RUHR

ECONOMIC PAPERS

Christoph Helbach

The Interplay of Standardized Tests and Incentives

An Econometric Analysis with Data from PISA 2000 and PISA 2009

#356

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #356

Responsible Editor: Volker Clausen

ISSN 1864-4872 (online) – ISBN 978-3-86788-410-5

Ruhr Economic Papers #356

Christoph Helbach

The Interplay of Standardized Tests and Incentives

An Econometric Analysis with Data from PISA 2000 and PISA 2009

UNIVERSITÄT DUISBURG ESSEN

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788410
ISSN 1864-4872 (online)
ISBN 978-3-86788-410-5

Christoph Helbach[1]

The Interplay of Standardized Tests and Incentives – An Econometric Analysis with Data from PISA 2000 and PISA 2009

Abstract

Since their first implementation in 2000, the PISA studies have attracted public attention and spurred the demand for institutional changes in schooling systems. The introduction of standardized student tests and of incentives for schools and teachers are notable examples of such institutional changes. This paper examines the effects of these particular developments. Identification is based on within-country variation between PISA 2000 and PISA 2009. The results indicate that comparing schools by means of standardized student test results is a promising measure, while evaluating teachers this way decreases the overall performance of a schooling system. The discussion provides possible explanations for these ambiguous findings.

JEL Classification: I20, I21, I29

Keywords: Education economics; incentives; standardized tests; PISA

July 2012

1 University of Duisburg-Essen. – I am grateful to Jeannette Brosig-Koch, Matthias Giesecke, Timo Heinrich, Laura Helbach, Klemens Keldenich, Miriam Krieger, and seminar participants at University of Duisburg-Essen for helpful comments. – All correspondence to: Christoph Helbach, Faculty of Economics and Business Administration, University of Duisburg-Essen, Universitätsstr. 12, 45117 Essen, Germany, E-Mail: christoph.helbach@ibes.uni-due.de.

1. Introduction

Various studies have shown the remarkable impact of schooling quality on economic growth[1] and during the last decade governments the world over have made school education policies a focus of their attention. The Economist (2011) refers to the latter as "The great schools revolution" and emphasizes the crucial role that the PISA studies have played in this context by providing both internationally comparable data and eye-opening analyses. Since the first PISA implementation in 2000 researchers have used these rich datasets to identify key elements of successful education policies. One of the most important policy implications is summarized by Hanushek (2006, p. 866) who states that "[…] evidence […] suggests that pure resource policies that do not change incentives are unlikely to be effective." There are two implications from this conclusion.

First, incentives may serve as a useful tool to improve an educational system's outcome. On principle, every agent who is involved in educational production can be incentivized. Recent research, however, has focused on incentives for students, teachers, and schools.[2] While the literature on student incentives yields rather inconclusive results,[3] direct incentives for schools and in particular for teachers tend to affect students' outcomes positively. This holds true for studies in the United States (Figlio & Kenny, 2007), in Israel (Lavy, 2002, 2009), in India (Muralidharan & Sundararaman, 2011) as well as between countries (Wößmann, 2011). However, some papers show that teachers respond "too well" to incentives and consequently disregard non-incentivized goals. Reback (2008), Jacob (2005), Eberts, Hollenbeck & Stone (2002), and Glewwe, Ilias & Kremer (2010) find that teachers focus only on the incentivized target group and the success of incentivized measures. Goals that are not part of the incentive scheme but may still be desirable from society's point of view may not benefit or may even suffer from the incentive schemes.

Second, the institutional framework of a schooling system may matter more than the mere availability of resources. Among the institutions that have attracted analysts' attention are accountability systems such as curriculum-based external exit exams. Figlio & Loeb (2011) discuss the incentive effect of such systems on schools if the schools' future development – implicitly or explicitly – depends on the performance measures. Furthermore, external exams may provide a guideline for agents in the education system and thus reduce insecurity of students, teachers, and schools.[4] Cross-country evidence suggests that centralized exit exams indeed affect student's outcomes positively (Bishop, 1997; Wößmann, Lüdemann, Schütz & West, 2009). Jürges & Schneider

[1] See Hanushek & Wößmann (2008) for a survey.

[2] Of course, teachers are part of the school and incentives for schools and teachers may consequently not be separable. Please refer to the hypotheses and the analysis of the estimation results for a further discussion on this topic.

[3] While girls respond to incentives in some cases (Angrist & Lavy, 2009; Kremer, Miguel & Thornton, 2009), in others they do not (Fryer Jr., 2011). For boys the consequences of explicit incentives seem even less promising (Angrist & Lavy; Fryer Jr.). In particular the study by Fryer Jr. points out how much the reaction of students to incentives may depend on the details of the respective setup.

[4] Another aspect is highlighted by De Paola & Scoppa (2010) for the case of centralized exit exams. These can on the one hand serve as an incentive device for the schools. On the other hand, standardized tests provide a less dispersed measure of graduates' abilities for employers. Consequently, students themselves could be incentivized by such an institution as they anticipate that prospective employers put more weight on results from standardized exams than from non-standardized ones.

(2010) as well as Jürges, Schneider, Senkbeil & Carstensen (2012) confirm this positive tendency by analyzing variation between German states. However, they also conclude that centralized exit exams can reduce students' interest, at least in mathematics. Furthermore, a meta study by Holme, Richards, Jimerson & Cohen (2010) casts some doubt on the robustness of the positive evidence mentioned above. While Figlio & Loeb share those doubts in particular when it comes to reading scores, they conclude that "school accountability improves average student performance in affected schools, at least in general" (p. 410).

While this literature suggests that both teacher and school incentives as well as accountability can improve a schooling system, new evidence by Hanushek, Link & Wößmann (2011) points to the need for a more detailed analysis. They combine the data from various PISA studies and, based on within-country variation, find that institutions – such as school autonomy in their paper – may work well in certain countries and cultures while they are counterproductive in others. Furthermore, the authors discuss the positive interaction of autonomy with accountability and thus point out the importance of the interplay of various institutions.

In the present paper data from PISA 2000 and PISA 2009 are combined to analyze the interplay of accountability and teacher/school incentives. The estimations reveal that standardized tests per se do not have a consistently significant impact on PISA scores. It is shown, however, that it is important how the results of these tests are used. While the combination with teacher evaluation yields a significant decrease of PISA scores, it seems useful if schools as a whole are compared and rated based on the results of standardized tests.

The paper proceeds as follows. The following section introduces the econometric approach, discusses the key variables, and states the hypotheses. In section 3 and 4 the estimation results are reported and discussed. Section 5 concludes.

2. Econometric approach

This paper is based on the datasets of PISA 2000 and PISA 2009 that are combined into one dataset to analyze institutional effects of within-country variation on PISA scores.[5] There are several reasons for this approach. First, the abovementioned literature on the effect of schooling quality on economic growth has shown the importance of PISA outcomes as schooling quality measures. Second, the rich PISA datasets include various student- and school-level variables that can be used as controls when analyzing the effects of standardized tests and teacher/school incentives on student outcomes. Third, the focus of PISA alternates between reading, mathematics, and science literacy. In 2009 reading was the first focus to be repeated, which results in very similar school and student questionnaires that simplify the merging process of the data. Fourth, the almost ten years that lay between the studies provide enough time for – potentially PISA-triggered – schooling reforms to affect student performance.

[5] For a detailed description of the PISA datasets and study framework please refer to Adams & Wu (2002) and Organisation for Economic Co-operation and Development (1999, 2009).

2.1 Data preparation

The datasets of PISA 2000 and PISA 2009 are reduced to students from countries that took part in both waves, aligned with regards to the variable names and codings, and finally merged into one school and one student dataset.[6] Within these datasets missing values are imputed following the approach described by Wößmann et al. and using Stata files by Royston (2009). As a first step, missing values for elementary variables are replaced by the group median, i.e. the school median for student variables and the country median of the respective wave for school variables.[7] Second, the variables to be imputed are regressed on these elementary variables, country dummies and a dummy variable indicating the PISA study a case stems from. Third, the missing values are predicted by the coefficients of these estimations. To make sure that none of the results of this paper is driven by the imputed values, every subsequent regression includes one dummy variable per explanatory variable that indicates whether the respective value is imputed. In addition, interaction terms between these dummy variables and the respective variable itself are included.[8] These procedures result in a dataset of 490,226 students from 38 countries.

2.2 Key variables and hypotheses

Three variables from the PISA school questionnaire are of primary interest here. The respective questions from PISA 2009 (Organisation for Economic Co-operation and Development, OECD, 2008) read as follows (extract):

Question 15

Generally, in your school, how often are students in <national modal grade for 15-year-olds> assessed using the following methods?

a) Standardised tests

Question 16

In your school, are assessments of students in <national modal grade for 15-year-olds> used for any of the following purposes?

d) To compare the school to <district or national> performance

f) To make judgements about teachers' effectiveness

The corresponding questions from the PISA 2000 school questionnaire are very similar. For the subsequent analysis the answers are coded as indicator variables with value one if standardized tests are carried out, if they are used to compare schools, and if they are used to make judgements about teachers' effectiveness, respectively.

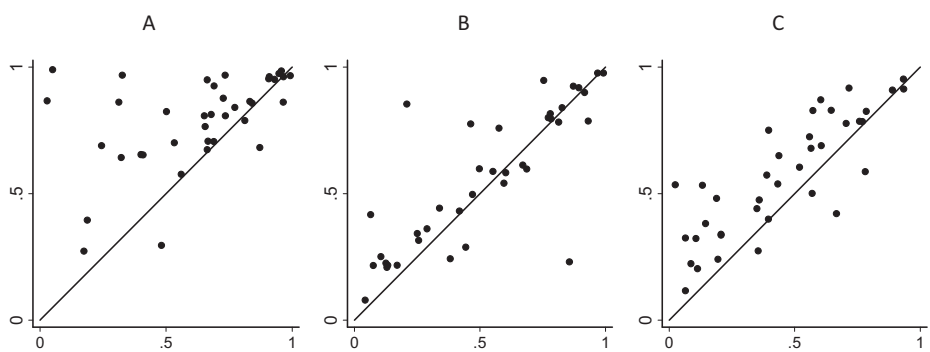
[6] Furthermore, students from Austria, Canada, France, and Korea are removed from the dataset due to missing key variables and student weights.

