position in the test) of test items because the booklets are randomly assigned to students (OECD, 2009). In other words, the random assignment of booklets ensures that the positioning of questions is unrelated to student characteristics. The results of balancing tests are consistent with random assignment. Table A.1 shows estimates from separate regressions of background characteristics on booklet indicators. Almost all of the coefficients of these indicators are statistically insignificant at conventional levels and the F-tests for joint significance never reject the null hypothesis.

The exogenous variation in question position can then be exploited to estimate the decline in performance during the test by using the following fixed effects model:

$$P[Y_{ij} = 1] = F\left(\alpha_0 + \alpha_1 Q_{ij} + \sum_{j=2}^{J} \mu_j\right) \tag{2}$$

where Y_{ij} is an indicator for whether student i answered question j correctly, Q_{ij} is the position of question j in the version of the test answered by student i and μ_j is a question fixed effect that takes account of unobservable characteristics of question j such as difficulty. Conditional on question fixed effects, the variation we are exploiting lies within a question across different students. As such, the identifying assumption

[4] This finding is consistent with Borghans, Golsteyn, Heckman, and Humphries (2011) and Heckman and Kautz (2012), who find that personality variables explain roughly a third of explained variance in achievement tests.

[5] In fact, the sequencing of items from easy to difficult is used as an explicit strategy for sustaining morale (Duckworth et al., 2011).

Table 1
Rotation design of the 13 PISA booklets.

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	Science 1	Science 2	Science 4	Science 7
2	Science 2	Science 3	Math 3	Reading 1
3	Science 3	Science 4	Math 4	Math 1
4	Science 4	Math 3	Science 5	Math 2
5	Science 5	Science 6	Science 7	Science 3
6	Science 6	Reading 2	Reading 1	Science 4
7	Science 7	Reading 1	Math 2	Math 4
8	Math 1	Math 2	Science 2	Science 6
9	Math 2	Science 1	Science 3	Reading 2
10	Math 3	Math 4	Science 6	Science 1
11	Math 4	Science 5	Reading 2	Science 2
12	Reading 1	Math 1	Science 1	Science 5
13	Reading 2	Science 7	Math 1	Math 3

Source: OECD (2009).

becomes random variation in the position of a question across different students. This assumption is met due to the random allocation of booklets to students: we are comparing the performance of identical students on the same question in four different positions. The estimated parameter α_1 will not be biased by unobserved factors and can be interpreted as the decline in performance during the test. The decomposition of the test scores into the starting performance and the performance decline is based on the estimation of Eq. (2). We estimate Eq. (2) separately for each country by using a probit model and use the PISA weighting factors to ensure that the sample is representative.[6] The parameter α_0 measures the starting performance of a specific country, because the question numbers have been rescaled such that the first item is numbered as 0 and the last item as 1. Both components are robust to the definition of the start of the test. For instance, including the first five questions in the starting performance does not affect the estimates of the two components. We use all test items for estimating Eq. (2). Skipped and non-reached items were coded as incorrectly answered questions. This allows us to stay closer to the framework of HW (2012) in which uncompleted items were interpreted as incorrectly answered to compute final test scores.[7]

We have also estimated Eq. (2) using the average performance on all test items within a cluster as the outcome variable. In this analysis the clusters have been rescaled such that the first cluster is numbered as 0 and the fourth cluster is numbered as 1. With this approach the unit of randomization exactly matches the unit of analysis. The results are very similar to the results from our main approach. We find a correlation of 0.936 for the estimates of the starting performance of the two approaches, and a correlation of 0.964 for the decline in performance.

3.1. Interpretation of the two components

The main purpose of the decomposition is to generate two components that capture both types of skills relatively well. The performance decline is a measure of noncognitive skills, where the starting performance provides a measure of cognitive skills that, differently from test scores, is not confounded by the personality factors that cause the decline in performance.

Conceptually, obtaining a clean measure of cognitive skills is difficult. Cunha and Heckman (2008) and Cunha et al. (2010) show that noncognitive skills positively affect the accumulation of cognitive skills during childhood. Moreover, in our case it might be that the performance at the start of the test is influenced by ex-ante motivation as well. Although the starting performance, arguably, provides a better measure of cognitive skills than the final test scores, we cannot rule out that it is fully devoid of noncognitive skills. The consequence would be an attenuated estimate of the effect of noncognitive skills in the growth regressions.

This does not challenge, however, our investigation on whether noncognitive skills are important for the relationship between test scores and economic growth. For this purpose we only need a measure of noncognitive skills that is not contaminated with cognitive skills. Therefore it is conceptually important that cognitive skills do not affect the accumulation of noncognitive skills (Cunha & Heckman, 2008; Cunha et al., 2010). With regard to our particular measure, let us label the possibility that "*the performance decline captures noncognitive skills but does not capture cognitive skills*" interpretation A, while interpretation B admits the possibility that "*the performance decline captures both cognitive and noncognitive skills*". Our efforts are concentrated towards obtaining a measure of noncognitive skills that fits with interpretation A. The arguments provided below are consistent with the performance decline providing such a measure.

Borghans and Schils (2012) provide four arguments in support of interpretation A. First, the performance decline differs from the students' performance at the beginning of the test, which indicates that the two components measure different types of skills. Second, they show that the two components are stable for the years 2003 and 2006 and that there are differences between countries. This suggests that the two components are able to measure stable traits of the 15-year-old population of a country. Third, they show that the performance decline is related to specific noncognitive skills. With the data collected in the Dutch Inventaar 2010 study they find that students with higher levels of agreeableness (a Big Five personality trait), ambition and motivation towards learning have a smaller performance decline. Fourth, using data from the British Cohort Study 1970, they show that the performance decline predicts future outcomes above and beyond the pure test score.

Additional evidence comes from Balart and Oosterveen (2017). These authors noted that girls typically score better on reading tests than boys, but perform worse in science and math (Caplan, Crawford, Hyde, & Richardson, 1997; Cornwell, Mustard, & Van Parys, 2013; Dee, 2007; Fryer & Levitt, 2010; Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 1988; Kimura, 2004; Quinn & Cooc, 2015). They argued that if the performance decline were in fact induced by cognitive skills, then we should observe girls experiencing a less pronounced decline in reading, while boys would have a less pronounced decline when answering math and science questions. Balart and Oosterveen (2017), however, found the opposite: girls exhibited a less pronounced decline than boys in both reading and in math/science. Specifically, using the PISA 2009, they find that in 66 (62) out of the 74 countries, girls perform better (worse) in reading (math-science) than boys at the start of the test. However, there is no single country in which boys exhibit a statistically significant lower decline in performance than girls either in reading or in math-science. Girls exhibit a less pronounced decline than boys in 68 countries for reading (statistically significant for 40) and in 68 countries for mathematics and science (statistically significant for 46). A smaller performance decline independent of one's ability in a topic strongly supports the argument that the decline is not driven by cognitive skills. Moreover, it is consistent with gender differences in noncognitive skills found in previous research.[8]

[6] Estimating Eq. (2) with OLS gives very similar results. In fact, the correlation of the components estimated with probit and OLS equals 0.996 for the starting performance and 0.969 for the performance decline. Notice that despite using a probit model with question fixed effects, the incidental parameter problem does not apply. The number of fixed effects to be estimated (questions) remains constant when increasing the number of observations (students).

[7] Borghans and Schils (2012) note that it is unclear which type of skills determine that test items are not reached. In Section 7 we will investigate the sensitivity of our results to alternative ways of dealing with non-reached questions.

[8] Girls are found to have more self-discipline (Duckworth & Seligman, 2006), have less behavioral problems (Jacob, 2002), to be less prone to overconfidence (Niederle & Vesterlund, 2007), show more developed attitudes towards learning (Cornwell et al., 2013) and report higher levels of extraversion, agreeableness, and conscientiousness (Schmitt, Realo, Voracek, & Allik, 2008).

Another element in support of interpretation A arises from the growing body of research that has used the decomposition strategy proposed by Borghans and Schils (2012). Using an epidemiological approach, Nollenberger and Rodríguez-Planas (2018) show that gender differences in the starting performance are related to gender equality of the country of origin of second generation immigrants while the same is not true for gender differences in the performance decline. They interpret that gender gaps in test scores are affected by social gender norms through cognitive skills rather than noncognitive skills. Zamarro, Hitt, and Mendez (2016) show that, at the country level, the performance decline is associated with other non-self reported measures of student effort such as careless answering patterns and non-response in the student background questionnaire after PISA. The non-challenging nature of filling out a questionnaire makes it unlikely that this is driven by cognitive skills (Hernández & Hershaff, 2015; Hitt, 2016; Hitt et al., 2016). As they argue, this is strengthened by the fact that careless answering patterns do not exhibit a higher correlation with test scores in reading than with non-reading ones.

Finally, following Linton (1945), Hofstede and McCrae (2004) and Benet-Martínez and Oishi (2008), we make use of the similarities between culture and noncognitive skills. Similar to noncognitive traits, culture is defined in terms of behavior and is transmitted from generation to generation.[9] Méndez (2015) directly associates culture with differences in noncognitive skills and exploits cultural variations in second-generation immigrants to show that differences in cultural values and accompanying noncognitive skills are related to differences in PISA test scores. By contrast, cognition does not have a strong link with culture. Indeed, psychologists distinguish two types of second-order factors of cognitive ability: fluid intelligence and crystalized intelligence (Cattell, 1987). Only the second is partially influenced by culture.[10] We provide additional evidence on the difference between the two components by regressing the cultural dimensions of Hofstede and Hofstede (2001) upon the standardized starting performance and performance decline in Table 2.[11]

We find that the performance decline has a stronger association with the cultural values than the starting performance. Whereas the starting performance is only significantly related to higher levels of individualism at the 10%-level, the performance decline shows strong associations with measures of uncertainty avoidance, long term orientation and indulgence.[12] These results suggest that the performance decline is smaller in countries with: (i) more thriftiness, perseverance for achieving results and higher efforts in modern education (long term orientation), (ii) less preference for leisure-time, more control on the gratification of desires and stricter social norms (indulgence) and (iii) more positive preference towards uncertainty, ambiguity and curiosity (uncertainty avoidance).[13] In sum, this can be interpreted as evidence that the performance decline is related to motivation, thriftiness, and less-preference towards leisure and certainty which are related to time- and risk-preferences. Remarkably, these are a large subset of the non-cognitive skills that Heckman (2008) lists as being related to earnings, employment, college attendance, and other socioeconomic outcomes.

3.2. Differences between countries and years

Results of the decomposition of the PISA test of 2006 are shown in Table A.2. Eq. (2) is used for computing the probability of correctly answering the first and the last question of the PISA test. Column (1) shows the average of the PISA 2006 test scores, column (2) shows the probability of correctly answering the first question, column (3) shows the probability of correctly answering the last question and column (4) shows the difference between these two probabilities. Column (2) can be interpreted as the starting performance of a country and column (4) can be interpreted as the performance decline. Countries are ranked with respect to the latter from high to low.

We observe that there are large differences between countries. Columbia and Uruguay have the largest decline in performance. That is, their probability to answer the last question correctly is 30 percentage points lower than their probability to answer the first question correctly. Within the top ten of countries with the largest decline we observe six countries from South America. Among the countries with the least pronounced performance declines, we observe Northern European and Asian countries. Moreover, Table A.2 indicates that these differences between countries are important for the total test score. We observe that Azerbaijan and Brazil have a very similar starting performance. However, the decline in performance for students in Brazil is much larger than for students in Azerbaijan. This translates into a difference on the PISA test of more than 19 points. This difference is shown in Fig. 1, where we use locally weighted scatterplot smoothing to visualize the performance decline and the starting performance for the two countries. This flexible nonparametric method also suggests that, without putting assumptions on the process that generated the data, the linear specification in Q_{ij} used in Eq. (2) seems to be a good approximation.

We have also decomposed the test scores for PISA 2003 and 2009 using the same procedure as was used for the PISA 2006. Table 3 shows the correlations between the different components for the three years. The correlations of the estimated starting performances (performance declines) over time are shown in bold. All of them are above 0.91. As indicated by Borghans and Schils (2012), a high correlation between the starting performances (performance declines) over the years suggests that these components capture some of the traits of the 15-year-old population of a country. Consistent with the literature on non-cognitive skills, the correlation between the starting performance and the performance decline is much lower, which indicates that the two components measure different traits.

