Empirical Economics (2007) 32:433–464
DOI 10.1007/s00181-006-0087-0

ORIGINAL PAPER

What accounts for international differences in student performance? A re-examination using PISA data

Thomas Fuchs · **Ludger Wößmann**

Revised: 17 July 2006 / Published online: 22 August 2006

Abstract We use the PISA student-level achievement database to estimate international education production functions. Student characteristics, family backgrounds, home inputs, resources, teachers and institutions are all significantly associated with math, science and reading achievement. Our models account for more than 85% of the between-country performance variation, with roughly 25% accruing to institutional variation. Student performance is higher with external exams and budget formulation, but also with school autonomy in textbook choice, hiring teachers and within-school budget allocations. Autonomy is more positively associated with performance in systems that have external exit exams. Students perform better in privately operated schools, but private funding is not decisive.

Keywords Education production function · PISA · International variation in student performance · Institutional effects in schooling

JEL classification I28 · J24 · H52 · L33

T. Fuchs (✉)
Ifo Institute for Economic Research,
Poschingerstr. 5, 81679 Munich, Germany
e-mail: fuchs@ifo.de
URL: www.cesifo.de/link/fuchs_t.htm

L. Wößmann
Ifo Institute for Economic Research,
University of Munich and CESifo,
Poschingerstr. 5, 81679 Munich, Germany
e-mail: woessmann@ifo.de
URL: www.cesifo.de/link/woessmann_l.htm

1 Introduction

The results of the Programme for International Student Assessment (PISA), conducted in 2000 by the Organisation for Economic Co-operation and Development (OECD), triggered a vigorous public debate on the quality of education systems in most participating countries. PISA made headlines on the front pages of tabloids and more serious newspapers alike. For example, *The Times* (December 6, 2001) in England titled, “Are we not such dunces after all?”, and *Le Monde* (December 5, 2001) in France titled, “France, the mediocre student of the OECD class”. In Germany, the PISA results hit the front pages of all leading newspapers for several weeks (e.g., “Abysmal marks for German students” in the *Frankfurter Allgemeine Zeitung*, December 4, 2001), putting education policy at the forefront of attention ever since. “PISA” is now a catch-phrase, known by every German, for the poor state of the German education system. While this coverage proves the immense public interest, the quality of much of the underlying analysis is less clear. Often, public assessments tend to simply repeat long-held believes, rather than being based on evidence produced by the PISA study. If based on PISA facts, they usually rest on bilateral comparisons between two countries, e.g., comparing a commentator’s home country to the top performer (Finland in the case of PISA reading literacy). And more often than not, they are bivariate, presenting the simple correlation between student performance and a single potential determinant, such as educational spending.

Economic theory suggests that one important set of determinants of educational performance are the institutions of the education system, because they set the incentives for the actors in the education process. Among the institutions that have been theorized to impact on the quality of education are public versus private financing and provision (e.g., Epple and Romano 1998; Nechyba 2000), centralization of financing (e.g., Hoxby 1999, 2001; Nechyba 2003), external versus teacher-based standards and examinations (e.g., Costrell 1994; Betts 1998; Bishop and Wößmann 2004), centralization versus school autonomy in curricular, budgetary and personnel decisions (e.g., Bishop and Wößmann 2004) and performance-based incentive contracts (e.g., Hanushek et al. 1994). In many countries, the impact that such institutions may have on student performance tends to be ignored in most discussions of education policy, which often focus on the implicitly assumed positive link between schooling resources and learning outcomes.

One reason for this neglect may be that the lack of institutional variation within most education systems makes an empirical identification of the impact of institutions impossible when using national datasets, as is standard practice in most empirical research on educational production (cf., Hanushek 2002 and the references therein). However, such institutional variation is given in cross-country data, and evidence based on previous international student achievement tests such as IAEP (Bishop 1997), TIMSS (Bishop 1997; Wößmann 2003a) and TIMSS-Repeat (Wößmann 2003b) supports the view that institutions play a key role in determining student outcomes. These international databases allow for multi-country multivariate analyses, which ensure that the impact of each determinant is estimated for observationally similar schools by holding the effects of other determinants constant.

In this paper, we use the PISA database to test the robustness of the findings of these previous studies of international education production functions.[1] Combining the performance data with background information from student and school questionnaires, we estimate the association between student background, schooling resources and schooling institutions on the one hand and international variations in students' educational performance on the other hand. In contrast to Bishop's (1997, 2006) country-level analyses, we perform the analyses at the level of the individual student, which allows us to take advantage of within-country variation in addition to between-country variation, vastly increasing the degrees of freedom of the analyses, at least to the extent that there is independence of error terms within countries.

Given its particular features, the rich PISA student-level database allows for a rigorous assessment of the determinants of international differences in student performance in general, and of the link between schooling institutions and student performance in particular. PISA offers the possibility to re-examine the validity of results of previous international studies in the context of a different subject (reading in addition to math and science), a different definition of required capabilities and of the target population, and to extend the examination by including more detailed family-background and institutional data. Among others, the PISA database distinguishes itself from previous international tests by providing data on parental occupation at the level of the individual student and on private versus public operation and funding at the level of the individual school.

Our results show that the PISA evidence underscores the importance of institutional features for international differences in student performance. Most notably, there are important interaction effects between external exit exams and several measures of school autonomy, with the association between school autonomy and student performance becoming more positive in school systems with external exit exams.

The remainder of the paper is structured as follows. Section 2 describes the database of the PISA international student performance study and compares its features to previous studies. Section 3 discusses the econometric model. Section 4 presents the empirical results. Section 5 reports a set of robustness

[1] Some economic research based on PISA data exists, but it is mostly on a national scale. Fertig (2003a) uses the German PISA sample to analyze determinants of German students' achievement. Fertig (2003b) uses the US PISA sample to analyze class composition and peer group effects. Wolter and Coradg Vellacott (2003) use PISA data to study sibling rivalry in Belgium, Canada, Finland, France, Germany and Switzerland. To our knowledge, the only previous study using PISA data to estimate multivariate education production functions internationally is Fertig and Schmidt (2002), who, sticking to reading performance, do not focus on estimating determinants of international performance variation but rather on estimating conditional national performance scores. Recently, Fuchs and Wößmann (2004b) analyze the association between computers and PISA performance in detail.

checks. Section 6 analyzes the explanatory power of the model and its different parts at the country level. Section 7 summarizes and concludes.

2 The PISA international student performance study

2.1 PISA and previous international student achievement tests

The international dataset used is the OECD Programme for International Student Assessment (PISA). In addition to testing the robustness of findings derived from previous international student achievement tests, the PISA-based analyses contribute several additional new aspects to the literature. First, PISA tested a new subject, namely reading literacy, in addition to math and science already tested in IAEP and TIMSS. This alternative measure of performance broadens the outcome of the education process considered in the analyses.

Second, particularly in reading, but also in the more traditional domains of math and science, “PISA aims to define each domain not merely in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in adult life” (OECD 2000, p. 8). That is, rather than being curriculum-based as the previous studies, “PISA looked at young people’s ability to use their knowledge and skills in order to meet real-life challenges” (OECD 2001, p. 16). For example, reading literacy is defined in terms of “the capacity to understand, use and reflect on written texts, in order to achieve one’s goals, to develop one’s knowledge and potential, and to participate in society” (OECD 2000, p. 10).[2] There is a similar real-life focus in the other two subjects. On the one hand, this focus should constitute the most important outcome of the education process, but on the other hand it bears the caveat that schools are assessed not on the basis of what their school system requires them to teach, but rather on what students might need particularly well for coping with everyday life.