[7] Elementary variables include e.g. age, gender, grade, migration background, number of books at home, and country of test for students and share of public funding, student-teacher ratio, size, shortage of computers, and country of test for schools. Elementary variables are missing for less than 3 % of the students and less than 10.7 % of the schools.

[8] For country level variables the respective country level averages of the indicator variables are constructed and used for the estimation. However, the key results of this paper do not depend on the inclusion of these variables.

Aside from the direct interpretation of the variable, the use of standardized tests for the assessment of 15 year olds can as well be interpreted as an indicator for the tendency to use centralized exams in a schooling system in general. Panel A of Figure 1 shows that there is a strong trend to increase the use of standardized tests at the country level, with only 6 of 38 countries reducing the average use of this assessment method. This trend is significant (p=0.000, Wilcoxon signed-rank test). From the point of view of a standard economic theory, this trend should only result in changes of PISA performance if two conditions are satisfied. First, the content of the standardized tests needs to be related to the skills assessed by PISA. Second, agents in the schooling system need to be motivated through incentives – explicit or implicit – to aim at higher student scores in standardized tests. In this study it is assumed that such incentives are given at least to some extent for schools and teachers if the answers to questions 16 d) and 16 f), respectively, are "yes". While the design – and thus the power – of these incentives remain unclear, it seems reasonable to assume that a school that is compared to other schools based on the results of standardized student assessments feels more bound to set standards than a school for which this is not the case. Similarly, a teacher whose effectiveness judgement depends on his students' performance in standardized tests is assumed to comply more with the set standards than a colleague for whom this is not the case. The variables are thus subsequently referred to as school incentives and teacher incentives, respectively. Panel B and C of Figure 1 display the country-level development of the incentive variables between 2000 and 2009. Both variables reveal a significant upwards shift (p=0.024 and p=0.000, respectively, Wilcoxon signed-rank test).

Figure 1: Weighted country level averages of key variables in 2000 and 2009



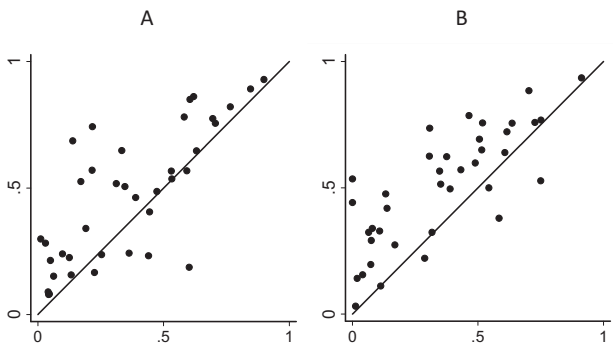
Notes: Weighted country level averages in 2000 (horizontal axis) and 2009 (vertical axis): Usage of standardized tests as assessments (Panel A), usage of assessments for teacher evaluation (Panel B) and for school comparison (Panel C). In addition, the 45 degree reference line is displayed.

Notably, the developments of the incentive variables are (weakly) significantly correlated with the developments in the use of standardized tests.[9] In Portugal, Poland, and Germany for instance an increase in the use of standardized tests between 2000 and 2009 came with an increase in the usage of student assessments for school comparison. Despite these examples, the variables are not per definition linked to each other, as can be seen from cases like Denmark and Latvia where school

[9] A spearman rank correlation at country level yields a coefficient of 0.410 (0.282) and a p-value of 0.01 (0.09) for the correlation of standardized tests and school incentives (teacher incentives).

incentives increased despite a constantly high level of standardised exams. Consequently, interaction terms for the variable from question 15 and the incentive variables are constructed and included in the subsequent analysis. The country-level averages of these interaction variables are displayed in Figure 2.

Figure 2: Weighted country level averages of interaction terms of key variables in 2000 and 2009



Notes: Weighted country level averages in 2000 (horizontal axis) and 2009 (vertical axis): Usage of standardized tests as assessments interacted with usage of assessments for teacher evaluation (panel A) and usage of assessments for school comparison (panel B), respectively. In addition, the 45 degree reference line is displayed.

As discussed before, research clearly indicates that standardized tests as well as teacher and school incentives tend to positively influence PISA performance. This leads to the first hypothesis:

H1: Standardized tests, teacher incentives, and school incentives all have a positive impact on PISA scores.

The present paper is the first to explicitly analyze the interplay of standardized tests and incentives. As expressed by H1, it is concluded from the literature that incentives affect agents' behaviour. Under the assumption that the content of the standardized tests is related to the content examined by PISA, this effect should be even larger if objective criteria such as standardized tests are used as performance measures. This leads to the second hypothesis:

H2: The combination of school and teacher incentives with standardized tests yields a further increase of PISA performance.

On the other hand, hypothesis H2 may be rejected if the standardized tests are not or poorly related to PISA assessments.

2.3 Regression models and identification

Similar to the framework of Hanushek et al., the analysis of this paper is based on the following education production function, subsequently referred to as the *general model*:

$$T_{cti} = \beta_F F_{cti} + \beta_S S_{cti} + \beta_P P_{ct} + \varepsilon_{cti}$$

where the achievement T of a student i in country c at a point of time t is the results of various inputs. These inputs stem from the student and her family (F), her school (S), the respective country's policies (P), and an error term (ε). Note however that a clean distinction between school properties and country policies is difficult or even impossible. Both public and private schools are dependent on laws and guidelines and thus even basic school properties such as location or size might be subject to changes for political reasons. For the estimations of the paper at hand the general model is converted to the *full model* that reads as follows:

$$T_{cti} = \beta_F F_{cti} + \beta_S S_{cti} + \beta_C D_c + \beta_T D_t + \theta_{cti}$$

Here, time invariant country properties such as robust policies or economic circumstances are captured by country fixed effects (D_c). Time fixed effects are controlled for by D_t. The PISA data provide rich background information on students and schools. The *full model* includes student variables on gender, age, family and migration background, parental education, and number of books at home. On the school level, various measures e.g. on general school properties, autonomy, equipment and staff, assessment, or admission are included as controls. In addition, GDP per capita is included at the country level (Heston, Summers & Aten, 2011).[10]

The key variables for this paper are measured at school level and are as such properties of a particular school. Identification in the *full model* is then based on the independence of the error term from the respective key variable, which cannot be taken for granted. On the contrary, it seems likely that student selection into schools is correlated with one or several of the key variables. In Germany for instance centralized exams for 15-year-olds have been implemented in half of the federal states between 2000 and 2009. However, not every student has to sit these exams. Particularly, in some states students who plan to proceed to the final secondary-school examinations (Abitur) – the prerequisite for tertiary education – do not sit the central exams at the age of 15.[11] Another potential selection bias could arise if schools are not obliged to publish results from standardized tests but can do so voluntarily. In such a scenario it is possible that only results from good schools are published and can be compared by parents. If parents then decide to send their children to the schools that provide this information, these schools may get the chance to choose more talented students. These exemplary selection effects are likely to bias the estimation. To avoid these biases the identification strategy of Hanushek et al. is implemented:

Key variables are measured as weighted averages at the country level and included in the regressions. Identification is then based on the variation of country-level averages between 2000 and 2009. Figure 1 indicates that this variation is substantial and thus can be used to identify the effects of the interplay between standardised exams and school and teacher incentives. This model is subsequently referred to as *full model key.*[12] Note however, that this approach may lead to a different kind of endogeneity problem. If changes in the key variables at the country level come with

[10] Liechtenstein is excluded from the regression due to missing GDP information.

[11] I am grateful to Andrea Schwermer from the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (Kultusministerkonferenz) for this pointer.

[12] The indicator variables for missing values are also averaged at the country level and included in the model both as such and interacted with the respective key variable averages. For a descriptive overview of the country level averages of the key variables please refer to the appendix.

changes in non-observable variables the estimated coefficients may again be biased. This would be the case for example if other reforms took place between 2000 and 2009. To reduce this bias other school-level variables are as well included in addition as country level averages. Even though this approach may suffer from very few degrees of freedom and may leave some variables unobserved, it seems likely to tackle large parts of the endogeneity problem as various possible reforms are controlled for. Among others, measures on school autonomy (taken from Hanushek et al.), budget, and admission criteria are included at the state level in these regressions. This model is subsequently referred to as *full model all.*[13]

All models are estimated by least squares regressions weighted with sampling probabilities. Robust standard errors are clustered at country level.

3. Estimation results

The results of the estimations are reported in Table 1.[14] As a first observation it can be noted that the two models yield qualitatively similar results. Despite the few degrees of freedom the *full model all* estimations tentatively confirm the results of the *full model key* regressions. The discussion is thus mainly based on the latter.