4. Estimation of the relationship between skills and economic growth

The starting point of our empirical analysis of the effect of skills on economic growth is the standard cross-country growth regression as shown by Eq. (1). The main previous studies aggregate scores from all available international cognitive tests and use this as a measure for cognitive skills (see Section 2.1). We label the aggregate test score from HW (2012) as the HW-index. In this study we decompose the scores on an international cognitive test into the starting performance (S_c) and the performance decline during the test (PD_c). Therefore, instead of using test scores as a unidimensional proxy for human capital (H_c), we use the two components: $H_c = f(S_c, PD_c) + \nu_c$. We include these two components into the cross-country growth regression to re-estimate Eq. (1):

[9] For instance, Guiso, Sapienza, and Zingales (2006) or Fernandez and Fogli (2009) define culture as customary beliefs, values and actions that social groups transmit fairly unchanged from generation to generation. Intergenerational transfers of noncognitive skills is argued by Heckman (2008), he shows enhancements of family environments (socioemotional nurturing) improve child outcomes, of which personality traits are the most important channel.

[10] For instance, the Raven Progressive Matrices Test, a test commonly used to measure fluid intelligence, is referred to as “culture-free”

[11] We use PISA 2009 to maximize the number of countries. Using the other waves of PISA does not change our results. We include initial GDP per capita in 1990 to control for economic development. Controlling for an OECD indicator or GDP per capita in 1960 gives almost identical results, but controlling for GDP per capita in 1960 sharply decreases our sample size. The components are standardized with mean 0 and standard deviation 1.

[12] Interestingly, Gorodnichenko and Roland (2016) provide evidence for a causal effect of individualism on long run growth. Gorodnichenko and Roland (2011) show that individualism is the only measure of culture, among the ones they consider, that has a robust effect on growth.

[13] See the Appendix for explanations on the six cultural dimensions of Hofstede and Hofstede (2001).

Table 2
Regressions of Hofstede's cultural dimensions on the starting performance and performance decline.

	(1) Power distance	(2) Individualism	(3) Masculinity	(4) Uncertainty avoidance	(5) Long term orientation	(6) Indulgence
Starting performance	−1.626	5.818*	0.751	2.579	6.406	0.644
	(−0.33)	(1.81)	(0.19)	(0.48)	(1.37)	(0.16)
Performance decline	2.426	1.759	0.303	−7.516**	11.37***	−9.712***
	(0.89)	(0.70)	(0.11)	(−2.10)	(4.68)	(−3.37)
N	53	53	53	53	56	55
Adj. R^2	0.296	0.469	−0.056	0.165	0.255	0.225
Std.Dev. *Y*	22.23	24.19	20.58	53.28	21.97	19.86

Notes: t statistics in parentheses, heteroskedasticity robust standard errors * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ Regressions include a constant and initial GDP per capita. The starting performance and performance decline are standardized with mean 0 and standard deviation 1.

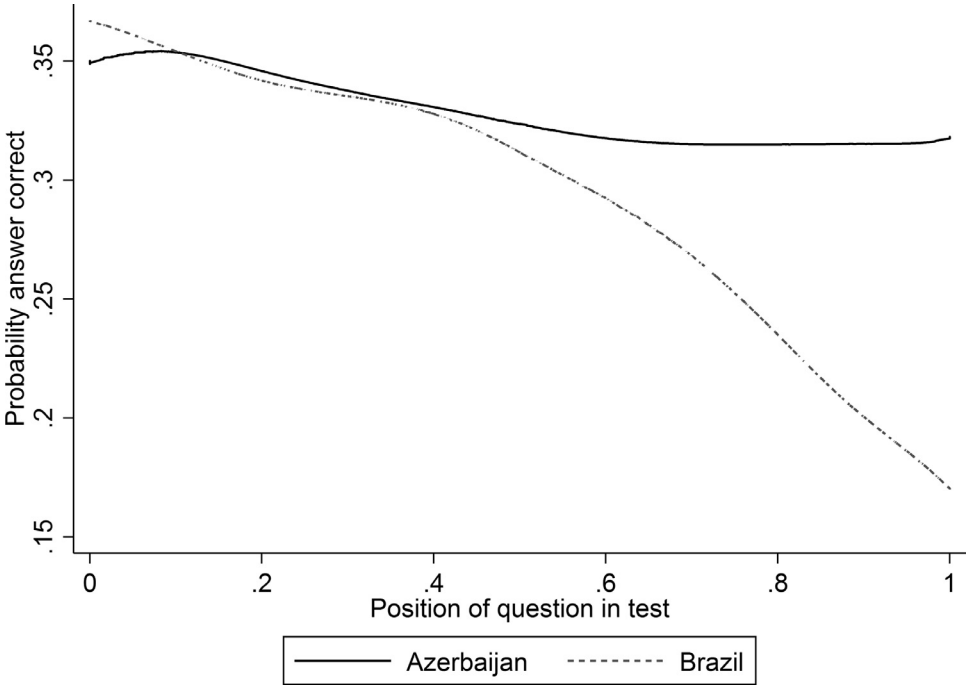

Fig. 1. The decline in performance for Azerbaijan and Brazil.
Notes: The figure is based upon PISA 2006 and uses LOWESS to visualize the relationship between *Y* and *Q* of Eq. (2) with a bandwidth of 0.8*N*, the program default. We do not use PISA weights or question fixed effects for the computation of this figure.

Table 3
Correlations between the starting performance and performance decline for PISA 2003, 2006 and 2009.

	Starting performance 2003	Decline 2003	PISA 2006	Starting performance 2006	Decline 2006	Starting performance 2009	Decline 2009
Starting performance 2003	1.000						
Decline 2003	0.426	1.000					
PISA 2006	0.860	0.745	1.000				
Starting performance 2006	**0.967**	0.521	0.920	1.000			
Decline 2006	0.526	**0.947**	0.716	0.584	1.000		
Starting performance 2009	**0.950**	0.480	0.856	**0.912**	0.501	1.000	
Decline 2009	0.539	**0.911**	0.760	0.631	**0.923**	0.462	1.000

Notes: The components are estimated using Eq. (2) with PISA weights.

$$G_c = \beta_0 + \beta_1 S_c + \beta_2 PD_c + \beta_3 GDP_{0c} + \sum_n \delta_n Z_{nc} + \varepsilon_c \quad (3)$$

When estimating Eq. (3), we try to stay as close as possible to HW (2012). We use the same data on economic growth, the same growth period (1960–2000), identical covariates, estimate the same model specifications, and start with the same sample of countries. However, the decomposition method that we apply in this paper exploits a specific feature of the PISA test, namely the random allocation of the PISA booklets (see Section 3). Hence, we can apply the decomposition method only to one of the tests included in the HW-index. This has two implications for the estimations. First, the sample of countries that participated in the PISA test differs from the sample used in HW (2012). As a first step in our analysis, we check whether the reduction in sample size from 50 to 37 countries, due to our reliance on the PISA, changes the results obtained in HW (2012). The second implication is that we use the PISA test only for measuring skills, and not the complete set of tests used for the HW-index. However, the PISA scores are highly correlated with the HW-index ($r = 0.91$). Further we will show below that re-estimating the main models from HW (2012) with PISA scores instead of the HW-index delivers very similar results. This suggests that PISA scores are a good proxy for the HW-index and, therefore, we use the PISA scores for estimating Eq. (1). Next, we decompose these PISA scores into the two components and we use these two components for

estimating Eq. (3). We estimate Eq. (3) with OLS and report robust standard errors. This analysis naturally relies upon the assumption that the distribution of both the starting performance and the performance decline across countries remained relatively stable over time, which is supported by Table 3.

As we are using a two-step estimation approach it could be argued that the standard errors should be adjusted because the regressors are not fixed (see e.g. Murphy & Topel, 2002). However, due to the large number of observations used in the estimation of Eq. (2), which is the number of students times the number of test items, the estimates for the starting performance and the performance decline are very precise, and can be considered as fixed (see Table A.3 for the standard errors of the two components and the number of students participating in PISA 2006 per country).[14]

5. Data

The data used in the analysis come from various sources. The HW-index is from HW (2012). This index aggregates all available math, science and reading scores from international cognitive tests that took place in the period 1964–2003 for 50 countries.[15]

In addition, we use data collected by the Programme for International Student Assessment. PISA is a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. The key subjects of the test are reading, science and math. The first PISA study took place in 2000. The method for decomposing test scores into a cognitive and a noncognitive component is applied for countries that participated in PISA 2003, 2006 and 2009.[16] We start our analysis with PISA 2006 which allows us to include 37 countries that were included by HW (2012). We standardize the decomposed test scores and the total PISA score separately to set the mean and standard deviation equal to the HW-index, allowing us to directly compare the size of our estimates to those of the HW-index.

We follow HW (2012) for sources on the other data. Real GDP per capita comes from version 7.1 of the Penn World Tables (Aten, Heston, & Summers, 2009).[17] Data on years of schooling are taken from the most recent version of the Barro and Lee dataset (Barro & Lee, 2013, version 2.1). Further control variables used by HW (2012) are regional indicators and two proxies for the quality of economic institutions: openness of the economy and protection against expropriation. For the regional indicators we follow the classification of HW (2012). The measure of openness is the Sachs, Warner, Åslund, and Fischer (1995) index reflecting the fraction of years between 1960 and 1992 that a country was classified as having an economy open to international trade.[18] For the data on protection against expropriation Acemoglu, Johnson, and Robinson (2001) is followed, the measure is an index between 0 and 10 averaged over 1985–1995. A higher score on this index means that there is more protection against expropriation. Two other controls that are used are fertility, obtained from World Development Indicators WorldBank (2002), and tropical location measured as the proportion of a countries' area located in the tropics (Gallup, Sachs, & Mellinger, 1999). Table A.3 provides the data per country on GDP growth, the HW-index and the two components of the PISA test for the sample of 37 countries used in Section 6.2.

6. Main estimation results

This section shows the main estimation results in three steps. First, we replicate the main analysis of HW (2012) for the sample of countries for which it is possible to decompose the PISA test. Second, we include the two components from the decomposition in the main estimation models.[19] Third, we repeat the latter analysis and extent the sample towards 55 countries.

6.1. Replication of previous cross-country growth regressions using PISA

In the first step of our analysis we check whether the estimation results obtained by HW (2012) change when we use scores of PISA 2006 instead of the HW-index. The results could, in theory, change because we are going from 50 to 37 countries, or because we use the PISA score instead of the HW-index. To show that none of these changes drive our results, we replicate the main models from HW (2012) using the sample of 37 countries. Panel A of Table 4 shows the results from models that use the HW-index, Panel B shows the results when using the PISA 2006 scores.

Panel A of Table 4 shows that the results for the growth regressions with the HW-index for the restricted sample are very similar to the results for the unrestricted sample in Table 1 of HW (2012). Column (1) shows the effect of years of schooling on economic growth. The estimated effect is statistically significant and suggests that an additional year of schooling increases the average annual growth rate in GDP per capita across 40 years with 0.2 percentage point. Column (2) shows the results from a model in which the HW-index is used as a proxy for human capital instead of years of schooling. A one standard deviation increase in cognitive test scores is associated with 2.3 percentage point increase in the annual growth rate of GDP per capita over 40 years. Similar to what HW (2012) found, replacing years of schooling with cognitive test scores increases the explained variance from one to three quarters. In column (3), we report results from a model that includes both proxies of human capital. The estimate of cognitive test scores is similar to that in column (2), but the estimated coefficient of years of schooling is no longer statistically significant. In columns (4)–(9) we report estimates from alternative specifications of the model; column (4) uses average years of schooling over the period 1960–2000 instead of the years of schooling in 1960, column (5) controls for outliers, column (6) includes eight regional indicators, column (7) includes measures for the openness of the economy and protection of property rights, column (8) adds fertility and tropical location as additional controls and column (9) controls for GDP per capita in logs instead of levels. These various estimates confirm that cognitive test scores are associated with economic growth and the results for our sample of 37 countries are very similar to the results for the full sample used by HW (2012).