Third, rather than targeting students in specific grades as in previous studies, PISA’s target population are 15-year-old students in each country, regardless of the grade they currently attend. This target population not only assesses young people near the end of compulsory schooling, but also captures students of the very same age in each country independent of the structure of national school systems. By contrast, the somewhat artificial grade-related focus of other studies may be distorted by differing entry ages and grade-repetition rules.

Fourth, the PISA data provide more detailed information than previous studies on some institutional features of the school systems. For example, PISA provides data on whether schools are publicly or privately operated, on which share of their funding stems from public or private sources and on whether schools can fire their teachers. These background data provide improved internationally comparable measures of schooling institutions.

[2] See OECD (2000) for further details on the PISA literacy measures, as well as for sample questions.

Fifth, the PISA data also provide more detailed information on students' family background. For instance, there is information on parental occupation and the availability of computers at home. This should contribute to a more robust assessment of potential determinants of student performance. Finally, reading literacy is likely to depend more strongly on family-background variables than performance in math and science. Hence controlling for a rich set of family-background variables should establish a more robust test of the institutions-performance link when reading performance is the dependent variable.

Taken together, the PISA international dataset allows for a re-examination of results based on previous international tests using an additional subject, real-life rather than curriculum-based capabilities, an age-based target population and richer data particularly on family background and institutional features of the school system.

2.2 The PISA database

The PISA study was conducted in 2000 in 32 developed and emerging countries, 28 of which are OECD countries, in order to obtain an internationally comparable database on the educational achievement of 15-year-old students in reading, math and science. The study was organized and conducted by the OECD, ensuring as much comparability among participants as possible and a consistent and coherent study design.[3] Table 1 reports the countries participating in the PISA 2000 study.[4]

As described above, PISA's target population were the 15-year-old students in each country. More specifically, PISA sampled students aged from 15 years and 3 months to 16 years and 2 months at the beginning of the assessment period. The students had to be enrolled in an educational institution, regardless of the grade level or type of institution. The average age of OECD-country students participating in PISA was 15 years and 8 months, varying by a maximum of only 2 months among the participating countries.

The PISA sampling procedure ensured that a representative sample of the target population was tested in each country. Most PISA countries employed a two-stage sampling technique. The first stage drew a (usually stratified) random sample of schools in which 15-year-old students were enrolled, yielding a minimum sample of 150 schools per country. The second stage randomly sampled 35 of the 15-year-old students in each of these schools, with each 15-year-old student in a school having equal selection probability. This sampling procedure typically led to a sample of between 4,500 and 10,000 tested students in each country.

[3] For detailed information on the PISA study and its database, see OECD (2000, 2001, 2002), Adams and Wu (2002) and the PISA homepage at http://www.pisa.oecd.org.

[4] Liechtenstein was not included in our analysis due to lack of internationally comparable country-level data, e.g. on educational expenditure per student. Note that there were only 326 15-year-old students in Liechtenstein in total, 314 of whom participated in PISA.

Table 1 Descriptive country statistics of selected variables

	Test score			ISEI	External Exit exam[a]	School autonomy in			Publicly managed school	Governm. funding
	Math	Science	Reading			Deter-mining course content	Establish. teachers' starting salaries	Choosing textbooks		
Australia	530.1	523.7	526.9	45.5	0.8	92.1	64.2	100.0	NA	69.5
Austria	513.0	517.2	508.0	45.6	0.0	70.0	42.0	100.0	87.3	90.4
Belgium	516.4	495.9	509.8	44.8	0.0	72.1	97.7	99.6	26.0	87.5
Brazil	362.8	386.0	403.4	39.9	0.0	93.7	42.4	99.8	86.8	78.0
Canada	528.3	526.3	532.4	45.5	0.5	59.0	86.6	93.1	NA	90.0
Czech Republic	495.7	509.8	494.9	42.7	1.0	88.7	97.5	100.0	93.9	94.9
Denmark	513.1	483.7	495.9	44.2	1.0	97.0	97.6	100.0	75.3	94.2
Finland	532.3	533.4	544.1	44.5	1.0	99.3	61.4	100.0	97.2	99.8
France	514.5	499.7	506.1	43.7	1.0	60.2	21.9	100.0	78.4	75.5
Germany	488.0	488.5	487.8	44.0	0.4	62.8	25.8	100.0	95.6	97.3
Greece	453.6	465.3	474.6	43.4	0.0	97.3	85.4	96.3	92.8	83.7
Hungary	486.5	495.2	484.3	42.1	1.0	99.6	100.0	100.0	94.7	87.3
Iceland	513.0	495.9	505.6	46.5	1.0	88.4	100.0	99.3	99.2	99.4
Ireland	501.2	512.6	524.3	42.3	1.0	49.1	88.6	100.0	39.1	91.1
Italy	460.2	479.6	489.2	43.5	1.0	93.1	10.3	100.0	94.2	75.2
Japan	553.0	543.6	521.2	43.8	1.0	99.3	34.5	99.3	69.8	72.5
Korea	542.0	546.1	523.0	41.4	1.0	99.4	32.3	99.4	47.7	49.1
Latvia	469.4	463.9	460.7	40.8	0.5	75.9	100.0	70.4	99.3	95.6
Luxembourg	452.2	450.6	447.5	42.3	1.0	0.0	0.0	0.0	87.9	100.0
Mexico	400.8	431.0	422.7	39.8	0.0	61.4	60.5	83.2	84.6	37.2
Netherlands	556.0	527.4	530.1	46.6	1.0	97.5	100.0	100.0	26.3	94.7
New Zealand	532.9	524.3	527.6	45.1	1.0	99.1	100.0	100.0	95.4	80.2
Norway	498.7	500.0	502.2	48.5	1.0	NA	NA	NA	98.5	99.5
Poland	475.6	484.9	481.1	40.1	1.0	NA	NA	NA	97.1	92.3

Table 1 continued

	Test score					School autonomy in				
	Math	Science	Reading	ISEI	External Exit exam[a]	Deter-mining course content	Establish. teachers' starting salaries	Choosing textbooks	Publicly managed school	Governm. funding
Portugal	457.7	463.0	472.2	41.0	0.0	26.5	21.1	100.0	92.7	87.9
Russian Fed.	479.3	463.8	463.3	41.3	1.0	94.6	99.6	97.4	100.0	93.5
Spain	477.1	492.1	493.1	42.4	0.0	89.4	37.7	99.6	60.7	82.9
Sweden	509.2	510.3	514.5	45.2	0.5	95.1	100.0	100.0	96.6	99.9
Switzerland	526.4	496.2	495.4	45.3	0.0	52.9	95.7	75.1	93.6	93.8
United Kingdom	526.6	528.3	522.1	46.1	1.0	96.4	99.7	100.0	91.4	89.8
United States	494.5	500.4	503.2	46.1	0.1	86.3	98.9	95.3	93.4	91.6
Total Average	496.1	494.4	495.3	43.7	0.6	78.0	68.5	93.5	82.9	86.7

Notes: Country means, based on non-imputed data for each variable, weighted by sampling probabilities. *ISEI* international socio-economic index. The institutional variables are reported as shares within each country (in percent). *NA* not available. [a]In math.