Table 1: Impact of key variables on PISA scores

	full model key			*full model all*		
Dependent variable	RL	ML	SL	RL	ML	SL
Standardized tests	25.01	32.01	88.86**	40.57	45.97	11.10
	(50.65)	(44.65)	(40.92)	(26.08)	(29.76)	(33.26)
School incentives	-181.90**	-147.41*	-56.30	-100.74**	-88.26*	-136.13**
	(75.98)	(86.66)	(75.09)	(47.88)	(44.35)	(65.31)
Standardized tests x school incentives	258.33***	312.40***	117.30	169.01***	159.98***	205.98***
	(77.99)	(94.91)	(76.78)	(56.72)	(55.82)	(71.39)
Teacher incentives	63.26	119.86**	107.08**	-23.75	96.05***	51.52
	(62.45)	(57.47)	(49.55)	(36.11)	(34.41)	(45.58)
Standardized tests x teacher incentives	-135.06*	-264.52***	-206.94***	-19.73	-167.77***	-108.43*
	(71.56)	(68.60)	(57.84)	(39.82)	(42.77)	(55.29)
Observations	489,583	409,884	409,736	469,593	394,060	393,922
Countries	37	37	37	35	35	35

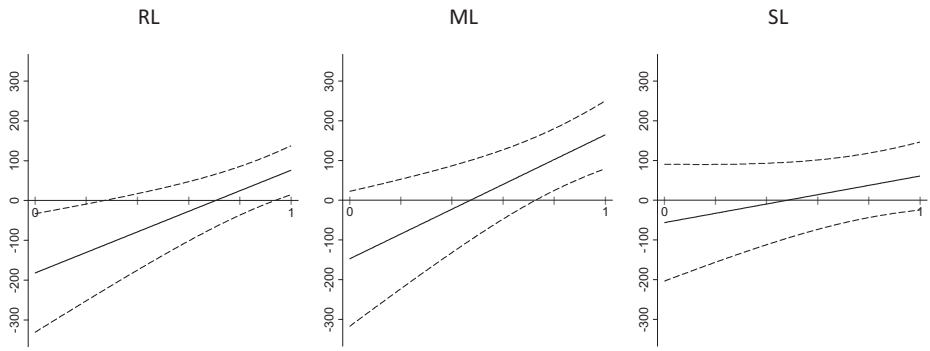
Notes: Impact of the key variables measured as country-level averages on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL). Estimates of *full model key* and *full model all* by least squares weighted with sampling probabilities. Controls include country fixed effects, time fixed effects, controls for missing values, GDP per capita, school level variables (e.g. size, location, staff, funding, autonomy, admission rules), and student level variables (e.g. gender, age, education of parents, migration background). Robust standard errors clustered at country level in parentheses. Significance levels: * 1 %, ** 5 %, * 10 %.**

[13] Controlling for missing value averages at the country level is not possible in this model because of too few degrees of freedom. Furthermore, students from Albania, and Peru are excluded in this model due to missing values at the country level.

[14] For the complete results please refer to the appendix.

The results shown in Table 1 translate into the marginal effects of school incentives and teacher incentives illustrated in Figures 3 and 4.[15] The graphs show that changes in school incentives tend to increase PISA scores if the incentives are linked to high levels of assessment standardization. On the other hand, such changes reduce PISA scores if linked to low levels of assessment standardization. These effects are most pronounced and widely significant for reading and mathematical literacy while they are not significant for scientific literacy. As for teacher incentives, the results are contrary. While evaluating teachers via student assessments in general improves PISA scores (significantly so for mathematical and scientific literacy), connecting evaluations to standardized tests is counterproductive for all three literacies.[16]

Figure 3: Marginal effects of school incentives as a function of standardized testing



Notes: Marginal effects of school incentives (solid line ———) on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL) as a function of standardized testing. The dashed lines (– – –) display 95 % confidence intervals.

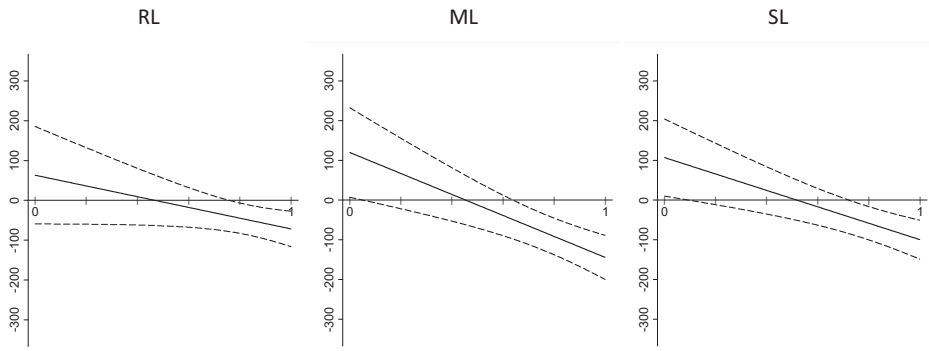
Note that leaving the interaction terms out of the regressions entangles the results and yields widely insignificant effects for standardized exams, positive effects of school incentives and negative effects of teacher incentives (cf. Table A5 in the Appendix). Altogether, these results lead to a partial rejection of all stated hypotheses.

Quantitatively, the large coefficients for the variables, particularly for the interaction terms, underline their importance. It should be noted, however, that the variation within countries and between points of time that is used for identification in these regressions does not cover the whole possible range. The interaction terms for example do not change between 2000 and 2009 by more than 0.54 in absolute value in any country. The medians of changes in the interactions of standardized tests with teacher incentives and school incentives are 0.06 and 0.13, respectively. These changes imply an average decrease (increase) of PISA scores by 12.1 % (29.8 %) of a standard deviation as a result of combining standardized tests with teacher incentives (school incentives).

[15] Brambor, Clark & Golder (2006) underline the importance of considering marginal effects in interaction models and provide useful tools for doing so.

[16] To test whether the results are driven particularly by low or high performing students, the sample is split at the country-by-wave median for each literacy category and *full model key* is then re-estimated for both subsamples. The results are very similar to those shown in Figures 3 and 4, although somewhat less pronounced for the subsample comprising the top 50 % of the students (cf. Appendix, Figures A1 to A4).

Figure 4: Marginal effects of teacher incentives as a function of standardized testing



Notes: Marginal effects of teacher incentives (solid line ——) on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL) as a function of standardized testing. The dashed lines (– – –) display 95 % confidence intervals.

4. Discussion

The results show that standardized tests per se do not consistently increase PISA outcomes.[17] It is then crucial to analyze what the test results are used for. The key to understanding the presented results is the question why a link between standardized tests and comparing (and thus incentivizing) schools is fruitful while a link with evaluating (and thus incentivizing) teachers is not.

The pertinent literature indicates consistently that teachers respond to incentives. However, this also includes possible side effects of teacher incentives. Jacob & Levitt (2003) for example find indication for cheating by teachers and administration despite standardized student testing. They show that the prevalence of cheating strongly responds to changes in incentives. In addition to such unintended behavioural consequences of incentives, the results at hand could also be explained if the content of the standardized tests were not closely related to the competence based measures of PISA. If teachers then focus too much on their students' performance in the incentivized test PISA performance may suffer. Against the background of research supporting the relevance of PISA and what it measures (Hanushek & Wößmann, 2008; Schleicher, 2007) and the present paper, this interpretation should lead to an adjustment of the standardized tests or a detachment of their results and teacher incentives.

A third and complementary explanation for the overall negative effect of teacher incentives is provided by the method used in this paper. Identification here is based on policy changes within countries rather than variation between countries (as it is used e.g. by Wößmann). Thus, history may play a role. For example, the self selection of people into the teaching body is likely to be driven not only by the profession itself but also by aspects such as job security. Dohmen & Falk (2010) find that teachers in Germany are more risk averse than the average German population. Such preferences

[17] Note that this result does not necessarily contradict results from other research. First, other studies often use curriculum based external exit exams as accountability measure. They may thus include an additional positive effect through student incentives to perform well in these exams (cf. footnote 4). Furthermore, some of the regressions presented here also reveal a significantly positive effect.

are likely to influence the effect of incentives. In a laboratory experiment in Germany, Helbach & Keldenich (2012) find that prospective teachers respond differently (i.e. more weakly) than other students to incentive scheme variations. Even though this interpretation by itself is unlikely to lead to negative effects of incentives it is possible that it reduces the positive incentive effects that are shown in Table 1 and thus contributes to the negative overall effect of teacher incentives.

In contrast to teachers, schools as a whole may not be able to influence the style and content of teaching to the same extent. Instead it is likely that incentivized schools are able to improve general measures such as atmosphere, learning environment, and general discipline (e.g. truancy). Claes, Hooghe & Reeskens (2009) provide evidence that schools play a major role in reducing truancy rates. These general improvements are likely to improve schooling outcomes – independently of the respective curriculum. However, schools are only incentivized to exert the required effort if they are held accountable for their performance. The results show that standardized tests can serve as an accountability measure while other student assessments cannot (cf. Table 1). This seems reasonable as results from non-standardized tests may not convince the public of the good or bad quality of a school. Furthermore, as schools may be able to affect the framing of such results, they have no incentive to sustain high schooling quality. Instead, they may even be encouraged to develop tools for disguising deficits.

5. Conclusion

In Finland, the average results of the central exit exams (ylioppilastutkinto) of every Finnish high school are published every year by various media. The Supreme Administrative Court (Korkein Hallinto-Oikeus) of Finland underlined the importance of these publications by judging that the results need to be made available to the public in electronic form as well (Korkein Hallinto-Oikeus, 2007).[18] It is this kind of comparison that imposes pressure – and hence incentives – on the schools. At the same time, teachers can be rewarded – and hence incentivized – by individual cash bonuses that are not necessarily based on their students' performance in centralized exams. Based on the previous discussion, this is a sound incentive mix that may be one reason why Finland has proven successful in every PISA study so far.

The analysis has shown that school comparisons are a main channel through which standardized tests take effect. It is thus promising that both institutions are strongly correlated. In contrast, the correlation of the use of teacher incentives and standardized tests seems counterproductive in the light of the analysis while teacher incentives in general remain a way to achieve better schooling outcomes.

[18] The ranking of 2011 is available for example on the website of Helsingin Sanomat, the largest newspaper in Finland (Helsingin Sanomat, 2011, 2012).