In Panel B of Table 4 we show estimates of models that use PISA 2006 scores instead of the HW-index. We find that the estimated effects are very similar to those in Panel A. In fact, the estimates for the PISA scores are always within the 95% confidence interval of the estimates for the HW-index, which can be explained by the high correlation ($r = 0.91$) between the PISA scores and the HW-index. This indicates that PISA 2006 is a good proxy for the HW-index in models that explain

[14] The maximum likelihood estimation of Eq. (2) gives us consistent estimates. Since the number of observations is large, we can be confident that the ML-estimates have reached their true values. For computational tractability, standard errors in Table A.3 are computed using sample PISA weights but not their 80 replicates. Using the 80 replicates does not substantially increase the size of standard errors. For instance, the standard error of the starting performance and the performance decline of Iceland increase from 0.0437 to 0.0476 and from 0.0101 to 0.0132, respectively.

[15] See the Appendix of HW (2012) for further details on the computation of this measure.

[16] We choose not to use the two most recent PISA waves (2012 and 2015) because fewer countries participated in these waves and to mitigate concerns regarding reverse causality.

[17] Real GDP per capita for Tunisia was not available for 1960, so we used data from 1961 onwards.

[18] Because Romania was not available in Sachs et al. (1995), we used Romanian data from Sachs and Warner (1997) for the period 1965–1990.

[19] We start this analysis using test scores from the PISA 2006, but our results do not change when we use the PISA 2003, although the number of countries does decrease to 31. Using the PISA 2009, the sample increases to 40 countries and the results remain qualitatively unchanged.

Table 4
Growth regressions with the HW-index and PISA scores using the PISA sample.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Panel A: HW-index as a measure of human capital with restricted sample									
HW-index		2.256***	2.260***	2.310***	2.186***	1.144**	1.399***	1.378***	2.213***
		(9.15)	(9.22)	(9.14)	(9.59)	(2.71)	(3.74)	(3.74)	(9.92)
Years of schooling	0.187*		−0.00375	−0.0320	−0.0661	0.0420	0.0582	0.0115	−0.0250
	(1.80)		(−0.05)	(−0.49)	(−1.02)	(0.51)	(0.85)	(0.15)	(−0.30)
N	37	37	37	37	37	37	36	36	37
Adj. R^2	0.208	0.730	0.722	0.723	0.809	0.784	0.770	0.799	0.717
Panel B: PISA 2006 as a measure of human capital with restricted sample									
PISA 2006		2.282***	2.245***	2.235***	2.223***	1.181***	1.265***	1.220**	2.299***
		(7.98)	(7.59)	(6.84)	(7.16)	(3.53)	(2.83)	(2.74)	(9.73)
Years of schooling	0.187*		0.0426	0.0316	−0.0241	0.0654	0.104	0.0526	0.0305
	(1.80)		(0.53)	(0.38)	(−0.28)	(0.84)	(1.46)	(0.63)	(0.35)
N	37	37	37	37	37	37	36	36	37
Adj. R^2	0.208	0.700	0.694	0.692	0.691	0.803	0.754	0.781	0.728

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable: average annual growth rate in GDP per capita, 1960–2000 Regressions include a constant and GDP per capita in 1960.

[a] Measure of years of schooling refers to the average over the period 1960–2000.

[b] Controlling for outliers by using rreg command in Stata.

[c] Includes indicators for the eight world regions.

[d] Controlled for openness of economy and protection against expropriation.

[e] Controls in *d* plus fertility and tropical location.

[f] GDP per capita 1960 measured in logs.

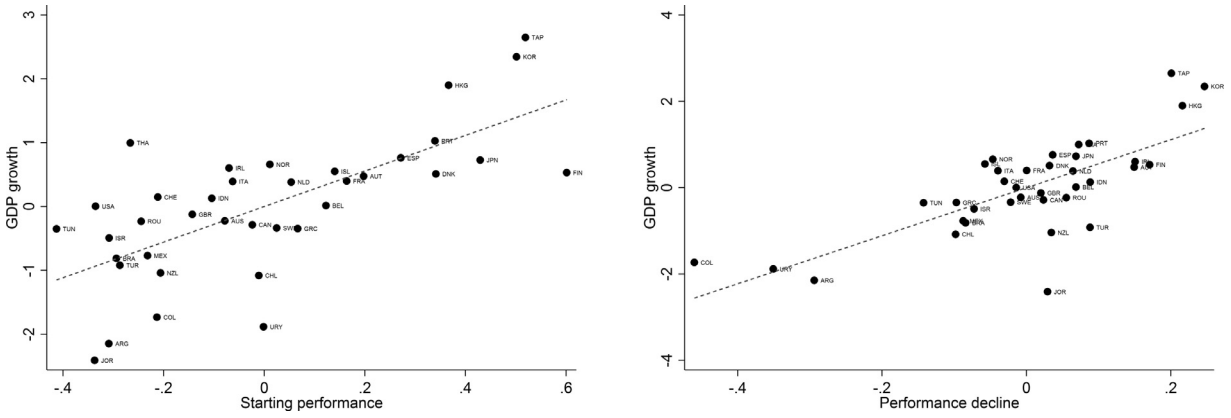

Fig. 2. The association between the conditional starting performance and the conditional decline in performance with economic growth for the period 1960–2000.

differences in economic growth between countries.

In sum, we find that the results obtained by HW (2012) are robust to using the sample of countries participating in PISA 2006 and to using PISA scores instead of the HW-index, suggesting that, within the framework of HW (2012), we can use the PISA scores as a proxy for the HW-index.

6.2. The relationship of the starting performance and the performance decline with economic growth

In this section, we present the main estimation results of models that include the two components obtained by decomposing the PISA scores. Fig. 2 provides a first impression of the relationship between these two components and economic growth, conditional on initial GDP and years of schooling. The left panel shows a positive association between the starting test score level and GDP growth. However, the right panel shows a very similar association between the decline in performance and GDP growth, which suggests that noncognitive skills are also related to economic growth. Below, we confirm that the association is not solely driven by the three Asian countries in the upper right corner and the three Southern American countries in the lower left corner of Fig. 2.

Table 5 replicates the model from Table 4 using the starting performance and the decline in performance as the main explanatory variables. Columns (1) and (2) show estimates of the relationships presented in Fig. 2. The starting performance is positively and significantly associated with economic growth. The estimated effect is somewhat smaller than the previous estimate from the model that uses the PISA score in Table 4, suggesting that the starting performance is less confounded by personality factors than the PISA score.

The performance decline is also positively and significantly associated with economic growth. Moreover, the size of this association is quite similar to that obtained for the starting performance. A comparison of columns (1) and (2) also reveals that years of schooling is associated with economic growth only in column (2). Years of schooling is more highly correlated with the starting performance ($r = 0.64$) than with the performance decline ($r = 0.38$), perhaps indicating that the latter also captures factors that are independent of what is learned at school, which is consistent with the idea that noncognitive skills are more affected by out-of-school influences than cognitive skills.[20]

In column (3) of Table 5, we report estimates from a model that includes both components on the right-hand side. We find that the starting performance and performance decline are positively and

[20] The estimated effects become somewhat smaller but remain statistically significant at the 1% significance level if potential outliers are excluded from the analysis (Taiwan, Hong Kong, Korea, Columbia, Uruguay and Argentina).

Table 5
Regressions of economic growth on the starting performance and performance decline.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	2.143***		1.330***	1.345**	1.049***	0.766*	0.701	0.249	1.934***
performance	(4.84)		(2.76)	(2.69)	(3.05)	(2.00)	(1.38)	(0.49)	(4.45)
Performance		1.872***	1.300***	1.257***	1.467***	0.680*	0.917***	1.120***	0.701**
decline		(6.36)	(4.44)	(4.17)	(5.14)	(1.83)	(3.09)	(4.95)	(2.64)
Years of	0.0415	0.173*	0.0871	0.0651	0.0483	0.0793	0.131*	0.0839	0.0344
schooling	(0.53)	(1.86)	(1.23)	(0.96)	(0.64)	(1.00)	(1.93)	(1.14)	(0.44)
N	37	37	37	37	37	37	36	36	37
Adj. R^2	0.579	0.623	0.726	0.721	0.746	0.787	0.754	0.798	0.716

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1960–2000; Regressions include a constant and GDP per capita in 1960 [a–f] See the description of Table 4.

significantly associated with economic growth, but these estimates are considerably smaller than those reported in columns (1) and (2). The estimated effect of cognitive skills (the starting performance), falls approximately 40 percent as compared to the estimate in column (1), implying that noncognitive skills are partially driving the relationship between test scores and economic growth found by previous studies. These results are consistent with the models of skill formation of Cunha and Heckman (2008) and Cunha et al. (2010): noncognitive skills appear to be important for the development of cognitive skills, but not vice versa. By using the starting performance rather than test scores, we are correcting for the measurement error induced by noncognitive skills in the generation of current test scores but not for the accumulated effect of noncognitive ability on the development of cognitive skills. As a consequence, once we include a measure of noncognitive skills in the growth regression model, we are explicitly accounting for the association between growth and cognitive skills for which noncognitive skills are responsible. This explains why the coefficient of the corrected measure for cognitive skills is only marginally smaller in column (1), while it is reduced by approximately 40 percent after including noncognitive skills in the growth regressions in column (3).

In the remaining columns of Table 5, we report results from using the different specifications introduced in Table 4. In general, the results are quite robust to these sensitivity tests. Controlling for average years of schooling over the period 1960–2000 does not change the estimates. Moreover, the results do not appear to be driven by outliers or by countries that belong to certain regions. However, when including regional indicators, noncognitive skills are only significant at the 10% level, which is consistent with the idea that cultural differences are an important determinant of noncognitive skills embedded in the performance decline. The estimated effect of the performance decline is robust to the inclusion of additional controls in columns (7) and (8). We observe that the starting performance is no longer a significant determinant of growth when controlling for the quality of economic institutions. A possible explanation for this result is that better institutions go hand in hand with better schools, capturing some of the effects of cognitive skills that the starting performance is intended to measure. Finally, controlling for the initial GDP level in logs in column (9) increases the estimated effect of starting performance and reduces the estimated effect of the performance decline. Both components, however, remain significant at conventional levels. This same pattern of results for cognitive skills was documented by HW (2012) and in Table 4. HW (2012) noted that specification (9) is more consistent with neoclassical growth models in which human capital affects steady-state levels of income but not growth rates.

By using PISA 2009 and the average economic growth for the period 1970–2010, we can extend our sample to 55 countries.[21] Table 6 shows the results are virtually identical compared to Table 5.[22] Not only are the estimates for both components of comparable magnitude, but the relationship between the performance decline and economic growth appears to be more robust to the inclusion of controls related to the quality of economic institutions than is the starting performance. The only difference is a marginally insignificant coefficient for the starting performance when we control for the regional dummies. Although the point estimate for the starting performance is somewhat larger than the estimated effect of the performance decline in column (6), it is less precise.

In sum, we find that both the starting performance and the performance decline are positively and significantly associated with economic growth. The estimated effects are similar in terms of magnitude, where the differences between the two components in both Tables 5 and 6 are statistically insignificant except in columns (8) and (9). The estimated effect of the performance decline is more robust to the inclusion of the quality of economic institutions than is the effect of the starting performance level. Finally, the estimate for the starting performance drops by roughly 40 percent after the inclusion of the performance decline, which implies that noncognitive skills are partly responsible for the relationship between test scores and economic growth.

7. Robustness checks

In this section we perform two types of analyses. First, we focus on measurement issues and test whether our results are robust to more restrictive (and alternative) computations of the performance decline. We use our large sample because it gives us more statistical power. Second, we perform two tests of whether the observed associations reflect a relationship from skills to growth, or from growth to skills.

7.1. Stricter measures of the performance decline

A concern with the interpretation of our previous analysis is that cognitive skills might have a direct effect on the performance decline. The correlation between the two components is 0.46, which could indicate that the performance decline is also capturing cognitive skills. We address this concern by using two stricter measures of the decline in performance. The first of these measures only exploits variation that is orthogonal to the starting performance. More precisely, we regressed the performance decline on the starting performance for all the countries participating in PISA 2009 and used the residuals of this regression as a corrected measure. As personality factors can boost the acquisition of cognition (Cunha & Heckman, 2008), the estimates obtained when using this new measure in Eq. (3) can be thought of as a lower bound for the relationship between noncognitive skills and economic growth.

Table 7 shows that the results are qualitatively similar to those in

[21] PISA 2009 was the first wave where countries where allowed to include a subset of seven "easier" booklets. These seven easier booklets were identical to the standard booklets, but each had one reading cluster replaced by an easier reading cluster. We computed the performance decline on the set of 13 standard booklets that where the same for every participating country.