Table 2 Variance decomposition

	Math	Science	Reading
Variance between students	100.0	100.0	100.0
Variance within schools	56.2	63.0	59.7
Variance between schools	43.8	37.0	40.3
Variance between countries	16.1	10.7	9.6

Note: Share of total variance in test scores occurring between (resp. within) the respective group (in percent)

The performance tests were paper and pencil tests, lasting a total of two hours for each student. Test items included both multiple-choice items and questions requiring the students to construct their own responses. The PISA tests were constructed to test a range of relevant skills and competencies that reflected how well young adults are prepared to analyze, reason and communicate their ideas effectively. Each subject was tested using a broad sample of tasks with differing levels of difficulty to represent a coherent and comprehensive indicator of the continuum of students' abilities. Using item response theory, PISA mapped performance in each subject on a scale with an international mean of 500 and a standard deviation of 100 test-score points across the OECD countries. The main focus of the PISA 2000 study was on reading, with two-thirds of the testing time devoted to this subject. In the other two subjects, smaller samples of students were tested. The correlation of student performance between the subjects is substantial, at 0.700 between reading and math (96,913 joint observations), 0.718 between reading and science (96,815) and 0.639 between math and science (39,079).

To give an idea of the international structure of the data, Table 1 reports country means of test-score performance in the three subjects. Mean performance on the reading literacy test ranges from 403.4 test-score points in Brazil to 544.1 in Finland. Table 2 reports results of a decomposition of the total international variance of student-level test scores into components within schools, between schools and between countries. In each subject, more than half of the total variance occurs within individual schools. The variance component that occurs between countries ranges from 9.6% in reading to 16.1% in math.

In addition to the performance tests, students as well as school principals answered respective background questionnaires, yielding rich background information on students' personal characteristics and family backgrounds as well as on schools' resource endowments and institutional settings. Combining the available data, we constructed a dataset containing 174,227 students in 31 countries tested in reading literacy. In math, the sample size is 96,855 students, and 96,758 students in science. In our estimations, we drop students in extremely outlying grades, namely grades 6 or lower and grades 12 or higher, which reduces the samples by between 342 and 609 students in the three subjects. The dataset combines student test scores in reading, math and science with students' characteristics, family-background data and school-related variables of resource use and institutional settings.[5] For estimation purposes, a variety of qualitative variables were transformed into dummy variables.

Table 1 contains country means of selected country characteristics. Table 3 reports international descriptive statistics for all the variables employed in this paper. It also includes information on the amount of original versus missing data for each variable. To be able to use a complete dataset of all students with performance data and at least some background data, we imputed missing values using the method described in the Appendix. Given the large set of explanatory variables considered and given that each variable is missing for some students, dropping all student observations that have a missing value on at least one variable would mean a severe reduction in sample size. While the percentage of missing values of each individual variable ranges from 0.9 to 33.3% (cf. Table 3), the percentage of students with a missing value on least one variable is 72.6% in reading. That is, the sample size in reading would be as small as 47,782 students from 20 countries (26,228 students from 19 countries in math and 24,049 students from 19 countries in science). Apart from the general reduction in sample size, dropping all students with a missing value on at least one variable would delete information available on other explanatory variables for these students and introduce bias if values are not missing at random. Thus, data imputation is the only viable way of performing the broad-based analyses. As described in Sect. 3.1 below, the estimations we employ ensure that the effects estimated for each variable are not driven by imputed values.

In addition to the rich PISA data at the student and school level, we also use country-level data on countries' GDP per capita in 2000 (measured in purchasing power parities (PPP), World Bank 2003), average educational expenditure per student in secondary education in 2000 (measured in PPP, OECD 2003)[6] and the existence of curriculum-based external exit exams (in their majority kindly provided by John Bishop; cf. Bishop 2006).

3 Econometric analysis of the education production function

3.1 Estimation equation, covariance structure and sampling weights

The microeconometric estimation equation of the education production function has the following form:

[5] We do not use data on teaching methods, teaching climate or teacher motivation as explanatory variables, because we view these mainly as outcomes of the education system. First, such measures are endogenous to the institutional surrounding of the education system. This institutional surrounding sets the incentives to use specific methods and creates a specific climate, thereby constituting the deeper cause of such factors. Second, such measures may be as much the outcome of students' performance as their cause, so that they would constitute left-hand-side rather than right-hand-side variables.

[6] For the three countries with missing data in OECD (2003), we use comparable data for these countries from World Bank (2003) and data from both sources for countries where both are available to predict the missing data for the three countries by ordinary least squares.

Table 3 International descriptive statistics

	Mean	SD	Source	Imputed
Test scores				
Math	496.1	102.6	St	0.0
Science	494.3	102.1	St	0.0
Reading	495.4	101.3	St	0.0
Institutions				
Testing				
External exit exam	0.578		C	0.0
Standardized tests	0.602		Sc	12.2
School autonomy				
Determining course content	0.780		Sc	9.6
Establishing teachers' starting salaries	0.265		Sc	8.0
Choosing textbooks	0.935		Sc	9.1
Deciding on budget allocations within school	0.946		Sc	8.9
Formulating school budget	0.762		Sc	9.3
Hiring teachers	0.685		Sc	7.9
Public vs. private operation and funding				
Publicly managed school	0.829		Sc	22.1
Government funding (share)	0.867	0.237	Sc	5.9
Resources and teachers				
Educational expenditure per student (1,000$)	5.664	2.627	C	0.0
Class size				
Math	23.6	8.3	St	8.7
Science	22.8	9.1	St	11.2
Reading	24.6	8.4	St	5.2
Student-teacher ratio	13.7	6.7	Sc	24.6
Instructional material				
Not at all lacking	0.502		Sc	3.8
Strongly lacking	0.049		Sc	3.8
Instruction time (1,000 minutes per year)				
Math	7.346	2.921	Sc	5.0
Science	7.473	4.319	Sc	5.0
Reading	7.558	3.183	Sc	5.0
Teacher education (share at school)				
Masters in pedagogy	0.616	0.392	Sc	29.9
Teacher certificate	0.844	0.266	Sc	32.7
Masters in math	0.734	0.338	Sc	31.4
Masters in science	0.789	0.319	Sc	32.6
Masters in language	0.774	0.315	Sc	33.3
Student characteristics				
Grade				
6th or lower	0.001		St	1.4
7th	0.012		St	1.4
8th	0.059		St	1.4
9th	0.388		St	1.4
10th	0.470		St	1.4
11th	0.069		St	1.4
12th or higher	0.002		St	1.4
Country's school entry age	6.163	0.572	C	0.0
Age (months)	188.5	3.4	St	1.0
Female	0.501		St	0.9

Table 3 continued

	Mean	SD	Source	Imputed
Family background				
Born in country				
Student	0.927		St	4.5
Mother	0.864		St	4.7
Father	0.863		St	5.6
Living with				
No parent	0.011		St	1.8
Single father	0.021		St	1.8
Single mother	0.132		St	1.8
Both parents	0.836		St	1.8
Parents' education				
None	0.011		St	6.8
Primary	0.075		St	6.8
Lower secondary	0.137		St	6.8
Upper secondary 1	0.149		St	6.8
Upper secondary 2	0.245		St	6.8
University	0.383		St	6.8
Parents' work status				
None working	0.066		St	1.9
At least one half-time	0.065		St	1.9
At least one full-time	0.492		St	1.9
Both full-time	0.378		St	1.9
Parents' job				
Blue collar	0.098		St	4.2
White collar	0.522		St	4.2
Books at home				
None	0.018		St	2.8
1–10 books	0.090		St	2.8
11–50 books	0.199		St	2.8
51–100 books	0.210		St	2.8
101–250 books	0.212		St	2.8
251–500 books	0.155		St	2.8
More than 500 books	0.117		St	2.8
International socio-economic index (ISEI)	43.807	16.712	St	6.5
School's community location				
Village or rural area (<3,000)	0.111		Sc	21.8
Small town (3,000–15,000)	0.236		Sc	21.8
Town (15,000–100,000)	0.314		Sc	21.8
City (100,000–1,000,000)	0.209		Sc	21.8
City center of city with > 1 million people	0.063		Sc	21.8
Elsewhere in city with > 1 million people	0.066		Sc	21.8
GDP per capita (1,000)	22.050	9.504	C	0.0
Home incentives and inputs				
Parental support				
Strongly lacking	0.189		Sc	3.4
Not at all lacking	0.065		Sc	3.4
Homework				
Math: <1 hour per week	0.418		St	2.9
Math: >1 and <3 hours per week	0.401		St	2.9
Math: >3 hours per week	0.182		St	2.9
Science: <1 hour per week	0.500		St	4.6
Science: >1 and <3 hours per week	0.337		St	4.6
Science: >3 hours per week	0.163		St	4.6