References

Adams, R. J. & Wu, M. (2002). *PISA 2000 technical report*. Paris: OECD.

Angrist, J. & Lavy, V. (2009). The effects of high stakes high school achievement awards: evidence from a randomized trial. *American Economic Review*, *99*(4), 1384–1414.

Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. *American Economic Review*, *87*(2), 260–264.

Brambor, T., Clark, W. R. & Golder, M. (2006). Understanding interaction models: improving empirical analyses. *Political Analysis*, 14(1), 63–82.

Claes, E., Hooghe, M. & Reeskens, T. (2009). Truancy as a contextual and school-related problem: a comparative multilevel analysis of country and school characteristics on civic knowledge among 14 year olds. *Educational Studies*, *35*(2), 123–142.

De Paola, M. & Scoppa, V. (2010). A signalling model of school grades under different evaluation systems. *Journal of Economics*, *101*(3), 199–212.

Dohmen, T. & Falk, A. (2010). You get what you pay for: incentives and selection in the education system. *Economic Journal*, *120*(546), F256–F271.

Eberts, R., Hollenbeck, K. & Stone, J. (2002). Teacher performance incentives and student outcomes. *Journal of Human Resources*, *37*(4), 913–927.

Figlio, D. & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. J. Machin, & L. Wößmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 383–421). Amsterdam: Elsevier.

Figlio, D. N. & Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics*, *91*(5-6), 901–914.

Fryer Jr., R. G. (2011). Financial incentives and student achievement: evidence from randomized trials. *Quarterly Journal of Economics*, *126*(4), 1755–1798.

Glewwe, P., Ilias, N. & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*(3), 205–227.

Hanushek, E. A. (2006). School resources. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 2, pp. 865–908). Amsterdam: Elsevier.

Hanushek, E. A., Link, S. & Wößmann, L. (2011). Does school autonomy make sense everywhere? panel estimates from PISA. *National Bureau of Economic Research Working Paper Series*, *17591*.

Hanushek, E. A. & Wößmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, *46*(3), 607–668.

Helbach, C. & Keldenich, K. (2012). Teaching in the lab: financial incentives in the education process. *Working Paper*. *Ruhr Economic Papers,* 328.

Helsingin Sanomat. (2011). Lukioiden menestys paremmuusjärjestyksessä kevään 2011 ylioppilaskirjoituksissa. Retrieved March 5, 2012, from http://www.hs.fi/kotimaa/artikkeli/1135266565425

Helsingin Sanomat. (2012). International edition - about. Retrieved February 29, 2012, from http://www.hs.fi/english/about/

Heston, A., Summers, R. & Aten, B. (2011). Penn world table version 7.0. *Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania*.

Holme, J. J., Richards, M. P., Jimerson, J. B. & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. *Review of Educational Research*, *80*(4), 476 –526.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5–6), 761–796.

Jacob, B. A. & Levitt, S. D. (2003). Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, *118*(3), 843–877.

Jürges, H. & Schneider, K. (2010). Central exit examinations increase performance... but take the fun out of mathematics. *Journal of Population Economics*, *23*(2), 497–517.

Jürges, H., Schneider, K., Senkbeil, M. & Carstensen, C. H. (2012). Assessment drives learning: the effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, *31*(1), 56–65.

Korkein Hallinto-Oikeus. (2007). KHO:2007:39. Retrieved March 5, 2012, from http://www.kho.fi/paatokset/39605.htm

Kremer, M., Miguel, E. & Thornton, R. (2009). Incentives to learn. *Review of Economics and Statistics*, *91*(3), 437–456.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy*, *110*(6), 1286–1317.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, *99*(5), 1979–2011.

Muralidharan, K. & Sundararaman, V. (2011). Teacher performance pay: experimental evidence from India. *Journal of Political Economy*, *119*(1), 39–77.

Organisation for Economic Co-operation and Development. (1999). *Measuring student knowledge and skill - a new framework for assessment*. Paris: OECD.

Organisation for Economic Co-operation and Development. (2008). *Programme for international student assessment: school questionnaire for PISA 2009*. Paris: OECD.

Organisation for Economic Co-operation and Development. (2009). *PISA 2009 assessment framework - key competencies in reading, mathematics and science*. Paris: OECD.

Reback, R. (2008). Teaching to the rating: school accountability and the distribution of student achievement. *Journal of Public Economics*, *92*(5-6), 1394–1415.

Royston, P. (2009). Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. The Stata Journal, 9(3), 466–477.

Schleicher, A. (2007). Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess? *Journal of Educational Change*, *8*(4), 349–357.

The Economist. (2011). The great schools revolution: reforming education. *17 September*. Retrieved from http://www.economist.com/node/21529014

Wößmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review*, *30*(3), 404–418.

Wößmann, L., Lüdemann, E., Schütz, G. & West, M. R. (2009). *School accountability, autonomy and choice around the world*. Cheltham: Edward Elgar Publishing.

Appendix

Table A1: Weighted country-level averages of key variables

	Standardized test		Teacher incentives		School incentives	
	2000	2009	2000	2009	2000	2009
Albania	0.664	0.950	0.917	0.900	0.705	0.777
Argentina	0.667	0.707	0.595	0.541	0.355	0.274
Australia	0.534	0.701	0.339	0.443	0.520	0.604
Belgium	0.174	0.273	0.256	0.315	0.065	0.117
Brazil	0.652	0.807	0.463	0.775	0.397	0.750
Bulgaria	0.690	0.926	0.894	0.918	0.646	0.829
Chile	0.313	0.862	0.552	0.588	0.359	0.475
Czech Republic	0.727	0.877	0.671	0.613	0.437	0.650
Denmark	0.910	0.962	0.044	0.079	0.065	0.325
Finland	0.957	0.985	0.382	0.243	0.569	0.501
Germany	0.188	0.395	0.129	0.208	0.107	0.323
Greece	0.322	0.643	0.170	0.217	0.113	0.204
Hong Kong-China	0.327	0.968	0.576	0.759	0.206	0.339
Hungary	0.655	0.765	0.686	0.597	0.606	0.689
Iceland	0.965	0.862	0.443	0.288	0.781	0.587
Indonesia	0.930	0.952	0.969	0.977	0.770	0.784
Ireland	0.872	0.682	0.289	0.361	0.350	0.441
Israel	0.832	0.865	0.827	0.840	0.433	0.538
Italy	0.689	0.706	0.857	0.231	0.207	0.336
Japan	0.401	0.654	0.814	0.783	0.087	0.223
Latvia	0.966	0.962	0.872	0.925	0.717	0.918
Liechtenstein	0.560	0.577	0.132	0.217	0.132	0.533
Luxembourg	0.051	0.990	0.075	0.216	0.024	0.535
Mexico	0.812	0.789	0.771	0.800	0.559	0.725
Netherlands	0.842	0.857	0.471	0.497	0.667	0.421
New Zealand	0.679	0.813	0.498	0.599	0.934	0.914
Norway	0.908	0.954	0.106	0.251	0.565	0.679
Peru	0.244	0.689	0.783	0.796	0.396	0.399
Poland	0.735	0.968	0.932	0.787	0.389	0.573
Portugal	0.029	0.867	0.251	0.343	0.189	0.481
Romania	0.502	0.824	0.209	0.854	0.604	0.871
Russian Federation	0.773	0.841	0.992	0.976	0.785	0.825
Spain	0.482	0.296	0.418	0.431	0.194	0.241
Sweden	0.993	0.966	0.125	0.225	0.758	0.785
Switzerland	0.409	0.653	0.065	0.417	0.146	0.382
Thailand	0.736	0.807	0.755	0.948	0.573	0.828
United Kingdom	0.664	0.674	0.780	0.816	0.890	0.909
United States	0.948	0.975	0.603	0.583	0.934	0.953

Table A2: Weighted country-level averages of interaction terms of key variables

	Standardized tests x teacher incentives		Standardized tests x school incentives	
	2000	2009	2000	2009
Albania	0.605	0.850	0.519	0.757
Argentina	0.444	0.406	0.288	0.221
Australia	0.191	0.340	0.351	0.514
Belgium	0.046	0.081	0.013	0.031
Brazil	0.334	0.648	0.307	0.625
Bulgaria	0.620	0.861	0.464	0.786
Chile	0.171	0.526	0.138	0.419
Czech Republic	0.534	0.536	0.375	0.623
Denmark	0.044	0.079	0.065	0.324
Finland	0.363	0.243	0.544	0.500
Germany	0.041	0.089	0.041	0.156
Greece	0.062	0.151	0.019	0.141
Hong Kong-China	0.217	0.743	0.079	0.339
Hungary	0.474	0.486	0.432	0.571
Iceland	0.440	0.232	0.750	0.528
Indonesia	0.900	0.928	0.727	0.759
Ireland	0.254	0.237	0.317	0.324
Israel	0.706	0.755	0.390	0.496
Italy	0.602	0.187	0.169	0.274
Japan	0.312	0.518	0.074	0.196
Latvia	0.846	0.892	0.703	0.884
Liechtenstein	0.132	0.156	0.132	0.476
Luxembourg	0.051	0.214	0.000	0.535
Mexico	0.632	0.647	0.490	0.599
Netherlands	0.389	0.463	0.585	0.379
New Zealand	0.347	0.507	0.637	0.755
Norway	0.099	0.240	0.515	0.650
Peru	0.216	0.570	0.075	0.292
Poland	0.696	0.774	0.348	0.566
Portugal	0.011	0.298	0.000	0.442
Romania	0.139	0.686	0.308	0.736
Russian Federation	0.765	0.821	0.615	0.721
Spain	0.226	0.166	0.113	0.111
Sweden	0.125	0.225	0.751	0.768
Switzerland	0.031	0.282	0.108	0.329
Thailand	0.582	0.780	0.506	0.692
United Kingdom	0.531	0.567	0.607	0.639
United States	0.593	0.569	0.913	0.935