[22] In this sample we also include China, India and Venezuela. These three countries only sample students within certain regions for PISA 2009. Results are qualitatively similar when we exclude these three countries.

Table 6
Regressions of economic growth on the starting performance and performance decline, maximizing our sample size.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting performance	1.674*** (3.40)		0.952** (2.61)	0.959** (2.62)	0.743* (1.98)	0.580 (1.61)	0.636 (1.28)	−0.00459 (−0.01)	1.577*** (5.02)
Performance decline		1.616*** (4.69)	1.299*** (4.32)	1.301*** (4.30)	1.287*** (4.45)	0.535* (1.73)	1.094** (2.13)	1.264** (2.55)	0.669*** (2.70)
Years of schooling	−0.00391 (−0.03)	−0.0407 (−0.43)	−0.0580 (−0.59)	−0.05298 (−0.56)	0.00447 (0.05)	0.0335 (0.34)	0.0341 (0.34)	−0.00407 (−0.04)	−0.0107 (−0.13)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.415	0.525	0.568	0.567	0.573	0.725	0.594	0.652	0.677

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table 7
Regressions of economic growth on components of test scores using an orthogonal measure of the performance decline.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting performance	1.674*** (3.40)		1.599*** (4.13) [4.07]	1.608*** (4.03) [3.98]	1.384*** (4.09) [2.38]	0.847** (2.52) [2.12]	1.181* (1.82) [1.82]	0.625 (0.91) [0.88]	1.910*** (5.85) [5.18]
Performance decline		1.176*** (3.55) [3.72]	1.119*** (4.32) [3.76]	1.121*** (4.30) [3.78]	1.109*** (4.45) [3.22]	0.461* (1.73) [1.52]	0.942** (2.13) [2.01]	1.088** (2.55) [2.33]	0.576*** (2.70) [2.42]
Years of schooling	−0.00391 (−0.03)	0.00206 (0.02) [0.02]	−0.0580 (−0.59) [−0.58]	−0.05298 (−0.56) [−0.55]	0.00447 (0.05) [0.04]	0.0335 (0.34) [0.30]	0.0341 (0.34) [0.33]	−0.00407 (−0.04) [−0.04]	−0.0107 (−0.13) [−0.13]
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.415	0.397	0.568	0.567	0.573	0.725	0.594	0.652	0.677

Notes: t statistics in parentheses, heteroskedasticity robust standard errors. Bootstrapped z statistics in squared brackets, based on 1000 replications; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table 6.[23] The estimated effect of the performance decline is statistically significant in all specifications but, as a lower bound, the estimates are somewhat smaller than the corresponding estimates in Table 6. The effect of the starting performance is statistically significant in all but one specification and the estimated coefficients are larger than those in Table 6. Because the association between the starting performance and economic growth is more easily distinguished if the starting performance is uncorrelated with the performance decline, the estimates in Table 7 can be interpreted as an upper bound of the correlation between cognitive skills and economic growth.

Our second restrictive measure of the performance decline concerns non-reached questions. In the PISA 2009, the average number of non-reached questions per student is 1.83.[24] However, this number differs per country with a standard deviation of 1.65. It is unclear what factors are driving non-reached questions. They could be a consequence of cognitive and/or noncognitive skills. To take all questions into account, we coded non-reached questions in our main results as incorrectly answered. Alternatively, we can code them as missing, which has the effect of making the performance decline (α_1) weaker for all countries, whereas the starting performance (α_0) is hardly affected. The standard deviation for the performance decline also decreases by 57 percent, which makes it more difficult to find potential effects. Moreover, as noncognitive skills are potentially the cause of non-reached questions, this can be considered as another lower bound for the effect of non-cognitive skills.

Recognizing this, we show results for this second stricter measure in Table 8. The estimated effects of the starting performance are similar in magnitude or somewhat larger than the effects shown in Table 6, consistent with the finding that this component is unaffected by considering non-reached questions as missing. The estimate for the performance decline is reduced in magnitude, but remains statistically significant in most specifications. The estimated coefficient of the performance decline is small and insignificant in the specification that controls for regional indicators (column (6)), which is consistent with the idea that cultural differences are important determinants of non-cognitive skills. Despite the notable reduction in the estimates observed after excluding non-reached questions, this alternative measure does confirm that the relationship between test scores and economic growth is partially mediated by noncognitive skills. Table A.4 in the Appendix shows this conclusion does not change when combining the two restricted measures (orthogonal correction and non-reached questions as missing).

The Appendix reports results for several alternative measures of the starting performance and the decline in performance. Table A.5 shows the results when the two components are obtained through OLS instead of using a probit. As suggested by the correlations of 0.96 and higher of the two components when estimated with the two different methods, the results are essentially unchanged compared to Section 6.2. Subsequently, in Table A.6, we use the probability of answering the first question correctly as the starting performance and the probability of answering the last minus the probability of answering the first question correctly to measure the performance decline. Despite being a non-linear function of the original probit-estimates α_0 and α_1, the results are robust to using these probabilities. Table A.7 reports regression results using a measure of the starting performance that incorporates performance on the first five questions, which might be considered a more

[23] The reported t-statistics are based on robust standard errors, but the results do not qualitatively change if the standard errors are bootstrapped. Bootstrapped z-statistics are shown in square brackets. We used the bootstrap procedure for the two-step estimator as described in Cameron and Trivedi (2005). Bootstrapping is more relevant for the analysis in this section than for the analysis in Section 6.2 because we can only use 73 observations in the first step of the estimation, making the argument for consistency less plausible.

[24] PISA distinguishes between non-reached and skipped test items. Non-reached questions are defined as all consecutive unanswered questions clustered at the end of test, except for the first missing answer.

Table 8
Regressions of economic growth on components of test scores coding non-reached questions as missing.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.832***		1.528***	1.526***	1.328***	0.733**	0.790	0.0344	1.912***
performance	(3.84)		(3.53)	(3.48)	(3.11)	(2.11)	(1.26)	(0.05)	(6.00)
Performance		1.034***	0.616**	0.618**	0.548*	0.164	0.484	0.661*	0.306
decline		(2.93)	(2.63)	(2.62)	(1.72)	(0.91)	(1.67)	(1.97)	(1.58)
Years of	−0.0193	0.0308	−0.0244	−0.1792	−0.000662	0.0427	0.0760	0.0295	0.0135
schooling	(−0.17)	(0.30)	(−0.22)	(−0.17)	(-0.01)	(0.42)	(0.73)	(0.26)	(0.16)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.454	0.354	0.486	0.486	0.417	0.717	0.553	0.593	0.669

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

accurate measure of the performance at the beginning of the test.[25] Results do not change when using this alternative measure of starting performance.

We have also computed the decomposition on the cluster level as explained in Section 3 so that the unit of analysis matches the unit of randomization and the starting performance corresponds to the first cluster of the test. Whereas the estimates reported in Table A.8 are qualitatively similar to those discussed in Section 6.2, the estimate for the starting performance (performance decline) is somewhat larger (smaller) and more (less) significant. This pattern of results can be explained by the fact that this version of the starting performance already includes the first part of the performance decline. Next, in Table A.9, we report estimates from using the average of the components for PISA 2003, 2006 and 2009. As both the performance decline and the starting performance are being estimated, albeit precisely, this mitigates measurement error.[26] We find qualitatively identical estimates to those discussed in Section 6.2. For a similar reason, we repeat our main analysis weighting the observations by the inverse of the standard error of the performance decline in Table A.10, decreasing the weight put on observations for which the performance decline is measured imprecisely. The results are qualitatively unchanged. Repeating these weighted regressions coding non-reached questions as missing confirms that the lower-bound estimate for the effect of noncognitive skills is statistically significant (see Table A.11). In sum, we find that the relationship between the two components and economic growth is remarkably robust and we conclude that noncognitive skills are an important mediator in the relationship between test scores and economic growth.

7.2. From skills to growth or from growth to skills

We have applied the decomposition strategy to the PISA test, which is mostly a post period measure. Despite the starting performance and the performance decline being stable over time and the PISA 2006 being a good proxy for the HW-index, the use of a post period measure raises obvious concerns about reverse causality. In other words, our analyses may be capturing the effect of growth on the accumulation of skills. To address this issue, we test for the presence of a reversed channel from growth to skills and apply the decomposition method to an international test administered in 1991.

Most importantly, growth might provide a country with resources that are invested in human capital:

$$R_c = \eta_0 + \eta_1 G_c + \sum_n \kappa_n X_{nc} + u_c \tag{4}$$

$$\begin{pmatrix} S_c \\ PD_c \end{pmatrix} = \begin{pmatrix} \pi_0^S \\ \pi_0^{PD} \end{pmatrix} + \begin{pmatrix} \pi_1^S \\ \pi_1^{PD} \end{pmatrix} R_c + \begin{pmatrix} V_c^S & 0 \\ 0 & V_c^{PD} \end{pmatrix} \begin{pmatrix} \gamma^S \\ \gamma^{PD} \end{pmatrix} + \begin{pmatrix} v_c \\ \xi_c \end{pmatrix} \tag{5}$$

Where R_c are the (educational) resources in country c, and X_c and V_c are vectors containing control variables, where the latter could potentially be different for the starting performance versus the performance decline. Consequently, the estimate of β_1 and β_2 in Eq. (3) could also reflect the effect of economic growth on the starting performance and the performance decline through its effect on resources. In particular, if we assume that Eqs. (3) to (5) only include a constant and, respectively, the starting performance, economic growth and the (educational) resources as explanatory variables, the estimate for the starting performance in Eq. (3) equals:

$$\widehat{\beta_1} = \beta_1 + \frac{\frac{\pi_1^s \eta_1}{1 - \pi_1^s \eta_1 \beta_1} var[\epsilon_c]}{var[S_c]}$$

However, this also shows that if π_1^s is equal to zero, it strongly reduces the concerns for reverse causality.[27] Previous studies failed to find consistent, strong evidence that test performance is affected by real classroom resources, financial aggregates, and other facilities such as availability of a laboratory or the size of the library (Hanushek, 2002; Hanushek & Kimko, 2000; Hanushek & Woessmann, 2011a; Lee & Barro, 2001; Woessmann, 2003). We revisit this issue by regressing the starting performance and the performance decline in the PISA 2009 on educational expenditures, which we collected from the World Development Indicators. Within our framework (i.e., if growth affects the two PISA components only through resources), consistent estimates are obtained if u is uncorrelated with v and ξ. Table 9 reports the results of regressing the starting performance and the performance decline on two measures of educational expenditures, average government expenditure on education as percentage of GDP for the period 1970–2009, and the average pupil-to-teacher ratio in primary school for the same period.[28] We control for economic development by including an OECD indicator. Although there is little evidence of an association between educational expenditures and the two components of the PISA-test, columns (3) and (6) show that the pupil-to-teacher ratio is negatively related to the starting performance (p-value = 0.059). In column (6), however, the two measures are jointly insignificant. Moreover, these estimates are likely

[25] In particular, we set Q_{ij} of Eq. (2) equal to zero for any item j that was ordered in any of the first five positions in the test.

[26] We include a country in this regression if it participated in at least one of three PISA waves. Repeating this analysis with countries that participated in all three PISA waves restricts us to 31 countries, but does not change our results.

[27] As π_1^s, η_1, and β_1 are expected to be non negative, for a shock to die out the term $\pi_1^s \eta_1 \beta_1$ must be less than 1. Note that one can obtain a similar expression for the bias of the performance decline, where Eq. (3) is only a function of the performance decline.

[28] We experimented with other educational expenditure measures, such as percentage of qualified teachers in primary education, government expenditure per primary student, secondary student, tertiary student, and expenditure on education as a percentage of total government expenditure. These measures give us the same results, but are available for less countries and a shorter time period (mostly from 1998 onwards).

Table 9
Regressions of the two PISA components on measures for educational expenditures.

	(1) Starting performance	(2) Performance decline	(3) Starting performance	(4)[a] Starting performance	(5) Performance decline	(6) Starting performance	(7) Performance decline
Gov. exp. % of GDP	0.00214 (0.06)	0.0129 (0.78)				−0.0289 (−0.84)	0.0102 (0.54)
Pupil-to-teacher ratio			−0.0117* (−1.93)	0.00512 (0.59)	−0.00526 (−1.40)	−0.0118* (−1.94)	−0.00293 (−0.74)
N	59	59	60	60	60	53	53
Adj. R^2	0.480	0.227	0.438	0.327	0.204	0.460	0.204
F-test						0.1249	0.5633

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Regressions include a constant and an OECD indicator.