Table 3 continued

	Mean	SD	Source	Imputed
Reading: <1 hour per week	0.478		St	2.6
Reading: >1 and <3 hours per week	0.397		St	2.6
Reading: >3 hours per week	0.126		St	2.6
Computers at home				
None	0.221		St	2.7
One	0.250		St	2.7
More than one	0.529		St	2.7

Notes: Mean: International mean, based on non-imputed data for each variable, weighted by sampling probabilities. SD: International standard deviation (only for metric discrete variables). Source: Data source and thus level of observation: St = student achievement test or student background questionnaire; Sc = school background questionnaire; C = country-level variable (see text for specific sources). Imputed: Fraction of students with missing and thus imputed data (in percent), weighted by sampling probabilities

$$T_{is} = B_{is}\beta_1 + R_{is}\beta_2 + I_s\beta_3 + D^B_{is}\beta_4 + \left(D^B_{is}B_{is}\right)\beta_5 + D^R_{is}\beta_6 + \left(D^R_{is}R_{is}\right)\beta_7 + D^I_s\beta_8 + \left(D^I_sI_s\right)\beta_9 + \varepsilon_{is}, \tag{1}$$

where T_{is} is the achievement test score of student i in school s. B is a vector of student background data (including student characteristics, family background and home inputs), R is a vector of data on schools' resource endowment and I is a vector of institutional characteristics. Because of a particular interest in interactions between external exit exams and other institutional features, I will include institutional interaction terms. The parameter vectors $\beta_{\{1\}}$to $\beta_{\{9\}}$ will be estimated in the regression. Note that with the exception of the institutional interaction terms, this specification of the international education production function restricts each effect to be the same in all countries, as well as at all levels (within schools, between schools and between countries). While it might be interesting to analyze the potential heterogeneity of certain effects between countries and between levels, regarding the object of interest of this paper it seems warranted to abstain from this effect heterogeneity and estimate a single effect for each variable.[7]

As discussed in the previous section, some of the data are imputed rather than original. Our imputation method is based on conditional mean imputation (cf. Little and Rubin 1987), which predicts the conditional mean for each missing

[7] Wößmann (2003a) compares this restricted specification to an alternative two-step specification, discussing advantages and drawbacks particularly in light of potential omitted country-level variables, and favoring the specification employed here. The first, student-level step of the alternative specification includes country fixed effects in the estimation of (1). These country fixed effects are then regressed in a second, country-level step on averages at the country level of relevant explanatory variables. Wößmann (2003a) finds that the substantive results of the two specifications are virtually the same. Furthermore, our results presented in Sect. 6 below show that our restricted model can account for more than 85% of the between-country variation in test scores in each subject. Therefore, the scope for obvious unobserved country-specific heterogeneity, and thus the need for the country-fixed-effects specification, seems small.

observation on the independent variables using non-missing values of the specific variables and a set of independent variables observed for all students (see Appendix). Schafer and Schenker (2000) show that conditional mean imputation combined with appropriately corrected standard errors yields an unbiased and efficient estimator which outperforms the multiple stochastic imputation estimator (Rubin 1987). Because imputed values are estimated quantities, the statistical inference has to take account of the uncertainty involved in imputation. The required correction procedure for the standard errors employed in this paper accounts for the degree of variability and uncertainty in the imputation process and for the share of missing data (cf. the discussion of the Schafer and Schenker 2000 procedure in the Appendix).

If values are not missing conditionally at random, estimates could still be biased. For example, if among observationally similar students the probability of a missing value for a variable depends on an unobserved student characteristic that also influences achievement, imputation would predict the same value of the variable for students with a missing value that was observed for the other students, which would result in biased coefficient estimates. To account for this possibility of non-randomly missing observations and to make sure that the results are not driven by imputed data, three vectors of dummy variables D^B, D^R and D^I are included as controls in the estimation. The D vectors contain one dummy for each variable in the three vectors B, R and I that takes the value of 1 for observations with missing and thus imputed data and 0 for observations with original data. The D vectors allow the observations with missing data on each variable to have their own intercepts. The interaction terms between imputation dummies and data vectors, $D^B B$, $D^R R$ and $D^I I$, allow them to also have their own slopes. These imputation controls for each variable with missing values ensure that the results are reasonably robust against possible bias arising from data imputation.

Owing to the complex data structure produced by the PISA survey design and the multi-level nature of the explanatory variables, the error term ε of the regression has a non-trivial structure. Given the possible dependence of students within the same school, the use of school-level variables and the fact that schools were the primary sampling unit (PSU) in PISA (see Sect. 2.2), there may be unobservable correlation among the individual error terms ε_i at the school level s (cf. Moulton 1986). Therefore, we report standard errors that are clustered at the school level, lifting the classical independence assumption to the level of schools.

Finally, PISA used a stratified sampling design within each country, producing varying sampling probabilities for different students. To obtain nationally representative estimates from the stratified survey data at the within-country level, we employ weighted least squares (WLS) estimation using sampling probabilities as weights. WLS estimation ensures that the proportional contribution to the parameter estimates of each stratum in the sample is the same as would have been obtained in a complete census enumeration (DuMouchel and Duncan 1983; Wooldridge 2001). Furthermore, at the between-country level, we weight each of the 31 countries equally.

3.2 Cross-sectional data and potential endogeneity

The econometric estimation of the PISA dataset is restricted by its cross-sectional nature, which does not allow for panel or value-added estimations (cf., e.g., Hanushek 2002; Todd and Wolpin 2003). Because of unobserved student abilities, cross-sectional analyses can give rise to omitted variable bias when the variables of interest are correlated with the unobserved characteristics. In this paper, we hope to minimize such biases due to unobserved student heterogeneity by including a huge set of observed abilities, characteristics and institutions which reduce potential biases. Estimates based on cross-sectional data will be unbiased under the conditions that the explanatory variables of interest are unrelated to features that still remain unobserved, that they are exogenous to the dependent variable and that they and their impact on the dependent variable do not vary over time.

It seems straightforward that the student-specific family background B_{is} is exogenous, in a narrow sense, to the students' educational performance. Furthermore, most aspects of the family background B_{is} are time-invariant, so that the characteristics observed at the given point in time of the PISA survey should be consistent indicators for family characteristics in the past. Therefore, student-related family background, as well as other student characteristics like area of residence, affect not only the educational value-added in the year of testing but rather educational performance throughout a student's entire school life. A level-estimation approach thus seems well-suited for determining the total association between family background and student-related characteristics on the one hand and students' achievements on the other hand. However, family background may be correlated with unobserved ability, which again may be correlated across generations. Therefore, a narrow causal interpretation of family-background effects is not possible. In the presented analyses, the large amount of family-background indicators mainly serves as control variables.

Many of the institutional features I_s of an education system may also be reasonably assumed to be exogenous to individual students' performance. The cross-country nature of the data allows the systematic utilization of country differences in institutional settings of the educational systems, which would be neglected in within-country specifications. However, a caveat applies here in that a country's institutions may be related to unobserved, e.g. cultural, factors which in turn may be related to student performance. To the extent that this may be an important issue, caution should prevail in drawing causal inferences and policy conclusions from the presented results.