Table A3: Further results of *full model key* regressions

Dependent variable	RL	*full modell key* ML	SL
Student-level variables			
Female	23.84*** (1.83)	-16.91*** (2.18)	-5.24*** (2.02)
Age (in months)	-0.91*** (0.3)	-0.96*** (0.34)	-0.63*** (0.3)
Grade	35.85*** (2.53)	34.85*** (2.52)	30.54*** (2.38)
Family members living at home			
Mother	22.9*** (3.18)	22.9*** (2.07)	23.51*** (1.37)
Father	2.54** (2.34)	6.38** (2.54)	6.08** (2.77)
Brother(s)	-4.40 (1.99)	-3.09 (1.88)	-6.01 (1.7)
Sister(s)	-2.16 (2.26)	-2.79 (2.91)	-5.02 (2.21)
Education of mother			
ISCED 0	-10.95*** (3.49)	-11.28*** (2.59)	-6.52*** (4.15)
ISCED 1	-8.3*** (3.02)	-10.17*** (2.68)	-6.23*** (3.65)
ISCED 2	-8.04** (3.2)	-7.40** (2.91)	-6.40** (4.17)
ISCED 3b / 3c	-1.78 (2.11)	-2.38 (2.56)	-1.44 (3.9)
ISCED 3a / 4	2.41 (2.35)	-0.13 (1.97)	1.33 (3.63)
Higher education	7.15*** (1.68)	6.01*** (2.15)	9.69*** (2.4)
Education of father			
ISCED 0	-8.44 (4.47)	-1.82 (3.59)	-9.17 (4.98)
ISCED 1	-1.36 (3.92)	-0.32 (2.8)	-4.32 (3.96)
ISCED 2	-2.32 (3.53)	1.37 (2.7)	-4.19 (3.48)
ISCED 3b / 3c	0.71** (3.37)	5.87** (2.38)	0.84** (3.28)
ISCED 3a / 4	5.66*** (3.32)	8.18*** (2.52)	3.24*** (2.94)
Higher education	11.76*** (2.41)	16.98*** (2.58)	11.88*** (2.85)
Number of books at home			
0 - 10	76.73*** (9.08)	55.82*** (8.09)	115.49*** (12.11)
11 - 100	97.13*** (8.72)	73.74*** (8.37)	134.84*** (11.77)
101 - 500	126.63*** (8.85)	102.53*** (8.78)	164.34*** (11.92)
>500	133.60*** (9.12)	111.17*** (9.18)	174.23*** (12.11)
Migration background			
Student born in test country	4.94 (2.77)	0.74 (2)	1.29 (2.61)
Mother born in test country	1.61* (3.23)	5.16* (2.97)	6.10* (2.26)
Father born in test country	7.30*** (2.86)	8.38*** (2.29)	9.21*** (2.48)
Test language used at home	7.66 (5.14)	1.62 (4.62)	5.92 (5.56)
School-level variables			
Share of public funding	-0.05** (0.04)	-0.07** (0.03)	-0.09** (0.04)
Student-teacher ratio	-0.05 (0.05)	-0.06 (0.05)	-0.05 (0.06)
School size	0.01** (0)	0.01** (0)	0.01** (0)
School location			
Small town	3.28 (2.29)	1.69 (3.05)	2.40 (3.03)
Town	9.72** (2.68)	6.25** (2.93)	5.92** (3.07)
City	14.36 (3.94)	8.23 (4.9)	7.15 (4.38)
Large city	15.28 (5.68)	8.21 (6.41)	6.38 (5.76)

Table A3 (continued)

	full modell key		
Dependent variable	RL	ML	SL
Shortage assessment of principal			
Science teachers			
A lot	-0.82 (3.33)	4.73 (4.09)	2.62 (2.95)
To some extent	-4.58 (1.98)	-3.19 (2.13)	-3.56 (1.4)
Little	-3.57 (2.19)	-4.45 (2.67)	-3.77 (1.75)
Math teachers			
A lot	-2.41 (5.42)	-6.75 (6.12)	-5.19 (5.24)
To some extent	0.91 (3.4)	-1.95 (4.99)	-0.54 (2.86)
Little	-2.95* (1.44)	-2.78* (1.58)	-2.61* (1.52)
Language teachers			
A lot	3.72 (5.37)	5.48 (4.91)	3.44 (4.64)
To some extent	2.34 (2.94)	1.91 (3.29)	1.36 (3.12)
Little	3.09 (2.34)	2.93 (2.4)	3.21 (2.94)
Supporting staff			
A lot	-1.02 (1.99)	-1.98 (1.75)	-1.81 (1.93)
To some extent	-0.91 (2.01)	-1.94 (1.97)	-1.36 (1.95)
Little	-0.45 (1.95)	0.12 (1.99)	-0.85 (2.38)
Laboratory material			
A lot	-5.5** (2.64)	-6.46** (3.12)	-5.7** (2.78)
To some extent	-7.44** (2.11)	-7.12** (2.66)	-5.42** (2.14)
Little	-2.81 (1.64)	-1.50 (2.03)	-1.13 (1.73)
Instructional material			
A lot	0.78 (3.74)	3.89 (3.72)	4.87 (4.81)
To some extent	0.81 (2.78)	2.36 (2.87)	1.48 (2.27)
Little	2.57 (2.52)	3.15 (2.51)	2.62 (2.19)
Computers			
A lot	1.08 (2.52)	0.03 (2.14)	-0.38 (2.58)
To some extent	6.36** (2.14)	5.99** (2.48)	5.78** (2.06)
Little	3.75** (1.92)	4.06** (1.93)	3.75** (1.87)
Library material			
A lot	-9.17*** (2.27)	-10.27*** (2.42)	-11.53*** (2.43)
To some extent	-1.59* (1.83)	-2.67* (1.53)	-3.80* (1.67)
Little	-2.19** (1.3)	-3.54** (1.33)	-5.05** (1.75)
Audio-visual resources			
A lot	-2.64* (2.08)	-4.37* (2.43)	-2.21* (1.84)
To some extent	-1.11 (1.88)	-3.23 (2.06)	-0.40 (1.35)
Little	0.49 (1.89)	-0.26 (1.33)	1.19 (1.52)
Assessments are used for			
Informing parents	6.37 (2.93)	3.45 (3)	6.19 (4.23)
Deciding about student's carreer	2.02 (2.24)	1.87 (2.63)	1.38 (2.73)
Grouping students	-0.07 (2.76)	-0.38 (2.62)	-1.06 (2.5)
Comparing school to others	2.05 (1.97)	1.74 (2.4)	1.64 (2.18)
Monitoring school	1.55 (2.42)	0.06 (2.38)	0.24 (2.04)
Judging teachers' effectiveness	-2.65 (1.86)	-2.73 (1.78)	-2.90 (2.11)

Table A3 (continued)

Dependent variable	RL	*full modell key* ML	SL
Autonomy, responsibility for			
Selecting teachers	2.03*** (1.68)	4.34*** (1.56)	2.12*** (1.52)
Firing teachers	0.47 (2.7)	-2.35 (2.71)	-0.89 (2.16)
Teachers' starting salary	-8.27** (1.98)	-5.52** (2.23)	-6.92** (1.65)
Teachers' salary increases	2.94 (3.13)	2.13 (2.86)	2.94 (2.52)
Formulating school budget	1.66 (1.39)	0.13 (1.51)	0.36 (1.51)
Allocating school budget	5.42*** (2.19)	5.14*** (1.85)	6.79*** (2.67)
Disciplinary policies	8.19** (5.29)	7.11** (3.31)	9.59** (4.8)
Assessment policies	0.55 (2.49)	1.78 (2.9)	1.49 (3.02)
Student admission	-5.78** (2.4)	-5.58** (2.47)	-6.81** (2.69)
Textbooks	2.09 (2.49)	1.14 (2.39)	4.30 (2.77)
Course content	-0.96 (1.82)	-0.62 (1.68)	-2.18 (1.9)
Course offers	2.77** (1.5)	3.72** (1.83)	4.08** (1.66)
Students are assessed by			
Standardized tests	-3.52 (2.35)	-2.98 (2.25)	-3.24 (2.42)
Teacher-developed tests	4.87 (3.97)	2.06 (4.19)	4.52 (6.27)
Teachers‘ judgemental ratings	1.29 (1.33)	1.64 (1.45)	1.27 (1.11)
Student <portfolios>	-1.89* (0.94)	-2.22* (1.22)	-2.41* (1.22)
Projects / Homework	4.81 (6.32)	7.18 (6.96)	5.65 (5.16)
Criteria for admission			
Residence			
Always	-1.72 (3.06)	-1.18 (2.78)	-1.06 (2.99)
Sometimes	-0.46 (1.41)	-0.86 (1.99)	-1.22 (1.54)
Student's record of performance			
Always	12.42*** (2.28)	12.06*** (2.18)	12.40*** (2.16)
Sometimes	3.46 (2.78)	3.19 (2.3)	3.12 (2.82)
Recommendation of feeder schools			
Always	-4.13 (3.21)	-4.48 (3.61)	-3.60 (3.85)
Sometimes	-1.27 (1.71)	-1.36 (2.41)	0.10 (2.32)
Parents‘ endorsement of philosophy			
Always	3.79 (1.99)	1.99 (1.76)	1.74 (1.94)
Sometimes	-0.06 (1.27)	-0.65 (1.31)	-1.06 (1.38)
Student's interest in programme			
Always	-7.44*** (1.96)	-8.15*** (2.83)	-6.12*** (2.33)
Sometimes	-1.67 (1.66)	-1.79 (2.38)	0.25 (2.09)
Family member of current student			
Always	-1.28 (2.29)	-2.41 (1.53)	-2.07 (2.22)
Sometimes	-0.99 (1.65)	-2.17 (1.86)	-2.18 (2.25)
Learning of students is hindered by			
Teachers' low expectations			
A lot	-11.19*** (4.28)	-11.74*** (3.97)	-12.24*** (5.63)
To some extent	-9.00** (3.05)	-9.04** (3.46)	-8.69** (3.37)
Little	-2.39 (1.86)	-1.33 (2.41)	-2.20 (2.08)