[a] Controlled for regional indicators instead of an OECD indicator; We include the country in the regression if it has more than 25% nonmissing observations of the educational expenditure measure within the time period 1970–2009. As on average 75% of the variation of these two measures lies between countries (and 25% over time), we feel this criteria is strict enough in order to give a good picture of the average educational expenditures of a country. Results are robust to different criteria, also to including the country if we have just one observation within the whole time period.

Table 10
Regressions of economic growth on the starting performance and performance decline using an early test (RLS 1991).

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting performance	1.961 (1.50)		2.428** (2.51)	2.347** (2.66)	2.377** (2.51)	1.693 (1.50)	0.303 (0.27)	0.00903 (0.01)	0.927 (0.61)
Performance decline		1.413* (1.93)	1.677*** (2.89)	1.750*** (2.92)	1.467** (2.49)	0.998 (1.68)	1.798** (2.77)	1.469** (2.85)	1.402*** (3.07)
Years of schooling	0.00812 (0.05)	0.140 (0.95)	0.0969 (0.67)	0.2209 (1.36)	0.143 (1.09)	0.190 (1.01)	0.254 (1.07)	0.330 (1.60)	-0.00523 (-0.05)
N	23	23	23	23	23	23	21	21	23
Adj. R^2	0.034	0.090	0.304	0.350	0.261	0.349	0.339	0.380	0.195

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1995–2010; Regressions include a constant and GDP per capita in 1990; [a–f] See the description of Table 4. We use average years of schooling over the period 1990–2010 and GDP per capita in 1990 in logs

to be upward biased, as favorable, but omitted, state policies tend to be positively correlated with resource-usage (Hanushek, 2002). Including regional indicators does not change our results and, in fact, column (4) shows the marginally significant estimate of the pupil-to-teacher ratio of column (3) is insignificant after controlling for regional indicators.

A more direct approach to address the possibility of reverse causality is to apply the decomposition method to an "early" international test. The main advantage of this approach is being able to explore the effects of noncognitive skills among workers in the labor force on economic growth. However, potential problems of early international tests are low country participation, bad documentation and the absence of (exogenous) variation in the ordering of the questions. Despite these potential problems, we apply the decomposition method to the Reading Literacy Study (RLS), a test administered in 1991, which is also included in the HW-index. This is the first test with relatively high country participation and we were able to retrieve the order of the questions in the test. All 14-year-old pupils in the RLS were administered the same booklet, so we are unable to separate the performance decline from the difficulty of the question. To the contrary, Elley (1992) notes that the end of the RLS contains the longer and more difficult reading passages, which suggests the performance decline is contaminated by cognitive skills.

Recognizing these potential problems, we use Eq. (2) without question fixed effects to decompose the RLS into the starting performance and performance decline.[29] Then, we use the average economic growth in the period 1995–2010 as our outcome in the post period.[30] Table 10 reports our results, restricting the sample to countries examined by HW (2012). Consistent with the results discussed above, we find that both components are positively related to economic growth and the estimates are statistically significant. The size of the estimates must be interpreted with care, as the performance decline is identified without variation in the order of the questions and could, therefore, be influenced by cognitive skills. This, in fact, could explain why we find somewhat larger estimates for the performance decline in Table 10 as compared to the results in Tables 5 and 6. Nevertheless, the overall results are very similar to those discussed in Section 6.2. Again, the performance decline seems to be more resilient to the inclusion of controls related to the quality of economic institutions than the starting performance, and including regional indicators reduces its coefficient.

Elley (1992) notes that the more difficult questions are concentrated at the end of the test. Therefore, we also compute the performance decline while excluding the last (two blocks of) questions from the RLS as an extra robustness check, so that the performance decline is estimated based on a more homogeneous sample of questions.[31] Table A.12 in the Appendix shows that the results are insensitive to this change.

[29] The 14-year-old pupils make two separate booklets for the RLS of 40 and 49 multiple-choice questions, we take the average of the two components to reduce measurement error. We are unsure about the length of the break in between the two tests, but results are very similar if we use the components of both booklets separately. Similarly to the previous analysis using PISA, we use the students weights provided by the RLS-dataset to ensure a representative sample. We do not use the data for the test of the 9-year-old pupils in the RLS, as this would again introduce concerns for reverse causality.

[30] We have also considered different periods of economic growth, for example starting at 1990 or ending at 2007 to avoid an influence of the financial crisis, and results are robust. We control for initial GDP and years of schooling in 1990, as to coincide with the measurement of the starting performance and performance decline. Results are qualitatively similar controlling for the initial values in 1995, though the starting performance loses some significance.

[31] Elley (1992) does not contain specific information on when the more difficult questions are asked, so the choice of which questions to exclude is somewhat arbitrary. However, we exclude two blocks of questions that are centered around the same reading passage. These two blocks contain 9 and 13 questions, of in total 40 and 49 questions in test 1 and 2 respectively. Moreover, we did observe a somewhat sharper increase in the number of incorrectly answered questions at the start of these two blocks.

8. Conclusions

Previous studies have found a positive association between cognitive test scores and economic growth. Although this association is difficult to interpret because of the potential for reverse causality, omitted variables and measurement error, HW (2012) have found evidence consistent with a causal interpretation. The goal of the present study was to investigate whether the well-documented relationship between cognitive test scores and economic growth is, at least in part, driven by noncognitive skills. Specifically, we have applied a recently developed method for decomposing test scores into two components: the starting performance and the decline in performance during the test. Research by Borghans and Schils (2012), Balart and Oosterveen (2017) and Zamarro, Hitt et al. (2016), as well as our results reported in Section 7.1, suggest that the performance decline provides a measure of noncognitive skills that is not confounded by cognitive skills. Consequently, it allows us to analyze whether the relationship between test scores and economic growth is, at least in part, driven by noncognitive skills.

We find that both components are associated with economic growth. The estimated effect of the performance decline is approximately equal to the estimated effect of the starting performance. Moreover, we find that the effect of the starting performance is reduced by 40 percent after controlling for the decline in performance, implying that previous estimates of cognitive skills are biased upwards and that noncognitive skills are partly responsible for the relationship between test scores and economic growth. This result is consistent with those of other recent studies that have raised concerns about the size of the effects of cognitive skills on economic growth (Atherton, Appleton, & Bleaney, 2013; Breton, 2011; Levin, 2012). Our results are robust to using a variety of measures of the performance decline, to testing for the presence of a reverse channel from growth to skills, and to using a post period measurement of economic growth.

It would, of course, be ideal to use a more direct approach to test for causal effects (e.g. via the use of instrumental variables), but finding a foolproof instrument for cross-country growth regressions is extremely difficult. Table A.13 in the Appendix takes a first step towards this goal by exploring cultural measures as an instrument for the performance decline. Nunn (2012) and Guiso et al. (2006) argue that culture reflects customary beliefs and values inherited from previous generations and can therefore be seen as a source of exogenous variation. We exploit that Hofstede's measures are defined as stable and, specifically, that long term orientation is described as thrift and effort, something directly related to the noncognitive skills the performance decline is hypothesized to measure. When we estimate the first stage, long term orientation is strongly correlated with the decline in performance. The second-stage estimates confirm our main findings.

In this study, we have tried to stay as close as possible to the approach used in previous studies that have established a clear relationship between test scores and economic growth. It should be noted that we are not able to apply the decomposition method to the HW-index, used in the previous studies, but we have applied this method to the PISA test which is only one of the tests included in the HW-index. However, it is likely that the results are also relevant for the other tests that compose the HW-index. First, a large literature in psychology, dating back to test pioneers as Thorndike and Wechsler, and a more recent stream of studies in economics provide evidence for the importance of noncognitive skills for cognitive test scores. Second, we find a very high correlation between the HW-index and the PISA scores, and using PISA scores instead of the HW-index produces very similar results when models used by previous studies are re-estimated. Third, the components resulting from the PISA-decomposition are very stable between countries and over time. Fourth, applying the decomposition to the Reading Literacy Study 1991, an international test also included in the HW-index, gives similar results. Therefore, it seems not likely that the decomposition results found for the PISA test are relevant to this specific test only.

Given the different types of policy interventions required to foster cognitive and noncognitive skills (Cunha et al., 2010), it is important to have a good understanding of the consequences of each type of skill. This distinction has been largely studied at the microeconomic level. Our study provides a first attempt to explore the implications of distinguishing between cognitive and noncognitive skills at the macroeconomic level. Our findings imply that noncognitive skills are also important for explaining the relationship between test scores and economic growth.

Acknowledgments

We would like to thank Lex Borghans, Antonio Cabrales, Eric Hanushek, Sacha Kapoor and Trudie Schils for helpful comments and suggestions. Seminar participants at ASSET, CEMFI, the Economics of Education Association, ESPE, the Erasmus School of Economics, the Ministry of Education (the Netherlands), RIDGE, the Tinbergen Institute and the Trends and Challenges on Human Resources International Workshop are also gratefully acknowledged for their feedback. Balart thanks the Spanish Ministry of Economy and Competitiveness for its financial support through grant ECO2012-34581.

Appendix A

Cultural Dimensions of Hofstede

- **Power Distance**: this dimension expresses the degree to which the less powerful members of a society accept and expect that power is distributed unequally. The fundamental issue here is how a society handles inequalities among people. People in societies exhibiting a large degree of Power Distance accept a hierarchical order in which everybody has a place and which needs no further justification. In societies with low Power Distance, people strive to equalise the distribution of power and demand justification for inequalities of power.
- **Individualism**: the high side of this dimension, called individualism, can be defined as a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families. Its opposite, collectivism, represents a preference for a tightly-knit framework in society in which individuals can expect their relatives or members of a particular in-group to look after them in exchange for unquestioning loyalty. A society's position on this dimension is reflected in whether peoples self-image is defined in terms of "I" or "we".
- **Masculinity**: the Masculinity side of this dimension represents a preference in society for achievement, heroism, assertiveness and material rewards for success. Society at large is more competitive. Its opposite, femininity, stands for a preference for cooperation, modesty, caring for the weak and quality of life. Society at large is more consensus-oriented. In the business context Masculinity versus Femininity is sometimes also related to as "tough versus tender" cultures.
- **Uncertainty Avoidance**: the Uncertainty Avoidance dimension expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity. The fundamental issue here is how a society deals with the fact that the future can never be known: should we try to control the future or just let it happen? Countries exhibiting strong UAI maintain rigid codes of belief and behavior and are intolerant of unorthodox behavior and ideas. Weak UAI societies maintain a more relaxed attitude in which practice counts more than principles.
- **Long Term Orientation**: every society has to maintain some links with its own past while dealing with the challenges of the present and the

future. Societies prioritize these two existential goals differently. Societies who score low on this dimension, for example, prefer to maintain time-honoured traditions and norms while viewing societal change with suspicion. Those with a culture which scores high, on the other hand, take a more pragmatic approach: they encourage thrift and efforts in modern education as a way to prepare for the future. In the business context this dimension is related to as "(short term) normative versus (long term) pragmatic" (PRA). In the academic environment the terminology Monumentalism versus Flexhumility is sometimes also used.

- **Indulgence**: indulgence stands for a society that allows relatively free gratification of basic and natural human drives related to enjoying life and having fun. Restraint stands for a society that suppresses gratification of needs and regulates it by means of strict social norms.

Source: https://geert-hofstede.com/national-culture.html.
Retrieved: July 1st, 2016

Instrumental Variable Analysis. Despite the usual set of controls in the cross-country growth regressions, it is desirable to use a more direct approach to identify causal effects. To this end we use an instrumental variable approach. This small-sample analysis has to be interpreted with caution, IV estimators can have a finite-sample distribution that differ from the asymptotic distribution. As in Guiso et al. (2006), we use cultural measures as an instrument, in our case for the performance decline. They argue that culture reflects customary beliefs and values that are inherited by an individual from previous generations, rather than voluntarily accumulated. Because of the difficulty of changing culture, it is largely given to individuals throughout their lifetime and can therefore be seen as a source of exogenous variation. Similarly, Nunn (2012) conceptualizes culture through decision making heuristics or rules of thumb that have evolved given our need to make decisions in complex and uncertain environments. He argues these are typically slow-moving.