In terms of time variability, changes in institutions generally occur only gradually and evolutionary rather than radically, particularly in democratic societies. Consequently, the institutional structures of education systems are highly time-invariant and thus most likely constant, or at least rather similar, during a student's life in secondary school. We therefore assume that the educational institutions observed at one point in time persist unchanged during the students' secondary-school life and thus contribute to students' achievement levels, and not only to the change from one grade to the next. Still, institutional structures may differ between primary and secondary school, so that issues of omitted prior inputs in a students' life may still bias estimated institutional effects, generally in an attenuating way.

The situation is more problematic for schools' resource endowments R_{is}. For example, educational expenditure per student has been shown to vary considerably over time (cf. Gundlach et al. 2001). Still, as far as the cross-country variation in educational expenditure is concerned, the assumption of relatively constant relative expenditure levels seems not too implausible, so that country-level values of expenditure per student in the year of the PISA survey may yield reasonable proxies for expenditure per student over students' school life.

However, students' educational resource endowments are not necessarily exogenous to their educational performance. Resource endogeneity in the narrow sense should not be a serious issue at the country level, due to the lack of a supranational government body that would redistribute educational expenditures according to students' achievement and due to international mobility constraints. But within countries, endogenous resource allocations, both between and within schools, may bias least-squares estimates of the effects of resources on student performance. To avoid biases from within-school sorting of resources according to the needs and achievement of students, Akerhielm (1995) suggests an IV estimation approach that uses school-level variables as instruments for class size. Accordingly, in our regressions we use the student-teacher ratio at the school level as an instrument for the actual class size in each subject in a two-stage least squares (2SLS) estimation.[8] However, this approach may still be subject to between-school sorting of differently achieving students based on schools' resource endowments, e.g. caused by school-related settlement decisions of parents (cf. West and Wößmann 2006). To the extent that between-school sorting is unrelated to the family-background and institutional characteristics for which our regressions control, it might still bias estimated resource effects (cf. Wößmann and West 2006; Wößmann 2005). Furthermore, variation in individual students' resource endowments over time, e.g. class-size variation, may also bias levels-based estimates, generally resulting in a downward attenuation bias. The PISA data do not allow for overcoming these possibly remaining biases.

4 Estimation results

This section discusses the results of estimating (1) for the three subjects. The results are reported in Table 4. The discussion of results only briefly refers to selected control variables (see Fuchs and Wößmann 2004a for a more extensive discussion) and then focuses on the effects of institutional features.

[8] Note that this approach also accounts for measurement-error biases in the class-size variable.

Table 4 International education production functions

	Math		Science		Reading	
	Coef.	*SE*	Coef.	*SE*	Coef.	*SE*
Institutions						
Testing						
External exit exam (EEE)	−72.766**	28.517[a]	−79.663**	30.126[a]	−90.205***	30.350[a]
Standardized tests	−7.631***	2.500	−8.757***	2.314	−5.222**	2.484
School autonomy						
Determining course content	−5.488**	2.576	−4.831**	2.266	−8.963***	2.604
Establishing teachers' starting salaries	−21.559***	3.231	−12.861***	2.910	−11.514***	3.327
Choosing textbooks	1.820	5.472	4.488	4.388	3.721	6.192
Deciding on budget allocations within school	8.245*	4.785	13.039***	4.750	9.67*	5.391
Formulating school budget	−5.734**	2.924	−5.812**	2.697	−1.12	3.112
Hiring teachers	6.843**	2.902	−4.098*	2.377	−0.561	2.770
Public vs. private operation and funding						
Publicly managed school	−19.189***	3.857	−12.600***	2.975	−15.15***	3.346
Government funding (share)	3.929	4.926	−3.848	4.079	−7.443	4.733
Interactions with external exit exam (EEE)						
Standardized tests x EEE	11.109***	3.668	14.106***	3.599	9.675***	3.667
School autonomy						
Determining course content x EEE	16.688***	4.182	13.680***	4.342	18.453***	4.588
Establishing teachers' starting salaries x EEE	27.979***	4.424	13.552***	4.095	9.522**	4.432
Choosing textbooks x EEE	57.898***	10.955	63.433***	10.182	69.084***	12.010
Deciding on budget allocations x EEE	−3.202	4.378	−0.055	4.265	−3.863	4.711
Formulating school budget x EEE	8.513	7.313	2.419	7.053	0.995	7.294
Hiring teachers x EEE	−2.153	4.920	2.266	4.351	−3.847	4.687
Public vs. private operation and funding						
Publicly managed school x EEE	6.424	5.481	2.200	4.531	6.964	4.778
Government funding (share) x EEE	0.614	7.782	2.792	6.863	5.273	7.839
Resources and teachers						
Educational expenditure per student (1,000$)	7.908***	2.555[a]	3.988	2.594[a]	−1.667	3.922[a]
Class size (m/s/r) (instr. by student–teacher ratio)	0.879*	0.512	1.446***	0.499	0.36	0.421
Instructional material						
Not at all lacking	6.159*	1.354	6.401***	1.297	6.848***	1.402
Strongly lacking	−11.882***	3.540	−5.430	3.436	−5.098	3.923
Instruction time (1,000 min per year) (m/s/r)	0.830***	0.225	1.238***	0.211	−0.499***	0.178
Teacher education (share at school)						
Masters in pedagogy	1.822	2.660	8.338***	2.420	4.283*	2.520
Teacher certificate	11.178***	3.384	10.484***	3.445	6.471*	3.655
Masters in subject (m/s/r)	11.847***	2.963	10.101***	2.719	17.583***	3.248
Student characteristics						
Grade						
7th	−105.234***	5.693	−76.383***	4.218	−107.976***	4.880
8th	−77.491***	2.720	−65.305***	2.233	−88.668***	2.914