Table A3 (continued)

	full modell key		
Dependent variable	RL	ML	SL
Student absenteeism			
A lot	-21.17*** (2.53)	-24.47*** (2.53)	-19.73*** (3.05)
To some extent	-16.53*** (2.16)	-19.65*** (2.5)	-17.59*** (2.24)
Little	-5.20*** (2.25)	-8.17*** (2.46)	-6.82*** (2.15)
Poor student-teacher relations			
A lot	6.01 (3.39)	2.67 (5.1)	6.19 (5.67)
To some extent	5.40** (1.88)	5.74** (2.12)	5.64** (2.16)
Little	1.61 (1.5)	1.57 (1.89)	2.60 (1.35)
Disruption of classes			
A lot	-19.28*** (5.49)	-16.56*** (5.58)	-17.87*** (4.95)
To some extent	-16.99*** (3.91)	-16.82*** (4.04)	-17.51*** (3.89)
Little	-8.58** (3.86)	-8.62** (3.75)	-9.05** (3.42)
Teachers not meeting students' needs			
A lot	14.61*** (4.32)	11.82*** (4.18)	15.27*** (4.35)
To some extent	8.96*** (2.25)	9.81*** (2.28)	8.62*** (2.44)
Little	7.50*** (2.44)	6.37*** (2.13)	6.26*** (2.17)
Teacher absenteeism			
A lot	2.74 (4.25)	4.24 (3.43)	5.67 (3.36)
To some extent	1.30 (2.67)	2.66 (2.14)	2.44 (2.67)
Little	0.84** (1.63)	2.92** (1.3)	2.47** (1.43)
Students skipping classes			
A lot	-13.4*** (5.25)	-11.77*** (4.16)	-13.44*** (3.97)
To some extent	-14.71*** (4.3)	-13.55*** (3.84)	-13.51*** (3.31)
Little	-8.89** (3.65)	-9.38** (3.93)	-8.93** (3.23)
Lack of respect			
A lot	0.04 (2.26)	-0.56 (2.67)	-4.62 (2.7)
To some extent	-0.63 (2.87)	-1.86 (2.37)	-3.23 (2.77)
Little	0.05 (1.76)	-0.26 (1.21)	-1.27 (1.34)
Staff resisting change			
A lot	-0.07 (3.76)	-1.07 (4.41)	-2.38 (3.77)
To some extent	4.41 (2.06)	4.27 (2.62)	6.35 (2.36)
Little	3.72 (1.64)	2.83 (1.8)	3.67 (1.88)
Students using drugs			
A lot	5.66 (5.58)	3.43 (4.37)	6.89 (5.96)
To some extent	-0.69 (3.86)	-0.92 (3.5)	-0.59 (3.82)
Little	-2.55* (1.77)	-2.47* (1.22)	-3.11* (1.68)
Teachers beeing to strict			
A lot	7.28 (3.48)	7.87 (4.72)	3.22 (3.57)
To some extent	4.85* (2.6)	6.36* (3.69)	4.38* (2.92)
Little	0.97 (1.43)	1.26 (2.02)	0.32 (1.73)
Students bullying students			
A lot	-4.72 (2.74)	0.39 (2.31)	-1.88 (2.05)
To some extent	-5.11* (4.08)	-5.06* (2.96)	-5.16* (3.74)
Little	0.43 (2.08)	1.31 (2.47)	1.80 (2.52)

Table A3 (continued)

	full modell key		
Dependent variable	RL	ML	SL
Students not beeing encouraged			
A lot	5.09 (3.4)	4.99 (3.42)	6.62 (2.31)
To some extent	-1.14 (2.02)	-1.14 (1.95)	-1.63 (2.32)
Little	-0.46 (1.83)	-0.82 (1.95)	-0.82 (1.57)
Country-level variables			
Standardized tests	25.01 (50.65)	32.01 (44.65)	88.86** (40.92)
School incentives	-181.90** (75.98)	-147.41* (86.66)	-56.30 (75.09)
Standardized tests x school incentives	258.33*** (77.99)	312.40*** (94.91)	117.30 (76.78)
Teacher incentives	63.26 (62.45)	119.86** (57.47)	107.08** (49.55)
Standardized tests x teacher incentives	-135.06* (71.56)	-264.52*** (68.60)	-206.94*** (57.84)
GDP per capita	0.002 (0.002)	-0.002 (0.002)	0.001 (0.002)
Observations	489,583	409,884	409,736
Countries	37	37	37

Notes: Impact of the key variables measured as country-level averages on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL). Estimates of *full model key* by least squares weighted with sampling probabilities. Controls include country fixed effects, time fixed effects, and controls for missing values. Robust standard errors clustered at country level in parentheses. Significance levels: * 1 %, ** 5 %, * 10 %.**

Table A4: Further results of *full model all* regressions

Dependent variable	RL	*full modell all* ML	SL
Student-level variables			
Female	24.18*** (1.86)	-16.70*** (2.23)	-5.08*** (2.06)
Age (in months)	-0.90*** (0.31)	-0.97*** (0.34)	-0.62*** (0.31)
Grade	36.01*** (2.68)	34.98*** (2.64)	30.78*** (2.47)
Family members living at home			
Mother	24.49*** (3.05)	23.61*** (1.89)	23.93*** (1.16)
Father	2.98** (2.35)	6.62** (2.55)	6.14** (2.8)
Brother(s)	-5.02 (1.95)	-3.70 (1.85)	-6.36 (1.69)
Sister(s)	-2.79 (2.24)	-3.40 (2.87)	-5.45 (2.15)
Education of mother			
ISCED 0	-11.46*** (3.49)	-12.01*** (2.56)	-6.76*** (4.38)
ISCED 1	-9.08*** (2.96)	-10.80*** (2.52)	-6.52*** (3.76)
ISCED 2	-9.38** (3.17)	-8.16** (2.77)	-6.54** (4.18)
ISCED 3b / 3c	-2.84 (1.89)	-3.72 (2.18)	-1.76 (3.74)
ISCED 3a / 4	1.18 (2.27)	-0.95 (1.97)	1.31 (3.7)
Higher education	6.24*** (1.65)	4.86*** (2)	9.23*** (2.53)
Education of father			
ISCED 0	-9.34 (4.49)	-2.54 (3.74)	-9.15 (5.19)
ISCED 1	-1.75 (3.92)	-0.80 (2.89)	-4.03 (4.12)
ISCED 2	-2.94 (3.51)	0.77 (2.7)	-3.96 (3.64)
ISCED 3b / 3c	0.36** (3.38)	4.97** (2.35)	0.79** (3.43)
ISCED 3a / 4	5.19*** (3.31)	7.93*** (2.57)	3.68*** (3.06)
Higher education	11.36*** (2.5)	16.54*** (2.78)	12.17*** (2.99)
Number of books at home			
0 - 10	76.96*** (9.23)	53.32*** (6.81)	113.13*** (11.44)
11 - 100	96.87*** (8.76)	71.11*** (6.42)	132.52*** (10.73)
101 - 500	126.48*** (8.89)	99.99*** (6.75)	162.12*** (10.77)
>500	133.62*** (9.17)	108.85*** (7.14)	172.24*** (10.89)
Migration background			
Student born in test country	4.74 (2.75)	0.66 (1.98)	1.09 (2.55)
Mother born in test country	1.61* (3.18)	5.17* (2.94)	6.17* (2.25)
Father born in test country	7.03*** (2.84)	8.09*** (2.26)	9.10*** (2.48)
Test language used at home	7.42 (5.07)	1.33 (4.56)	5.47 (5.47)
School-level variables			
Share of public funding	-0.06** (0.04)	-0.08** (0.03)	-0.08** (0.04)
Student-teacher ratio	-0.03 (0.06)	-0.04 (0.05)	-0.04 (0.06)
School size	0.01** (0)	0.01** (0)	0.01** (0)
School location			
Small town	2.59 (2.3)	1.41 (3.04)	2.18 (3.06)
Town	9.10** (2.69)	5.74** (2.95)	5.29** (3.09)
City	13.20 (4.06)	7.11 (4.9)	6.06 (4.32)
Large city	14.44 (5.76)	7.16 (6.44)	5.56 (5.81)

Table A4 (continued)