The growing body of research investigating the effects of culture upon economic growth raises questions regarding the validity of using culture as an instrument (Gorodnichenko & Roland, 2011; Gorodnichenko & Roland, 2016; Tabellini, 2010). For example, Gorodnichenko and Roland (2016) use genes as an instrument to document a direct effect of individualism on growth. To the best of our knowledge, however, the cultural measures used below have not been well studied (also not by Gorodnichenko & Roland (2011)). Moreover, Guiso et al. (2006), arguably, define cultural measures as potential instruments, whereas this IV-analysis might also reduce the risk of reverse causality as we call upon slow-moving variation in the performance decline.

Table A.13 shows our results, where the upper and lower panel display the first- and second stage respectively. First, we exploit the Weber-hypothesis which states that the emergence of the spirit of capitalism, accumulation of wealth, and virtues of hard work can be attributed to Protestant work ethic (Nunn, 2012). We use the share of Protestantism in 2000 from Barro (2003) as an instrument for the performance decline (column (1)). We find a positive first-stage relationship, but the F-statistic reveals we are dealing with a weak instrument. The second stage shows an IV-estimate that is close to OLS, but it is very imprecisely estimated which can be explained by the amount of noise introduced in the first stage. Moreover, problems related to finite-sample bias and potential endogeneity of the instrument are magnified if the instrument is weakly correlated with the endogenous variable.

Table A.1
Randomization test.

	(1) Gender	(2) Mother highest schooling	(3) Father highest schooling	(4) Self born in country	(5) Mother born in country	(6) Father born in country	(7) Language at home	(8) Possessions desk	(9) Possessions own room	(10) How many books at home	(11) Age of student
Booklet = 2	0.0219**	0.00937	0.0216	−0.00430	−0.0000963	−0.000727	0.0110	0.00126	−0.00566	−0.0385	−0.00141
	(2.31)	(0.36)	(0.82)	(−1.12)	(−0.02)	(−0.13)	(1.05)	(0.15)	(−0.71)	(−1.40)	(−0.24)
Booklet = 3	0.0138	0.000973	0.0320	−0.000545	0.00458	0.00398	0.00969	0.00383	−0.00966	−0.0178	−0.00400
	(1.46)	(0.04)	(1.21)	(−0.13)	(0.80)	(0.69)	(0.92)	(0.48)	(−1.22)	(−0.65)	(−0.68)
Booklet = 4	0.0105	0.0254	0.0361	−0.00620*	−0.00460	0.000133	−0.00145	−0.00271	−0.00862	−0.0246	0.00145
	(1.11)	(0.99)	(1.39)	(−1.65)	(−0.83)	(0.02)	(−0.14)	(−0.35)	(−1.11)	(−0.91)	(0.24)
Booklet = 5	0.00380	0.000718	0.0132	−0.000969	−0.000655	0.00415	0.00862	0.00109	−0.00358	0.00290	0.00161
	(0.40)	(0.03)	(0.50)	(−0.23)	(−0.12)	(0.72)	(0.82)	(0.14)	(−0.45)	(0.11)	(0.28)
Booklet = 6	0.0151	0.00317	0.0463*	−0.000802	0.00137	0.00195	0.00322	0.0149*	−0.00731	0.0127	0.00453
	(1.60)	(0.12)	(1.76)	(−0.20)	(0.24)	(0.34)	(0.31)	(1.83)	(−0.91)	(0.47)	(0.51)
Booklet = 7	0.0112	−0.0139	0.0187	−0.000645	0.00157	0.000849	0.00214	0.00432	−0.0101	−0.0497*	0.00296
	(1.19)	(−0.55)	(0.71)	(−0.16)	(0.28)	(0.15)	(0.21)	(0.53)	(−1.28)	(−1.84)	(0.50)
Booklet = 8	0.0166*	0.0213	0.0361	−0.00442	−0.00145	0.000277	0.00387	0.00361	−0.00711	−0.0603**	−0.00148
	(1.77)	(0.83)	(1.37)	(−1.15)	(−0.26)	(0.05)	(0.38)	(0.45)	(−0.88)	(−2.23)	(−0.25)
Booklet = 9	0.00706	0.0371	0.0233	−0.00329	−0.00369	−0.000517	0.00476	−0.00314	0.00158	−0.0493*	0.00333
	(0.75)	(1.43)	(0.88)	(−0.84)	(−0.66)	(−0.09)	(0.47)	(−0.40)	(0.19)	(−1.81)	(0.56)
Booklet = 10	−0.000460	0.00574	0.0170	0.0000214	−0.000589	0.00270	0.00300	−0.00194	−0.00600	−0.0384	0.000419
	(−0.05)	(0.23)	(0.65)	(0.01)	(−0.10)	(0.47)	(0.29)	(−0.25)	(−0.76)	(−1.41)	(0.07)
Booklet = 11	0.00842	−0.000262	0.0146	−0.00526	0.00428	0.00426	0.00507	0.00435	0.00357	0.00683	−0.00255
	(0.90)	(−0.01)	(0.56)	(−1.36)	(0.73)	(0.73)	(0.49)	(0.54)	(0.44)	(0.25)	(−0.43)
Booklet = 12	0.0118	0.0253	0.00712	−0.00374	0.00228	0.00254	0.00654	−0.000107	−0.00620	−0.0125	0.00600
	(1.26)	(0.99)	(0.28)	(−0.95)	(0.40)	(0.44)	(0.63)	(−0.01)	(−0.78)	(−0.48)	(1.02)
Booklet = 13	0.0120	0.0248	0.0273	−0.00200	0.000499	0.00227	0.000449	0.00373	0.00191	−0.0213	−0.00285
	(1.28)	(0.96)	(1.04)	(−0.50)	(0.09)	(0.40)	(0.04)	(0.46)	(0.24)	(−0.78)	(−0.49)
Constant	1.485***	2.110***	2.074***	1.044***	1.092***	1.089***	1.177***	1.164***	1.240***	2.962***	15.78***
	(223.51)	(119.07)	(111.35)	(357.29)	(271.57)	(264.96)	(160.00)	(207.22)	(218.42)	(155.55)	(3713.17)
Observations	397,916	378,276	367,202	390,715	389,346	386,517	383,775	390,488	391,047	390,014	397,920
F-value	0.92	0.62	0.49	0.66	0.45	0.20	0.27	0.66	0.67	1.52	0.49
P-value	0.5267	0.8268	0.9234	0.7875	0.9448	0.9985	0.9941	0.7943	0.7816	0.1074	0.9200
Adjusted R^2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Regressions of background characteristics upon separate indicators for every booklet. The columns '*F*-value' and '*P*-value' refer to the tests for joint significance of the booklet indicators. PISA 2006 and PISA weights are used. *t* statistics in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2
The starting performance and decline in performance per country.

Country	(1) PISA score	(2) P[$Q_0 = 1$]	(3) P[$Q_1 = 1$]	(4) Decline	Country	(1) PISA score	(2) P[$Q_0 = 1$]	(3) P[$Q_1 = 1$]	(4) Decline
Colombia	381	.59	.243	.347	Lithuania	481.3	.737	.64	.097
Uruguay	422.7	.722	.423	.299	United States	481.5	.776	.679	.097
Argentina	382	.636	.359	.277	Luxembourg	485	.807	.71	.097
Tunisia	377	.454	.229	.225	Poland	500.3	.823	.726	.097
Brazil	384.3	.547	.331	.216	China, Macao	509.3	.828	.732	.096
Kyrgyzstan	306	.393	.185	.208	Hungary	492.3	.789	.694	.095
Mexico	408.7	.594	.387	.207	Slovakia	482	.816	.725	.091
Chile	430.3	.714	.508	.207	Sweden	504	.842	.752	.09
Qatar	326.3	.525	.337	.188	Japan	517.3	.877	.791	.086
Israel	445	.686	.506	.18	Azerbaijan	403.7	.584	.499	.085
Russia	465	.832	.654	.177	Canada	529.3	.832	.748	.084
Greece	464	.786	.609	.177	Ireland	508.7	.753	.67	.083
Jordan	402.3	.51	.339	.171	Australia	520	.836	.754	.082
Romania	409.7	.6	.432	.168	Belgium	510.3	.836	.755	.081
Thailand	418.3	.533	.368	.165	Denmark	501	.866	.787	.079
Bulgaria	416.3	.667	.502	.165	Taiwan	525.7	.833	.754	.078
Indonesia	392.3	.574	.411	.163	Czech Republic	502	.845	.766	.078
Italy	468.7	.735	.582	.153	New Zealand	524.3	.814	.737	.077
Turkey	431.7	.537	.388	.149	Slovenia	505.7	.819	.742	.077
Serbia	424	.709	.578	.131	Germany	505	.826	.753	.073
Latvia	485	.761	.638	.123	Estonia	515.7	.829	.756	.072
Portugal	470.7	.779	.658	.121	Netherlands	521	.83	.76	.07
Spain	476.3	.803	.682	.121	Hong Kong	541.7	.817	.746	.07
Montenegro	401	.583	.467	.117	Korea	541.7	.823	.755	.067
France	493	.807	.693	.114	Switzerland	513.7	.85	.789	.061
United Kingdom	501.7	.762	.652	.11	Liechtenstein	519	.885	.828	.057
Norway	487	.83	.721	.109	Austria	502	.843	.789	.054
Iceland	493.7	.85	.75	.1	Finland	552.7	.898	.856	.042
Croatia	479	.759	.66	.099					

Notes: Probabilities are based on the estimates from Eq. (2), using PISA 2006 and PISA weights.

Table A.3
Descriptive statistics.

Country	Initial GDP (1960)	GDP growth (1960–2000)	HW-index	Starting performance (st. error)	Performance decline (st. error)	Num. of students
Argentina	6.033	1.258	3.920	.3478 (.0427)	−.710 (.0118)	4339
Australia	15.20	2.061	5.093	.9789 (.0250)	−.292 (.0059)	14170
Austria	10.54	3.173	5.089	1.007 (.0418)	−.204 (.0094)	4908
Belgium	10.16	2.975	5.041	.9766 (.0310)	−.287 (.0072)	8685
Brazil	2.469	2.709	3.637	.1183 (.0338)	−.555 (.0096)	9295
Canada	12.90	2.382	5.037	.9610 (.0284)	−.292 (.0068)	22646
Chile	3.700	2.689	4.049	.5664 (.0358)	−.546 (.0095)	5233
Colombia	2.940	1.758	4.152	.2271 (.0459)	−.924 (.0127)	4478
Denmark	11.60	2.757	4.962	1.107 (.0421)	−.310 (.0096)	4532
Finland	9.034	3.149	5.126	1.271 (.0442)	−.207 (.0094)	4714
France	10.19	2.815	5.040	.8680 (.0369)	−.362 (.0090)	4716
Greece	5.588	3.428	4.607	.7919 (.0371)	−.515 (.0093)	4873
Hong Kong	3.289	5.633	5.194	.9025 (.0413)	−.239 (.0097)	4645
Iceland	14.07	2.584	4.935	1.038 (.0437)	−.363 (.0101)	3789
Indonesia	.6651	3.719	3.879	.1871 (.0343)	−.411 (.0101)	10647
Ireland	7.280	4.008	4.994	.6842 (.0363)	−.245 (.0093)	4585
Israel	6.989	3.133	4.686	.4848 (.0363)	−.470 (.0095)	4584
Italy	8.718	3.174	4.757	.6285 (.0249)	−.420 (.0064)	21773
Japan	5.594	4.521	5.310	1.160 (.0369)	−.349 (.0084)	5952
Jordan	2.721	.8659	4.263	.0257 (.0332)	−.441 (.0092)	6509
Korea	1.670	6.129	5.337	.9255 (.0365)	−.234 (.0090)	5176
Mexico	4.942	2.271	3.997	.2379 (.0278)	−.526 (.0076)	30971
Netherlands	13.43	2.606	5.114	.9557 (.0429)	−.249 (.0099)	4769
New Zealand	14.26	1.661	4.978	.8943 (.0388)	−.259 (.0093)	4823
Norway	12.50	3.286	4.830	.9542 (.0386)	−.369 (.0094)	4692
Portugal	4.181	4.134	4.563	.7694 (.0376)	−.362 (.0095)	5109
Romania	1.362	3.904	4.562	.2532 (.0478)	−.425 (.0127)	5118
Spain	6.333	3.809	4.829	.8522 (.0308)	−.379 (.0075)	19604
Sweden	14.31	1.912	5.013	1.001 (.0407)	−.320 (.0104)	4443
Switzerland	21.02	1.494	5.141	1.035 (.0339)	−.234 (.0077)	12192
Taiwan	1.858	6.459	5.451	.9650 (.0320)	−.276 (.0074)	8815
Thailand	.9620	4.713	4.564	.0835 (.0336)	−.420 (.0094)	6192
Tunisia	1.805	2.945	3.795	−.115 (.0373)	−.627 (.0105)	4640
Turkey	3.183	2.285	4.127	.0923 (.0370)	−.376 (.0103)	4942
United Kingdom	11.20	2.558	4.949	.7134 (.0304)	−.322 (.0077)	13152
United States	15.38	2.373	4.902	.7598 (.0364)	−.294 (.0098)	5611
Uruguay	5.010	1.562	4.300	.5873 (.0409)	−.782 (.0104)	4839

Notes: Descriptive statistics for the sample used in Table 5. GDP per capita in 1960 PPP adjusted (in 2005 international Dollars), shown in thousands. The PISA-components are related to the wave of 2006.