Table 4 continued

	Math		Science		Reading	
	Coef.	*SE*	Coef.	*SE.*	Coef.	*SE.*
9th	−35.110***	1.678	−30.770***	1.567	−40.447***	1.721
11th	28.782***	2.543	19.861***	2.744	23.133***	2.330
Country's school entry age	22.667***	5.886[a]	14.140**	6.328[a]	15.925**	7.771[a]
Age (months)	−0.845***	0.119	−0.165	0.119	−0.716***	0.102
Female	−16.896***	0.866	−3.977***	0.821	23.687***	0.796
Family background						
Born in country						
Student	4.097*	2.175	6.639***	2.064	11.311***	2.030
Mother	5.311***	1.555	7.163***	1.637	8.182***	1.352
Father	4.163**	1.557	10.295***	1.553	7.573***	1.359
Living with						
Single father	17.461***	3.710	16.587***	3.546	22.463***	4.183
Single mother	5.998	4.227	10.296**	4.288	10.851**	4.305
Both parents	12.532***	3.741	14.993***	3.639	19.317***	4.071
Parents' education						
Primary	11.846***	4.176	12.718***	4.298	18.631***	4.012
Lower secondary	13.756***	4.239	12.407***	4.179	18.962***	3.985
Upper secondary 1	17.100***	4.345	17.864***	4.363	26.556***	4.109
Upper secondary 2	20.063***	4.288	19.915***	4.301	26.615***	3.983
University	22.596***	4.333	24.374***	4.295	29.679***	4.012
Parents' work status						
At least one half-time	1.769	2.052	−1.860	2.064	0.057	1.758
At least one full-time	15.181***	1.669	10.469***	1.639	11.470***	1.457
Both full-time	14.857***	1.697	11.366***	1.686	11.527***	1.493
Parents' job						
Blue collar	−10.036***	1.378	−10.033***	1.354	−10.903***	1.179
White collar	9.054***	1.045	8.161***	0.998	11.284***	0.844
Books at home						
1–10 books	14.216***	3.378	12.911***	3.166	27.492***	3.833
11–50 books	28.059***	3.318	27.961***	2.970	43.055***	3.390
51–100 books	35.180***	3.373	35.604***	3.006	49.487***	3.230
101–250 books	50.039***	3.444	48.938***	3.059	65.547***	3.363
251–500 books	59.980***	3.474	59.853***	3.121	76.480***	3.473
More than 500 books	61.734***	3.547	61.479***	3.180	76.199***	3.489
International socio-economic index (ISEI)	0.428***	0.033	0.415***	0.033	0.548***	0.026
School's community location						
Small town (3,000–15,000)	2.152	2.854	3.457	2.735	2.813	2.765
Town (15,000–100,000)	4.140	3.048	3.850	2.952	6.600**	2.977
City (100,000–1,000,000)	6.243*	3.410	6.089*	3.237	9.999***	3.289
City center of city with >1 million people	7.473*	4.176	4.846	3.893	11.702**	4.010
Elsewhere in city with >1 million people	−0.344	3.797	2.060	3.713	6.489*	3.747
GDP per capita (1,000$)	0.431	0.635[a]	1.390**	0.543[a]	2.178***	0.459[a]
Home Incentives and Inputs						
Parental support						
Strongly lacking	−19.195***	1.941	−17.995***	1.859	−21.286***	2.037
Not at all lacking	12.008***	2.664	7.177***	2.315	8.356***	2.416
Homework						
>1 and <3 hours per week	8.551***	0.868	7.073***	0.941	9.046***	0.693
>3 hours per week	11.387***	1.122	8.407***	1.247	5.449***	1.067

Table 4 continued

	Math		Science		Reading	
	Coef.	*SE*	Coef.	*SE*	Coef.	*SE*
Computers at home						
One	−4.246***	1.187	−3.325***	1.104	−5.376***	1.056
More than one	−11.714***	1.172	−12.964***	1.181	−13.445***	1.065
Imputation dummies	Incl.		incl.		Incl.	
Students (units of observation)	96,507		96,416		173,618	
Schools (PSUs)	6,611		6,613		6,626	
Countries (strata)	31		31		31	
R^2	0.344		0.285		0.354	
R^2 (without imputation controls)	0.318		0.258		0.322	

Notes: Dependent variable: PISA international test score. 2SLS regression in each subject, with class size instrumented by schools' student-teacher ratio. Regressions weighted by students' sampling probabilities. Coef.: Coefficient estimate. SE.: Clustering-robust standard error (taking account of correlated error terms within schools), corrected for imputed data using the Schafer and Schenker (2000) procedure

Significance level (based on clustering-robust standard errors): ***1%:**5%:*10%. [a] Clustering-robust standard errors (and thus significance levels) based on countries rather than schools as clusters

4.1 Control variables: student, family and school characteristics

Dozens of variables of student characteristics, family background and home inputs show a statistically significant association with student performance on the three PISA tests. These include indicators of grade levels, age, gender, immigration status, family status, parental education and work, socio-economic background of the family, community location and home inputs (see Table 4).

Given that reading is a new subject on the PISA test, the findings on gender differences seem noteworthy. In math and science, boys perform statistically significantly better than girls, at 16.9 achievement points (AP) in math and 4.0 AP in science. The opposite is true for reading, where girls outperform boys by 23.7 AP. Given that test scores are scaled to have an international standard deviation among OECD countries of 100, estimate sizes can be interpreted as percentage points of an international standard deviation. As a concrete benchmark for size comparisons, the unconditional performance difference between 9th- and 10th-grade students (the two largest grade categories) in our sample is 30.3 AP in math, 32.4 in science and 33.2 in reading. That is, the boys' lead in math equals roughly half of this grade equivalent, and the girls' lead in reading roughly two thirds. As an alternative benchmark, when estimating the average unconditional performance difference per month between students of different age and extrapolating this to a performance difference per year of age, this equals 12.9 AP in math, 19.3 in science and 16.4 in reading.

With the exception of the gender effect, the results on associations with student and family characteristics in reading are qualitatively the same as in math and science. But most of the family-background effects tend to be larger in reading than in math and science. In general, the results are very much in line with results derived from previous international student achievement tests (e.g., Wößmann 2003a).

Two categories of variables that have not been available in the previous international achievement tests concern the work status of students' parents. First, students with at least one parent working full time perform statistically significantly better than students whose parents do not work. However, there is no statistically significant performance difference between students whose parents do not work and students whose parents work at most half-time. Neither is there a statistically significant performance difference depending on whether one or both parents worked full-time. Second, students' performance varies statistically significantly with their parents' occupation, expressed as blue-collar and white-collar dummies.[9] As throughout the paper, these effects are calculated holding all other influence factors constant. For example, they are estimated for a given level of parental education.[10]

The general pattern of associations of student performance with schools' resource endowments and teacher characteristics is that resources seem to be positively related to student performance, once family-background and institutional effects are extensively controlled for. This holds particularly in terms of the quality of instructional material and of the teaching force. By contrast, there are no positive associations with reduced class sizes, and the associations with expenditure levels are shaky and small, not guaranteeing that the benefits of the expenditures would warrant the costs.

4.2 Institutions

Economic theory suggests that external exit exams, which report performance relative to an external standard, may affect student performance positively (cf. Bishop and Wößmann 2004; Costrell 1994; Betts 1998). In line with this argument, students in school systems with external exit exams perform statistically significantly better by 19.5 AP in math than students in school systems without external exit exams (based on a specification without interaction terms, not reported in the table). This effect replicates previous findings based on other international studies (Bishop 1997; Wößmann 2003a, b). Wößmann (2003b) shows that the cross-country result is robust to the inclusion of a lot of other systemic and cultural features of a country, as well as to the inclusion

[9] White-collar workers were defined as major group 1–3 of the International Standard Classification of Occupations (ISCO), encompassing legislators, senior officials and managers; professionals; and technicians and associate professionals. Blue-collar workers were defined as ISCO 8–9, encompassing plant and machine operators and assemblers; and sales and services elementary occupations. The residual category between the two, ranging from ISCO 4–7, encompasses clerks; services workers and shop and market sales workers; skilled agricultural and fishery workers; and craft and related trades workers. The variable was set to while-collar if at least one parent was in ISCO 1–3, and to blue-collar if no parent was in ISCO 1–7.

[10] Parental education is measured by the highest educational category achieved by either father or mother, whichever is the higher one.

of regional dummies, suggesting that it is not strongly driven by severe cultural differences. Also, Bishop (1997) and Jürges et al. (2005) present within-country evidence that students perform better in regions with external exams in Canada and Germany, the two national education systems within which some regions feature external exams and others not.

The association between external exit exams and student performance in science is statistically significant at the 11% level. In reading, the relationship is also positive, but not statistically significant. However, there is no direct data available on external exit exams in reading, so that the measure used is a simple mean of math and science. Therefore, this smaller effect might be driven by attenuation bias due to measurement error. Furthermore, the low levels of statistical significance in all three subjects reflects the measurement of external exit exams at the country level, so that there are only 31 separate observations on this variable, which we account for by clustering the standard error of this variable at the country level. Still, the pattern of results provides a hint that external exit exams may be more important for performance in math and science than in reading (cf. also Bishop 2006). At the school level, there is information on whether standardized testing was used for 15-year-old students at least once a year. This alternative measures of external testing is statistically significantly related to better student performance in all three subjects.