Dependent variable	RL	*full modell all* ML	SL
Shortage assessment of principal			
Science teachers			
A lot	-2.14 (3.41)	2.90 (4.11)	0.95 (2.95)
To some extent	-5.26 (2.03)	-3.56 (2.23)	-3.79 (1.42)
Little	-4.39 (2.26)	-4.91 (2.7)	-4.21 (1.73)
Math teachers			
A lot	-1.69 (5.53)	-6.11 (6.17)	-4.84 (5.31)
To some extent	1.36 (3.46)	-1.50 (5.14)	-0.39 (2.94)
Little	-2.96* (1.46)	-2.66* (1.56)	-2.45* (1.49)
Language teachers			
A lot	3.98 (5.1)	6.09 (4.89)	4.04 (4.62)
To some extent	2.91 (2.95)	2.08 (3.26)	1.68 (3.14)
Little	2.99 (2.35)	2.80 (2.4)	3.15 (2.99)
Supporting staff			
A lot	-1.35 (2.05)	-2.70 (1.75)	-1.92 (2.09)
To some extent	-0.85 (2)	-2.22 (1.95)	-1.42 (1.97)
Little	-0.66 (1.96)	-0.30 (1.93)	-1.02 (2.37)
Laboratory material			
A lot	-5.49** (2.58)	-6.47** (3.06)	-6.09** (2.72)
To some extent	-7.72** (2.04)	-7.58** (2.61)	-5.89** (2.09)
Little	-3.00 (1.62)	-1.83 (1.99)	-1.36 (1.69)
Instructional material			
A lot	1.83 (3.76)	4.43 (3.71)	4.98 (4.9)
To some extent	1.32 (2.77)	2.96 (2.82)	1.74 (2.24)
Little	2.9 (2.52)	3.33 (2.54)	2.76 (2.2)
Computers			
A lot	1.07 (2.71)	0.14 (2.21)	-0.29 (2.68)
To some extent	6.51** (2.15)	6.30** (2.5)	6.18** (2.11)
Little	3.52** (1.9)	4.01** (1.89)	3.72** (1.87)
Library material			
A lot	-9.78*** (2.31)	-10.40*** (2.49)	-11.24*** (2.49)
To some extent	-1.71* (1.82)	-2.38* (1.52)	-3.48* (1.67)
Little	-2.36** (1.31)	-3.26** (1.31)	-4.72** (1.7)
Audio-visual resources			
A lot	-3.48* (2)	-5.45* (2.36)	-2.65* (1.77)
To some extent	-1.63 (1.84)	-4.05 (1.94)	-0.71 (1.22)
Little	0.34 (1.9)	-0.66 (1.28)	1.01 (1.49)
Assessments are used for			
Informing parents	6.98 (2.94)	4.45 (2.82)	7.07 (4.08)
Deciding about student's carreer	3.24 (2.38)	2.87 (2.79)	2.06 (2.94)
Grouping students	0.17 (2.81)	-0.15 (2.66)	-0.80 (2.52)
Comparing school to others	2.44 (1.96)	2.02 (2.47)	1.83 (2.23)
Monitoring school	1.02 (2.54)	-0.04 (2.45)	0.19 (2.11)
Judging teachers' effectiveness	-2.90 (1.92)	-3.17 (1.81)	-3.31 (2.11)

Table A4 (continued)

Dependent variable	RL	*full modell all* ML	SL
Autonomy, responsibility for			
Selecting teachers	1.93*** (1.75)	3.75*** (1.62)	0.91*** (1.69)
Firing teachers	0.06 (2.69)	-2.61 (2.63)	-0.64 (2.12)
Teachers' starting salary	-7.61** (1.94)	-5.19** (2.15)	-6.81** (1.53)
Teachers' salary increases	2.70 (3.03)	1.61 (2.65)	2.63 (2.35)
Formulating school budget	0.47 (1.41)	0.01 (1.59)	-0.18 (1.58)
Allocating school budget	5.96*** (2.31)	5.82*** (1.91)	7.70*** (2.85)
Disciplinary policies	8.75** (5.19)	7.89** (3.34)	10.76** (5.03)
Assessment policies	0.87 (2.54)	1.56 (2.98)	1.07 (3.14)
Student admission	-5.61** (2.43)	-5.48** (2.54)	-6.76** (2.74)
Textbooks	1.51 (2.62)	-1.25 (2.13)	2.13 (2.64)
Course content	-1.47 (1.87)	-0.15 (1.75)	-1.86 (1.97)
Course offers	2.91** (1.51)	3.74** (1.92)	3.60** (1.81)
Students are assessed by			
Standardized tests	-3.69 (2.45)	-3.27 (2.31)	-3.60 (2.47)
Teacher-developed tests	6.10 (4.33)	2.13 (4.18)	4.67 (6.58)
Teachers' judgemental ratings	1.71 (1.39)	1.42 (1.47)	0.97 (1.1)
Student <portfolios>	-2.22* (0.98)	-1.81* (1.18)	-1.94* (1.15)
Projects / Homework	3.24 (6.84)	6.92 (7.61)	6.28 (5.68)
Criteria for admission			
Residence			
Always	-1.98 (3.1)	-1.82 (2.83)	-1.63 (3.06)
Sometimes	-0.25 (1.44)	-0.76 (2.03)	-1.34 (1.54)
Student's record of performance			
Always	12.28*** (2.27)	11.83*** (2.11)	12.27*** (2.1)
Sometimes	2.95 (2.73)	3.03 (2.24)	3.07 (2.82)
Recommendation of feeder schools			
Always	-4.61 (3.28)	-4.92 (3.71)	-3.93 (3.88)
Sometimes	-1.13 (1.85)	-1.57 (2.52)	0.01 (2.34)
Parents‘ endorsement of philosophy			
Always	4.04 (2.03)	1.87 (1.81)	1.61 (2)
Sometimes	0.19 (1.29)	-0.48 (1.36)	-0.80 (1.41)
Student's interest in programme			
Always	-7.59*** (1.94)	-8.24*** (2.8)	-6.16*** (2.31)
Sometimes	-2.23 (1.59)	-2.23 (2.34)	-0.02 (2.1)
Family member of current student			
Always	-0.51 (2.45)	-1.53 (1.63)	-1.26 (2.34)
Sometimes	-0.62 (1.68)	-1.62 (1.94)	-1.61 (2.31)
Learning of students is hindered by			
Teachers' low expectations			
A lot	-11.14*** (4.41)	-12.05*** (3.95)	-12.57*** (5.79)
To some extent	-9.26** (3.17)	-9.51** (3.58)	-9.14** (3.47)
Little	-2.14 (1.91)	-1.01 (2.46)	-1.99 (2.1)

Table A4 (continued)

Dependent variable	RL	*full modell all* ML	SL
Student absenteeism			
A lot	-21.54*** (2.68)	-23.9*** (2.58)	-19.46*** (3.23)
To some extent	-16.87*** (2.17)	-19.27*** (2.52)	-17.27*** (2.34)
Little	-5.29*** (2.19)	-7.53*** (2.5)	-6.24*** (2.24)
Poor student-teacher relations			
A lot	7.21 (3.63)	2.93 (5.18)	5.49 (5.9)
To some extent	5.89** (1.89)	6.23** (2.18)	5.94** (2.25)
Little	2.00 (1.52)	1.77 (1.92)	2.64 (1.37)
Disruption of classes			
A lot	-19.86*** (5.57)	-16.71*** (5.59)	-18.17*** (4.94)
To some extent	-18.05*** (3.95)	-17.69*** (4.08)	-18.49*** (3.92)
Little	-8.96** (3.93)	-8.89** (3.84)	-9.48** (3.48)
Teachers not meeting students' needs			
A lot	14.77*** (4.24)	10.90*** (4.06)	14.38*** (4.3)
To some extent	9.44*** (2.13)	9.85*** (2.12)	8.55*** (2.35)
Little	7.94*** (2.34)	6.62*** (2.08)	6.38*** (2.13)
Teacher absenteeism			
A lot	3.09 (4.27)	4.05 (3.49)	5.34 (3.42)
To some extent	1.09 (2.59)	2.10 (2.08)	1.93 (2.63)
Little	0.96** (1.66)	2.71** (1.32)	2.30** (1.49)
Students skipping classes			
A lot	-12.70*** (5.16)	-12.07*** (4.11)	-13.85*** (3.88)
To some extent	-14.03*** (4.28)	-13.64*** (3.9)	-13.82*** (3.31)
Little	-8.24** (3.58)	-9.35** (3.95)	-8.98** (3.2)
Lack of respect			
A lot	-0.63 (2.21)	-1.16 (2.63)	-5.03 (2.72)
To some extent	-1.03 (2.89)	-2.15 (2.36)	-3.30 (2.79)
Little	-0.33 (1.85)	-0.69 (1.23)	-1.69 (1.36)
Staff resisting change			
A lot	-0.24 (3.66)	-1.12 (4.41)	-2.44 (3.81)
To some extent	4.49 (1.96)	4.43 (2.58)	6.44 (2.32)
Little	3.72 (1.59)	2.79 (1.76)	3.67 (1.9)
Students using drugs			
A lot	6.54 (5.76)	3.33 (4.41)	6.05 (5.9)
To some extent	-0.36 (4)	-0.78 (3.6)	-0.69 (3.92)
Little	-2.08* (1.71)	-2.03* (1.15)	-2.98* (1.66)
Teachers beeing to strict			
A lot	9.43 (3.74)	9.17 (4.99)	3.82 (3.76)
To some extent	5.77* (2.66)	7.37* (3.71)	5.27* (2.9)
Little	1.67 (1.51)	1.75 (2.11)	0.61 (1.79)
Students bullying students			
A lot	-4.03 (2.83)	0.86 (2.4)	-1.10 (2.13)
To some extent	-4.96* (4.07)	-4.94* (2.94)	-5.02* (3.78)
Little	0.45 (2.17)	1.41 (2.57)	1.82 (2.6)

Table A4 (continued)