Table A.4
Regressions of economic growth on components of test scores coding non-reached questions as missing and using an orthogonal measure of the performance decline.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.832***		1.778***	1.778***	1.551***	0.799**	0.987	0.303	2.036***
performance	(3.84)		(4.06)	(3.98)	(3.90)	(2.35)	(1.48)	(0.43)	(6.14)
			[4.28]	[4.17]	[2.33]	[1.87]	[1.49]	[0.41]	[5.51]
Performance		0.656**	0.565**	0.567**	0.503*	0.150	0.444	0.606*	0.280
decline		(2.19)	(2.63)	(2.62)	(1.72)	(0.91)	(1.67)	(1.97)	(1.58)
		[2.53]	[2.55]	[2.53]	[2.00]	[0.75]	[1.69]	[1.89]	[1.56]
Years of	−0.0193	0.0534	−0.0244	−.01792	−0.000662	0.0427	0.0760	0.0295	0.0135
schooling	(−0.17)	(0.52)	(−0.22)	(−0.17)	(−0.01)	(0.42)	(0.73)	(0.26)	(0.16)
		[0.50]	[−0.22]	[−0.17]	[−0.00]	[0.38]	[0.73]	[0.26]	[0.16]
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.454	0.274	0.486	0.486	0.417	0.717	0.553	0.593	0.669

Notes: t statistics in parentheses, heteroskedasticity robust standard errors. Bootstrapped z statistics in squared brackets, based on 1000 replications; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.5
Regressions of economic growth on components of test scores, where the two components are estimated via OLS instead of probit.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.697***		1.164***	1.169***	1.055***	0.709**	0.763	0.109	1.735***
performance	(3.45)		(3.35)	(3.28)	(2.87)	(2.10)	(1.43)	(0.20)	(5.62)
Performance		1.466***	1.158***	1.159***	1.168***	0.392	0.983**	1.123**	0.538**
decline		(4.15)	(3.96)	(3.94)	(4.20)	(1.35)	(2.14)	(2.42)	(2.16)
Years of	−0.0145	−0.0187	−0.0543	−0.0464	−0.00789	0.0350	0.0422	0.00136	−0.00891
schooling	(−0.13)	(−0.19)	(−0.54)	(−0.48)	(−0.09)	(0.35)	(0.42)	(0.01)	(−0.11)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.424	0.485	0.562	0.561	0.567	0.725	0.595	0.650	0.677

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.6
Regressions of economic growth on components of test scores using the probabilities related to the probit estimates.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.600***		0.892**	0.895**	0.779**	0.637*	0.595	0.00406	1.534***
performance	(3.20)		(2.62)	(2.58)	(2.13)	(1.76)	(1.21)	(0.01)	(5.20)
Performance		1.681***	1.380***	1.384***	1.387***	0.440	1.156**	1.300**	0.731***
decline		(4.67)	(4.42)	(4.41)	(4.67)	(1.32)	(2.15)	(2.47)	(2.78)
Years of	−0.0116	−0.0179	−0.0446	−0.04028	0.00493	0.0353	0.0474	0.0209	−0.0107
schooling	(−0.10)	(−0.19)	(−0.45)	(−0.42)	(0.06)	(0.35)	(0.48)	(0.20)	(−0.13)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.404	0.531	0.570	0.570	0.581	0.724	0.591	0.645	0.681

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.7
Regressions of economic growth on components of test scores including the first five questions of the test in the starting performance.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.718***		0.957**	0.965**	0.751*	0.586	0.644	−0.000494	1.594***
performance	(3.51)		(2.59)	(2.60)	(1.98)	(1.60)	(1.27)	(−0.00)	(4.98)
Performance		1.626***	1.285***	1.287***	1.279***	0.529	1.093**	1.277**	0.636**
decline		(4.72)	(4.26)	(4.25)	(4.36)	(1.67)	(2.13)	(2.60)	(2.56)
Years of	−0.00727	−0.0412	−0.0582	−0.0529	0.00388	0.0329	0.0331	−0.00422	−0.0112
schooling	(−0.06)	(−0.44)	(−0.60)	(−0.56)	(0.05)	(0.33)	(0.33)	(−0.04)	(−0.14)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.425	0.527	0.570	0.569	0.576	0.725	0.595	0.653	0.678

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.8
Regressions of economic growth on components of test scores, where the two components are estimated using the average score within a cluster as outcome variable and the position of the cluster in the test as explanatory variable.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.960***		1.372***	1.383***	1.313***	1.023***	1.239**	1.066	1.701***
performance	(4.87)		(3.74)	(3.67)	(3.45)	(2.95)	(2.26)	(1.59)	(5.70)
Performance		1.497***	0.865***	0.869***	1.001***	0.301	0.676*	0.763*	0.326
decline		(4.24)	(3.26)	(3.29)	(3.19)	(1.08)	(1.79)	(1.99)	(1.25)
Years of	−0.0607	−0.0282	−0.0760	−0.0734	−0.0204	0.00170	0.0107	0.00208	−0.0395
schooling	(−0.58)	(−0.29)	(−0.76)	(−0.77)	(−0.24)	(0.02)	(0.11)	(0.02)	(−0.48)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.535	0.491	0.590	0.589	0.626	0.765	0.649	0.687	0.680

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.9
Regressions of economic growth on the average of both components of the test score, using PISA 2003, 2006 and 2009.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.747***		1.056***	1.068***	0.886**	0.591	0.792	0.0523	1.618***
performance	(3.60)		(2.86)	(2.84)	(2.43)	(1.64)	(1.47)	(0.09)	(4.97)
Performance		1.649***	1.306***	1.315***	1.279***	0.655**	1.118**	1.353**	0.685***
decline		(4.67)	(4.30)	(4.30)	(4.48)	(2.06)	(2.10)	(2.58)	(2.73)
Years of	−0.0324	−0.0375	−0.0739	−0.07441	−0.0225	0.0236	0.0126	−0.0157	−0.0377
schooling	(−0.29)	(−0.41)	(−0.77)	(−0.79)	(−0.27)	(0.24)	(0.12)	(−0.14)	(−0.47)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.432	0.528	0.585	0.584	0.589	0.729	0.599	0.660	0.688

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.10
Regressions of economic growth on components of test scores using the standard error of performance decline as weights.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.815***		0.908**	0.939**	0.743*	0.470	0.598	0.133	1.633***
performance	(3.63)		(2.08)	(2.20)	(1.98)	(1.35)	(1.22)	(0.29)	(4.55)
Performance		1.764***	1.394***	1.385***	1.287***	0.608	1.367**	1.420**	0.731***
decline		(4.84)	(4.05)	(4.03)	(4.45)	(1.58)	(2.44)	(2.53)	(2.72)
Years of	−0.0391	−0.102	−0.101	−0.0897	0.00447	−0.0358	0.0280	0.00234	−0.0632
schooling	(−0.36)	(−1.03)	(−0.98)	(−0.85)	(0.05)	(−0.34)	(0.25)	(0.02)	(−0.81)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.487	0.578	0.604	0.600	0.573	0.744	0.639	0.670	0.715

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.11
Regressions of economic growth on components of test scores coding non-reached questions as missing and using the standard error of performance decline as weights.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	1.942***		1.549***	1.561***	1.328***	0.654*	0.916	0.313	1.998***
performance	(3.94)		(3.46)	(3.47)	(3.11)	(1.74)	(1.39)	(0.48)	(6.00)
Performance		1.095***	0.646**	0.646**	0.548*	0.214	0.646**	0.741*	0.329
decline		(3.01)	(2.58)	(2.57)	(1.72)	(0.88)	(2.10)	(1.98)	(1.51)
Years of	−0.0533	−0.0383	−0.0655	−0.0546	−0.000662	−0.0298	0.0852	0.0566	−0.0434
schooling	(−0.49)	(−0.38)	(−0.61)	(−0.51)	(−0.01)	(−0.29)	(0.80)	(0.50)	(−0.55)
N	55	55	55	55	55	55	47	47	55
Adj. R^2	0.518	0.449	0.548	0.546	0.417	0.738	0.592	0.611	0.711

Notes: t statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970; [a–f] See the description of Table 4. We use average years of schooling over the period 1970–2010 and GDP per capita in 1970 in logs

Table A.12
Regressions of economic growth on the starting performance and performance decline using an early test (RLS 1991) and excluding the last two blocks of questions.

	(1)	(2)	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[d]	(8)[e]	(9)[f]
Starting	2.133		1.936*	1.913*	0.707	0.973	0.579	0.0590	0.237
performance	(1.63)		(1.85)	(1.84)	(0.78)	(0.89)	(0.64)	(0.06)	(0.18)
Performance		1.781**	1.697**	1.680**	0.910*	1.185*	1.727**	1.412**	1.490**
decline		(2.68)	(2.63)	(2.51)	(1.96)	(2.12)	(2.45)	(2.34)	(2.67)
Years of	0.00556	0.164	0.113	0.120	0.144	0.222	0.332	0.391*	0.0105
schooling	(0.03)	(1.29)	(0.90)	(1.35)	(1.40)	(1.28)	(1.60)	(2.12)	(0.09)
N	23	23	23	23	22	23	21	21	23
Adj. R^2	0.061	0.232	0.359	0.388	0.154	0.437	0.378	0.399	0.262

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1995–2010; Regressions include a constant and GDP per capita in 1990; [a–f] See the description of Table 4. We use average years of schooling over the period 1990–2010 and GDP per capita in 1990 in logs

Table A.13
Growth regressions using instrumental variables for the performance decline and starting performance.

	(1)[a]	(2)[a]	(3)[a]	(4)[a]	(5)[b,c]	(6)[b,c]	(7)[b,d]	(8)[b,d]
				First stage				
Protestant	0.342							
share in 2000	(1.52)							
Long-term		0.0132***	0.00798**	0.0114***				
orientation		(5.19)	(2.72)	(4.15)				
Private					0.0046**	0.0022		
enrollment share					(2.58)	(1.38)		
Catholic							0.766	0.358
share in 1900							(1.31)	(0.73)
Years of					0.0433	0.0495	0.035	0.0624
schooling					(1.10)	(1.65)	(0.82)	(1.34)
Additional controls	No	No	Yes	No	No	Yes	No	Yes
F-test	2.31	26.94	7.40	17.22	3.74	3.05	1.12	1.12
				Second stage				
Performance	1.095	2.604***	2.998***	2.375***				
decline	(0.79)	(4.39)	(3.17)	(3.67)				
Starting				0.457	2.472*	3.474*	1.50	1.771
performance				(0.94)	(1.93)	(1.95)	(0.83)	(1.16)
Years of	0.006	−0.104	−0.081	−0.108				
schooling	(0.07)	(−1.01)	(−0.65)	(−1.08)				
N	53	51	45	51	21	21	42	41
Adj. R^2	0.495	0.415	0.504	0.461	0.395	0.152	0.617	0.666

Notes: *t* and *z* statistics in parentheses, heteroskedasticity robust standard errors; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$; Dependent variable: average annual growth rate in GDP per capita, 1970–2010; Regressions include a constant and GDP per capita in 1970. Additional controls include: openness of the economy and protection against expropriation (column (6) and (8)), plus fertility and tropical location (column (3));

[a] Dependent variable in first stage is the performance decline

[b] Dependent variable in first stage is the starting performance. We treat initial years of schooling (in 1970) as an extra instrumental variable

[c] Sample restricted to OECD countries

[d] Controlled for the share of Catholics in 1970 in both the first and second-stage

Next we exploit Hofstede's long term orientation cultural component as an alternative instrument for the performance decline. Hofstede explains his components as follows: "These relative scores have been proven to be quite stable over time. The forces that cause cultures to shift tend to be global or continent-wide. This means that they affect many countries at the same time, so if their cultures shift, they shift together and their relative positions remain the same".[32] Using long term orientation as an instrument for the performance decline gives a strong first-stage relationship with an F-statistic of 26.94 (column (2)). The second-stage estimate for the performance decline is statistically significant, the OLS-estimate falls within the 95% confidence interval of the IV-estimate. Regarding the validity of long term orientation as an instrument, the exclusion restriction is violated if culture affects economic growth through other channels than the performance decline. In particular, Nunn (2012) argues historical shocks can have a persistent effect upon culture (only) if formal institutions change with it. Column (3) and (4) respectively show our results are robust to the full set of controls, including the quality of economic institutions, and to controlling for the potentially endogenous starting performance, which addresses concerns on this particular violation of the exclusion restriction. Moreover, long term orientation is described as thrift and effort which are directly related to the noncognitive skills the performance decline is hypothesized to measure.[33]

Ideally we would instrument both of the potentially endogenous components within one model. To this end, column (5) until (8) investigate whether two of the instruments used by HW (2012) can be used for the starting performance. For comparability we perform this analysis on the same

[32] https://geert-hofstede.com/national-culture.html, retrieved July 1st, 2016.