From a theoretical point of view, one may expect institutional effects to differ between systems with and without external exams. External exams can mitigate informational asymmetries in the school system, thereby introducing accountability and transparency and preventing opportunistic behavior in decentralized decision-making. This reasoning leads to a possible complementarity between external exams and school autonomy in other decision-making areas, the extent of which depends on the incentives for local opportunistic behavior and the extent of a local knowledge lead in a given decision-making area (cf. Wößmann 2003b, c). Therefore, the model reported in Table 4 allows for heterogeneity of institutional effects across systems with and without external exams by including interaction terms between external exit exams and the other institutional measures.

The relationship between standardized tests and student achievement indeed differs strongly and statistically significantly between systems with and without external exit exams. If there are no external exit exams, regular standardized testing is statistically significantly negatively related to student achievement in all three subjects. That is, if the educational goals and standards of the school system are not clearly specified, regular standardized testing can backfire and lead to weaker student performance. But the relationship between regular standardized testing and student achievement in all three subjects turns around to be statistically significantly positive in systems where external exit exams are in place.[11] That is, what was only hypothesized in Wößmann (2003c), who

[11] The statement that the positive relationship between regular standardized testing and student achievement in external-exam systems—whose point estimate can be calculated from the results reported in Table 4 as, e.g., $5.3 = -8.8 + 14.1$ in science—differs statistically significantly from zero,

lacked relevant data to support the hypothesis, is now backed up by empirical evidence: regular standardized examination seems to have additional positive performance effects when added to central exit exams.

A similar pattern can be observed for school autonomy in determining course contents. In systems without external exit exams, students in schools that have autonomy in determining course contents perform statistically significantly worse than otherwise. That is, the effect of school autonomy in this area seems to be negative if there are no external exit exams to hold schools accountable for what they are doing. This effect turns around to be statistically significantly positive where schools are made accountable for their behavior through external exit exams. This pattern of results suggests that the decision-making on determining course contents entails substantial incentives for local opportunistic behavior as well as significant local knowledge lead (cf. Wößmann 2003b). The incentives for local opportunistic behavior stem from the fact that content decisions influence teacher workloads, which can account for the negative autonomy effect in systems without accountability. The local knowledge lead stems from the fact that teachers probably know best what specific course contents would be best suited for their specific students, which can account for the positive autonomy effect in systems where external exit exams mitigate the scope for opportunistic behavior.

The decision-making area of establishing teachers' starting salaries shows a similar pattern of results. A negative association between school autonomy in establishing teachers' starting salaries and student performance in systems without external exit exams turns around to be positive in systems with external exit exams in math, and gets reduced to about zero in the other two subjects.

In systems without external exit exams, there is no statistically significant association between school autonomy in choosing textbooks and student achievement in either subject. However, there is a substantial statistically significant positive association in systems with external exit exams in all three subjects. The estimated effect sizes seem particularly large, though, which may hint at possible remaining biases. The result pattern reflects the theoretical case where incentives for local opportunistic behavior are offset by a local knowledge lead (Wößmann 2003b). External exit exams suppress the negative opportunism effect and keep the positive knowledge-lead effect. Thus, it seems that the positive association of student performance with school autonomy in textbook choice prevails only in systems where schools are made accountable for their behavior through external exit exams.

Economic theory suggests that school autonomy in such areas as process operations and personnel-management decisions may be conducive to student performance by using local knowledge to increase the effectiveness of teaching. By contrast, school autonomy in areas that allow for strong local opportunistic

---

is based on an auxiliary specification (not reported) which defines the interaction terms the other way round and thus allows to give an assessment of the statistical significance of the relationship between regular standardized testing and student achievement in external-exam systems. The same is true for similar statements made below.

behavior, such as standard and budget setting, may be detrimental to student performance by increasing the scope for diverting resources from teaching (cf. Bishop and Wößmann 2004). The result on school autonomy in the process operation of choosing textbooks supports this view. Similarly, school autonomy in deciding on budget allocations within schools are statistically significantly associated with higher achievement in all three subjects, once the level at which the budget is formulated is held constant. This pattern does not differ significantly between systems without and with central exams, although it tends to get stronger in the latter case. The pattern suggests that this decision-making area features only small incentives for local opportunistic behavior, but a significant local knowledge lead (cf. Wößmann 2003b).

By contrast, the association between school autonomy in formulating their school budget and student achievement is statistically significantly negative in math and science, again not differing significantly between systems with and without external exams. This combination of effects, which suggests having the size of the budget externally determined while having schools decide on within-school budget allocations themselves, replicates and corroborates the findings reported in Wößmann (2003a). Furthermore, the PISA indicator of within-school budget allocation seems superior to the data previously used in TIMSS, where this indicator was not available and information on teachers' influence on purchasing supplies was used instead as a proxy.

Students in schools that have autonomy in hiring their teachers perform statistically significantly better in math, corroborating theory (Bishop and Wößmann 2004) and previous evidence (Wößmann 2003a). The association is insignificant in reading, though, and it is significantly negative in science, albeit only in systems without external exit exams.

PISA also provides school-level data on the public/private operation and funding of schools, not previously available at the school level in international studies. Economic theory is not unequivocal on the possible effects of public versus private involvement in education, but it often suggests that private school operation may lead to higher quality and lower cost than public operation (cf. Shleifer 1998; Bishop and Wößmann 2004), while reliance on private funding of schools may or may not have detrimental effects for some students (cf. Epple and Romano 1998; Nechyba 2000). In the PISA database, public schools are defined as schools managed directly or indirectly by a public education authority, government agency or governing board appointed by government or elected by public franchise. By contrast, private schools are defined as schools managed directly or indirectly by a non-government organization, e.g. a church, trade union or business. The results show that students in privately managed schools perform statistically significantly better than students in publicly managed schools, after controlling for the large set of background features. The effect does not differ significantly between systems with and without external exit exams.

In contrast to the management of schools, we find that the share of private funding that a school receives is not significantly related to student performance, once the mode of management is held constant. In PISA, public funding is defined as the percentage of total school funding coming from government sources at different levels, as opposed to fees, donations and so on. Thus, the results on public versus private involvement differ between the management and funding of schools.

Overall, the share of performance variation accounted for by our models at the student level is relatively large, at 31.8% in math, 25.8% in science and 32.2% in reading (not counting variation accounted for by the imputation dummies). This is substantially larger than in previous models using TIMSS data, where 22% of the math variation and 19% of the science variation could be accounted for at the student level (Wößmann 2003a).

In sum, our evidence corroborates the notion that institutions of the school system are important for student achievement. External and standardized examinations seem to be performance-conducive. The effect of school autonomy depends on the specific decision-making area and on whether the school system has external exams. School autonomy is mostly positively associated with student performance in areas of process and personnel decisions, which may contain informational advantages at the local level. By contrast, the association is negative in areas of setting standards and budgets, which may be prone to local rent-seeking activities. Furthermore, the general pattern of results suggests that the effects of school autonomy differ between systems with and without external exit exams. External exit exams and school autonomy are complementary institutional features of a school system. School autonomy tends to be more positively associated with student performance in all subjects when external exit exams are in place to hold autonomous schools accountable for their decisions. This evidence corroborates the reasoning of external exams as "currencies" of school systems (Wößmann 2003c) which ensure that decentralized school systems function in the interest of students' educational performance. Finally, private school operation is associated with higher student achievement, which is not true for private school financing.

5 Robustness of results

To test whether the reported results on institutional effects in our base specification are sensitive to the specific model specification, we have preformed numerous robustness checks in terms of the sensitivity of the imputation, the set of controls and the specific sample.