Dependent variable	RL	*full modell all* ML	SL
Students not beeing encouraged			
A lot	4.42 (3.22)	5.51 (3.31)	7.70 (2.22)
To some extent	-1.74 (2.07)	-1.24 (1.96)	-1.49 (2.29)
Little	-1.46 (1.84)	-1.26 (1.91)	-0.76 (1.54)
Country-level variables			
Standardized tests	40.57 (26.08)	45.97 (29.76)	11.10 (33.26)
School incentives	-100.74** (47.88)	-88.26* (44.35)	-136.13** (65.31)
Standardized tests x school incentives	169.01*** (56.72)	159.98*** (55.82)	205.98*** (71.39)
Teacher incentives	-23.75 (36.11)	96.05*** (34.41)	51.52 (45.58)
Standardized tests x teacher incentives	-19.73 (39.82)	-167.77*** (42.77)	-108.43* (55.29)
GDP per capita	0.002 (0.001)	0.000 (0.001)	0.001 (0.001)
Share of public funding	0.04 (0.5)	-0.32 (0.54)	-1.40 (0.65)
Student-teacher ratio	-0.11* (0.38)	-0.48* (0.42)	-0.49* (0.43)
School size	-0.02** (0.01)	-0.05** (0.01)	-0.04** (0.02)
Criteria for admission			
Residence	13.07*** (10.95)	6.39*** (10.87)	4.87*** (12.19)
Student's record of performance	-25.32*** (20.45)	-56.63*** (20.51)	-37.87*** (23.44)
Recommendation of feeder schools	34.12** (7.59)	30.20** (8.75)	41.56** (8.66)
Parents endorsement of philosophy	-66.85*** (27.23)	-5.22*** (32.95)	-33.59*** (32.18)
Student's interest in program	-23.26** (24.71)	-4.42** (21.22)	-37.70** (33.27)
Family member of current student	-21.53*** (13.74)	-37.04*** (15.79)	-26.00*** (16.21)
Students are assessed by			
Teacher-developed tests	-52.96 (89.33)	-201.14 (109.61)	-192.47 (109.75)
Student <portfolios>	11.74 (13.77)	42.44 (13.65)	48.79 (17.91)
Projects / Homework	168.82** (108.94)	291.17** (134.62)	357.36** (135.16)
Assessments are used for			
Informing parents	438.00 (155.97)	688.77 (136.7)	387.80 (216.2)
Deciding about student's career	-164.27 (23.49)	-151.89 (24.72)	-90.04 (30.25)
Grouping students	-4.76 (19.42)	51.69 (19.55)	7.01 (24.65)
Autonomy, responsibility for			
Content	2.30 (19.96)	4.54 (21.5)	26.16 (20.93)
Staff	-45.77*** (20.87)	-14.70*** (22.55)	22.44*** (25.39)
Budget	35.23*** (22.96)	39.26*** (24.36)	29.21*** (25.8)
Shortage assessment of principal			
Staff	56.33*** (20.29)	87.22*** (22.37)	59.18*** (27.1)
Material	-9.85*** (13.66)	16.69*** (16.12)	13.32*** (16.56)
Observations	469,593	394,060	393,922
Countries	35	35	35

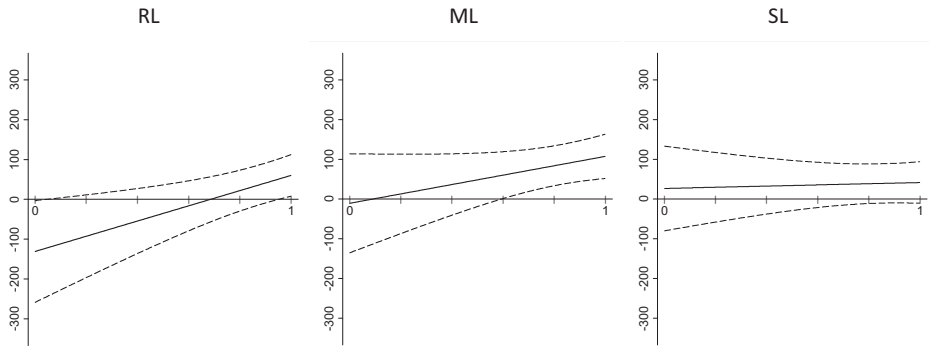
Notes: Impact of the key variables measured as country-level averages on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL). Estimates of *full model all* by least squares weighted with sampling probabilities. Controls include country fixed effects, time fixed effects, and controls for missing values. Robust standard errors clustered at country level in parentheses. Significance levels: * 1 %, ** 5 %, * 10 %.**

Table A5: Impact of key variables on PISA scores in regressions without interaction terms

	full model key			*full model all*		
Dependent variable	RL	ML	SL	RL	ML	SL
Standardized tests	4.95	-59.15	-16.61	64.03***	9.77	6.55
	(21.25)	(45.03)	(25.03)	(19.95)	(22.34)	(23.80)
School incentives	31.34	148.19***	84.45**	43.45	64.95**	49.27
	(31.35)	(47.07)	(34.39)	(29.28)	(30.31)	(34.23)
Teacher incentives	-33.26**	-49.06***	-63.11***	-37.72***	-23.21	-25.64
	(12.69)	(17.63)	(17.71)	(13.60)	(15.10)	(15.97)
Observations	489,583	409,884	409,736	469,593	394,060	393,922
Countries	37	37	37	35	35	35

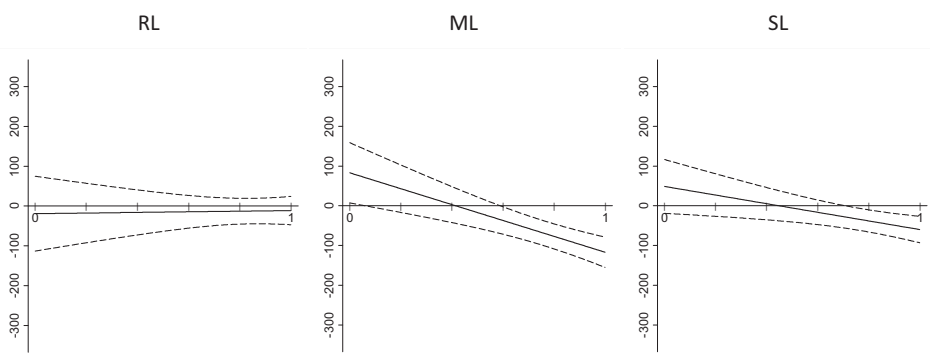
Notes: Impact of the key variables measured as country-level averages on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL). Estimates of *full model key* and *full model all* by least squares weighted with sampling probabilities. Controls include country fixed effects, time fixed effects, controls for missing values, GDP per capita, school level variables (e.g. size, location, staff, funding, autonomy, admission rules), and student level variables (e.g. gender, age, education of parents, migration background). Robust standard errors clustered at country level in parentheses. Significance levels: * 1 %, ** 5 %, * 10 %.**

Figure A1: Marginal effects of school incentives as a function of standardized testing, subsample comprising the top 50% of all students



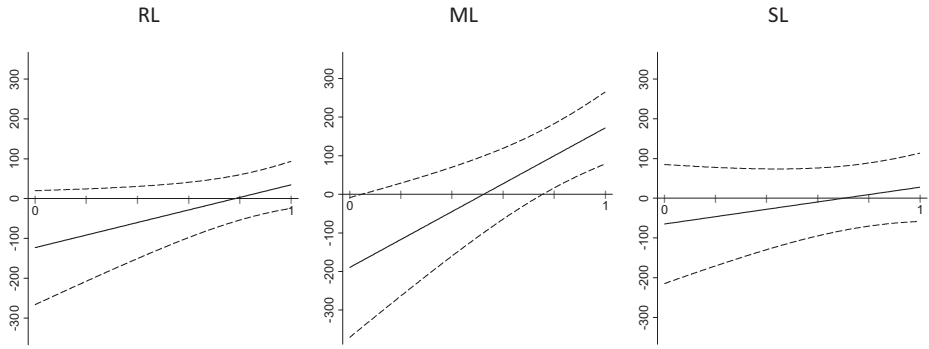
Notes: Marginal effects of school incentives (solid line ———) on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL) as a function of standardized testing. Results from a subsample comprising the top 50 % of all students in the respective literacy. The dashed lines (– – –) display 95 % confidence intervals.

Figure A2: Marginal effects of teacher incentives as a function of standardized testing, subsample comprising the top 50% of all students



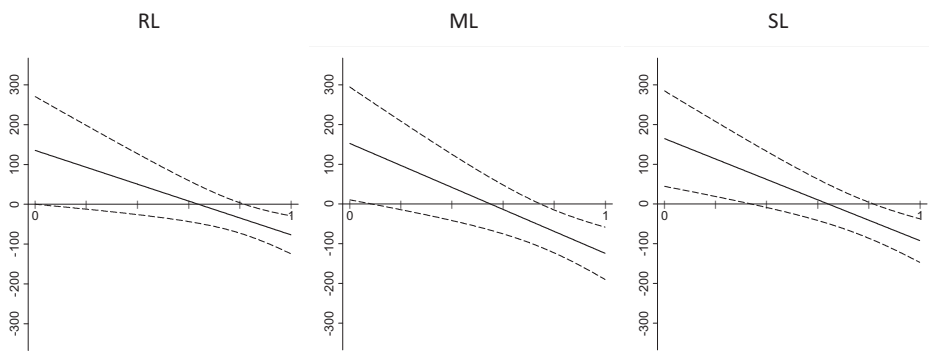
Notes: Marginal effects of teacher incentives (solid line ———) on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL) as a function of standardized testing. Results from a subsample comprising the top 50 % of all students in the respective literacy. The dashed lines (– – –) display 95 % confidence intervals.

Figure A3: Marginal effects of school incentives as a function of standardized testing, subsample comprising the bottom 50% of all students



Notes: Marginal effects of school incentives (solid line ———) on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL) as a function of standardized testing. Results from a subsample comprising the bottom 50 % of all students in the respective literacy. The dashed lines (– – –) display 95 % confidence intervals.

Figure A4: Marginal effects of teacher incentives as a function of standardized testing, subsample comprising the bottom 50% of all students



Notes: Marginal effects of teacher incentives (solid line ———) on reading literacy (RL), mathematical literacy (ML), and scientific literacy (SL) as a function of standardized testing. Results from a subsample comprising the bottom 50 % of all students in the respective literacy. The dashed lines (– – –) display 95 % confidence intervals.