[33] For the significant IV-estimates (columns (2), (3) and (4)) we tested for exogeneity of the performance decline using the Durbin-Wu and Hausman test. Under the assumption that the instrument is valid, we reject exogeneity at the 5%-level for all three specifications. While IV has the property of being consistent, keep in mind we are working with a small sample.

sample as HW (2012) and also use initial years of schooling as an additional instrument.[34] The potential instruments are private enrollment share in 1985 and the catholic share in 1900, which use the idea of private competition being beneficial for student achievement. HW (2012) state one can plausibly assume this institutional feature to be exogenous.[35] For both instruments the first stage shows weak F-statistics, there is only a significant relationship in column (5) at the 5%-level. This reduces the interest in the second stage, which show positive estimates that are either borderline significant or insignificant. We will refrain from an analysis with both components instrumented within one model.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.econedurev.2017.12.004.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *The American Economic Review, 91*(5), 1369–1401.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Chapter 1 – Personality psychology and economics. In E. A. Hanushek, S. Machin, & L. Woessmann (Vol. Eds.), *Handbook of the economics of education. 4. Handbook of the economics of education* (pp. 1–181). Elsevier.

Aten, B., Heston, A., & Summers, R. (2009). *Penn world table version 7.1.* Center for International Comparisons of Production, Income, and Prices at the University of Pennsylvania.

Atherton, P., Appleton, S., & Bleaney, M. (2013). International school test scores and economic growth. *Bulletin of Economic Research, 65*(1), 82–90.

Balart, P., & Oosterveen, M. (2017). Wait and see: Gender differences in performance during cognitive tests. *JOLE Working Paper* 17679.

Barro, R. (2003). Religion adherence data, http://scholar.harvard.edu/barro/publications/religion-adherence-data.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The Quarterly Journal of Economics, 106*(2), 407–443.

Barro, R. J. (2001). Human capital and growth. *The American Economic Review, 91*(2), 12–17.

Barro, R. J., & Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics, 104*, 184–198.

Benet-Martínez, V., & Oishi, S. (2008). *Culture and personality.* New York: Guilford.

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., Golsteyn, B. H., Heckman, J., & Humphries, J. E. (2011). Identification problems in personality psychology. *Personality and Individual Differences, 51*(3), 315–320.

Borghans, L., Meijers, H., & Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry, 46*(1), 2–12.

Borghans, L., & Schils, T. (2012). The leaning tower of pisa: Decomposing achievement test scores into cognitive and noncognitive components. JOLE Working Paper 13260.

Bosworth, B., & Collins, S. M. (2003). The empirics of growth: An update. *Brookings Papers on Economic Activity, 2003*(2), 113–206.

Breton, T. R. (2011). The quality vs. the quantity of schooling: What drives economic growth? *Economics of Education Review, 30*(4), 765–773.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications.* Cambridge University Press.

Caplan, P. J., Crawford, M., Hyde, J. S., & Richardson, J. T. (1997). *Gender differences in human cognition. Counterpoints: Cognition, memory, and language series.* ERIC.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action. 35*. Elsevier.

Cohen, D., & Soto, M. (2007). Growth and human capital: Good data, good results. *Journal of Economic Growth, 12*(1), 51–76.

Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human Resources, 48*(1), 236–264.

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653–665.

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

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica, 78*(3), 883–931.

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

Dohmen, T., Enke, B., Falk, A., Huffman, D., & Sunde, U. (2016). *Patience and the wealth of nations.* Mimeo.

Doménech, R., & De la Fuente, A. (2006). Human capital in growth regressions: How much difference does data quality make? *Journal of the European Economic Association, 4*(1), 1–36.

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.

Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*(1), 198.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251.

Elley, W. B. (1992). How in the world do students read? IEA study of reading literacy. *International Association for the Evaluation of Educational Achievement*.

Fernandez, R., & Fogli, A. (2009). Culture: An empirical investigation of beliefs, work, and fertility. *American Economic Journal: Macroeconomics, 1*(1), 146–177.

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*(2), 210–240.

Gallup, J. L., Sachs, J. D., & Mellinger, A. D. (1999). Geography and economic development. *International Regional Science Review, 22*(2), 179–232.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics,* 791–810.

Gorodnichenko, Y., & Roland, G. (2011). Which dimensions of culture matter for long-run growth? *The American Economic Review, 101*(3), 492–498.

Gorodnichenko, Y., & Roland, G. (2016). Culture, institutions and the wealth of nations. *Review of Economics and Statistics, 99*(3), 402–416.

Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives, 20*(2), 23–48.

Hanushek, E. A. (2002). Publicly provided education. *Handbook of Public Economics, 4*, 2045–2141.

Hanushek, E. A. (2013). Economic growth in developing countries: The role of human capital. *Economics of Education Review, 37*, 204–212.

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *The American Economic Review,* 1184–1208.

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature,* 607–668.

Hanushek, E. A., & Woessmann, L. (2011). Chapter 2 - The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Woessmann (Vol. Eds.), *Handbook of the economics of education. 3. Handbook of the economics of education* (pp. 89–200). Elsevier.

Hanushek, E. A., & Woessmann, L. (2011). How much do educational outcomes matter in oecd countries? *Economic Policy, 26*(67), 427–491.

Hanushek, E. A., & Woessmann, L. (Woessmann, 2011c). Sample selectivity and the validity of international student achievement tests in economic research. *Economics Letters, 110*(2), 79–82.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*(4), 267–321.

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *The American Economic Review, 103*(6), 2052–2086.

Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411–482.

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic Inquiry, 46*(3), 289–324.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19*(4), 451–464.

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

Hernández, M., & Hershaff, J. (2015). Skipping questions in school exams: The role of socio-emotional skills on educational outcomes. *Mimeo*.

Hitt, C. (2016). Just filling in the bubbles: Using careless answers patterns on surveys as a proxy measure of non cognitive skills. *EDRE Working Paper* 2015-06.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review, 52*, 105–119.

Hofstede, G., & McCrae, R. R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-Cultural Research: The Journal of Comparative Social Science*.

[34] As we have more instruments than endogenous variables, we can use the Sargan test to test whether the moment conditions are valid. For columns (5) until (8) we cannot reject the null-hypothesis of exogeneity for the instruments at conventional significance levels.

[35] See HW (2012) for details, but in particular they argue that many educational institutions are slow-moving and reflect long-standing policies that are not the outcome of economic growth. The data for the private enrollment share refers to the private enrollment as a percentage of total enrollment in general secondary education in 1985 and come from UNESCO (1998). The Catholic shares in 1900 and 1970 are obtained from Barro (2003).

Hofstede, G. H., & Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.* SAGE Publications.
Hübner, M., & Vannoorenberghe, G. (2015). Patience and long-run growth. *Economics Letters, 137*, 163–167.
Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155.
Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*(1), 53–69.
Jacob, B. A. (2002). Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review, 21*(6), 589–598.
Jamison, E. A., Jamison, D. T., & Hanushek, E. A. (2007). The effects of education quality on income growth and mortality decline. *Economics of Education Review, 26*(6), 771–788.
John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research, 2*(1999), 102–138.
Kautz, T., Heckman, J. J., Diris, R., ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. NBER Working Paper No. 20749.
Kimura, D. (2004). Human sex differences in cognition, fact, not predicament. *Sexualities, Evolution & Gender, 6*(1), 45–53.
Krueger, A. B., & Lindahl, M. (2001). Education for growth: Why and for whom? *Journal of Economic Literature, 39*(4), 1101–1136.
Lee, D. W., & Lee, T. H. (1995). Human capital and economic growth tests based on the international evaluation of educational achievement. *Economics Letters, 47*(2), 219–225.
Lee, J.-W., & Barro, R. J. (2001). Schooling quality in a cross–section of countries. *Economica, 68*(272), 465–488.
Levin, H. M. (2012). More than just test scores. *Prospects, 42*(3), 269–284.
Linton, R. (1945). *The cultural background of personality.* New York: Appleton-Century.
Sala-i-Martín, X., Doppelhofer, G., & Miller, R. I. (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *The American Economic Review, 94*(4), 813–835.
Méndez, I. (2015). The effect of the intergenerational transmission of noncognitive skills on student performance. *Economics of Education Review, 46*(C), 78–97.
Mueller, G., & Plug, E. (2006). Estimating the effect of personality on male and female earnings. *Industrial & Labor Relations Review, 60*(1), 3–22.
Murphy, K. M., & Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics, 20*(1), 88–97.
Nelson, R. R., & Phelps, E. S. (1966). Investment in humans, technological diffusion, and economic growth. *The American Economic Review, 56*(1/2), 69–75.
Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics,* 1067–1101.
Nollenberger, N., & Rodríguez-Planas, 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.
Nunn, N. (2012). Culture and the historical process. *Economic History of Developing Regions, 27*(Sup1), 108–126.
OECD (2009). *PISA 2006: Technical report.* Paris: OECD Publishing.
Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598–609.
Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psycholgy, 57*, 27–53.
Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school trends and predictors. *Educational Researcher, 44*(6), 336–346.
Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality, 43*(2), 137–145.
Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy, 98*(5), 71–102.
Sachs, J. D., Warner, A., Åslund, A., & Fischer, S. (1995). Economic reform and the process of global integration. *Brookings Papers on Economic Activity, 1995*(1), 1–118.
Sachs, J. D., & Warner, A. M. (1997). Fundamental sources of long-run growth. *The American Economic Review, 87*(2), 184–188.
Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*(1), 168–182.
Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58*(8), 1438–1457.
Sunde, U., & Vischer, T. (2015). Human capital and growth: Specification matters. *Economica, 82*(326), 368–390.
Tabellini, G. (2010). Culture and institutions: Economic development in the regions of Europe. *Journal of the European Economic Association, 8*(4), 677–716.
UNESCO (1998). *World education report: Teachers and teaching in a changing world.* Paris: UNESCO.
Wechsler, D. (1940). *Nonintellective factors in general intelligence. 37*, 444–445.
West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., & Gabrieli, J. D. (2016). Promise and paradox: Measuring students non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis, 38*(1), 148–170.
Woessmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics, 65*(2), 117–170.
WorldBank (2002). *World development indicators 2002.* World Bank Publications.
Zamarro, G., Cheng, A., Shakeel, M., & Hitt, C. (2016). Comparing and validating measures of character skills: Findings from a nationally representative sample. *EDRE Working Paper* 2015-08.
Zamarro, G., Hitt, C., & Mendez, I. (2016). When students don't care: Reexamining international differences in achievement and non-cognitive skills. *EDRE Working Paper* No. 2016-18.
Zamarro, G., Nichols, M., Duckworth, A., & D'Mello, S. (2017). Further validation of survey-effort measures of conscientiousness: Results from a sample of high school students. *The Character Assessment Initiative Working Paper*.