In terms of the sensitivity of imputations, we test our employed method against three alternative treatments of missing data. First, to test whether the reported coefficient for each variable hinges on the inclusion of students with missing data on the specific variable, we re-estimate the reported specification as many times as there are variables in the model, each time omitting the students for whom data is missing on a specific variable. In all cases, the estimates of the alternative model are of similar magnitude and statistical significance to the estimation on the full (imputed) student sample, confirming that reported coefficient estimates are not driven by the employed imputation method.

Second, rather than performing our conditional mean imputation method, we simply impute a constant for each missing value and add, as before, an imputation dummy for each variable that equals 1 if the respective variable has a missing value for the student and 0 otherwise. Again, the alternative procedure leads to estimates that are comparable in terms of magnitude and statistical significance to the base specification, with the sole exception of school autonomy over deciding on budget allocations within school. While the statistically significant positive estimate on this variable remains in the alternative imputation method, its interaction term with external exit exams turns statistically significantly negative. This negative interaction does not survive our more elaborate imputation method.

Third, we re-estimate our base specification under omission of the controls of imputation dummies [all terms containing *D*s in (1)]. Again, most results are very robust to this alternative specification, which implicitly assumes that observations are missing conditionally at random. Exceptions are that in math, the coefficient estimate on school autonomy in formulating the school budget loses statistical significance, while its interaction term with external exit exams gets statistically significant (negative), and that again only in math, the interaction term of government funding and external exit exams turns statistically significant. It seems that the assumption of data missing conditionally at random would lead to biased estimates in these few cases.

The next set of robustness checks relates to the controls included in the model. First, to account for group selection into schools based on observable characteristics, we estimate an alternative specification that controls for the composition of the student population, in terms of family background measures, such as median parental education or socio-economic status at the school level. The results confirm the institutional effects estimated in the base specification, suggesting that they are not driven by the differential composition of the student body. In particular, the performance differential between publicly and privately operated schools is robust to observable school composition effects, corroborating findings by Dronkers and Robert (2003). Of course, the differential may still be affected by selection on unobservable characteristics. The sole difference in findings to the base specification is that in math, the positive association between government funding share and student achievement reaches standard levels of statistical significance. The performance difference between fully privately financed and fully state financed schools is approximately 10 AP.

Second, we add a variable on the age of first ability-based tracking into different school types at the country level (from OECD 2001) as an additional institutional control. The coefficient estimates on the tracking variable are very close to zero and statistically highly insignificant, suggesting that tracking does not have a significant association with the level of student performance across countries. Third, we add a variable on teacher salaries, namely annual statutory salaries of teachers in public institutions of lower secondary education after 15 years of experience (measured in PPP, from OECD 2003). They turn out statistically insignificant, which might be related to the fact that internationally comparable salary data are missing for about one third of our sample.

Fourth, we initially included two additional institutional variables on school autonomy, namely in firing teachers and in determining teachers' salary increases. These two variables proof highly collinear with school autonomy in hiring teachers and in determining teachers' starting salaries. Because adding them to the base specification renders them statistically insignificant, we chose to drop them from the specification, noting that their collinear counterparts may capture some of their potential effects.

Fifth, it is not obvious that the international socio-economic index (ISEI), which is based on parental occupations, should be included as a separate family-background control. However, all results prove robust to its inclusion or exclusion in the specification. Even the coefficients on the dummies of parents' jobs show separate predictive power, although they get noticeably larger in a specification without the ISEI control.

Sixth, it is also not obvious that the regressions should control for grade levels. Given PISA's age-based target population, a student's grade will to some extent be endogenous to her performance, particularly in systems with common grade repetition. Also, international differences in school entry age are only controlled for at the country level. Unfortunately, there is no data on individual school entry age, nor on grade repetition. Thus, to check for robustness of our results, we repeated the regressions without the grade controls. There were few qualitative changes to the presented results, with two noteworthy exceptions. The first is the coefficient on countries' school entry age, which turns significantly negative in a specification without grade dummies, as might be expected. The second is the coefficient on the share of a school's government funding, which turns statistically significantly negative in science and reading without grade controls. This change is driven by a negative correlation between grade level and government funding, which reflects the fact that schools that serve higher grades are likely to depend more on private funding. On average, grades 10 and higher receive 8.4% points less public funding than grades 9 and lower. Thus, not controlling for grades might leave the coefficient on government funding biased by sorting of weaker students into schools with a higher share of government funding. Given these reasons to expect the government-funding result to be biased without grade controls, we decided to keep the grade controls in the base specification.

The final set of robustness checks relates to the definition of the student and country sample. The first two alternative sample specifications also refer to the specific grades included. First, while our base specification dropped the small number of students in grades six or lower and twelve or higher, who seem to be strong outliers, their inclusion in the model does not change any of the substantive results.

Second, to reduce the sample of students even further to the grade levels which students should normally attend based on country rules, we re-estimate our model on a sample that excludes all students who are not in the two grades with the largest share of 15-year-olds in each respective country. Again, this alternative sample yields qualitatively the same results as our base model, indicating that that grade repetition polices, which are less relevant in the two-grade specification, do not drive the results of the base specification.

Third, restricting the sample of countries to a more homogenous group does not change the qualitative findings, either. We restrict the sample to the group of economically developed OECD countries, which drops the three poorest countries from our base sample (Brazil, Latvia and Russia). Next, we even restrict the sample to countries with a PPP-measured GDP per capita in 2000 of at least 13,000, dropping altogether six countries from our base sample (Hungary, Mexico and Poland in addition to the previous three). It turns out that none of the reported institutional effects are driven by differences between developing and developed countries, but are robust to variation within these more homogeneous groups of countries.

Fourth, it seems that the positive association between educational expenditure per student and student performance in math is largely driven by a few countries with very low spending, but is not existent among the developed OECD countries. Once the four countries with particularly low spending levels are excluded from the sample, the coefficient on educational expenditure per student at the country level gets statistically insignificant also in math.[12]

In sum, the results reported for our base specification prove remarkably robust to changes in the imputation specification, additional controls and different samples.

6 Explanatory power at the country level

In our regressions, the five categories of variables—student characteristics; family background; home incentives and inputs; resources and teachers; and institutions and their interactions—all add statistically significantly to an explanation of the variation in student performance. To assess how much each of these categories, as well as the whole model combined, adds to an explanation of the between-country variation in student performance, we do the following exercise. First, we perform the student-level regression reported in Table 4, equivalent to (1), only without the imputation dummies:

$$T_{is} = S_{is}\beta_{11} + F_{is}\beta_{12} + H_{is}\beta_{13} + R_{is}\beta_2 + I_s\beta_3 + \varepsilon_{is}, \tag{2}$$

where the student-background vector B is subdivided into three parts as in Table 4, namely student characteristics S, family background F and home inputs H.

[12] The OECD (2001) reports a measure of cumulative educational expenditure per student for 24 countries, which cumulates expenditure over the theoretical duration of education up to the age of 15. Using this alternative measure of educational expenditure, we find equivalent results of a statistically significant relation with math performance, a weakly statistically significant relation with science performance, and a statistically insignificant relation with reading performance. Given that this measure is available for only 24 countries, and given that it partly proxies for the duration of schooling in different countries, we stick to our alternative measure of average annual educational expenditure per student in secondary education.

Next, we construct one index for each of the five categories of variables as the sum of the products between each variable in the category and its respective coefficient β. That is, the student-characteristics index S^I is given by

$$S_{is}^{I} = S_{is}\beta_{11}, \tag{3}$$

and equivalently for the other four categories of variables. Note that, as throughout the paper, this procedure keeps restricting all coefficients β to the ones received in the student-level international education production function, abstaining from any possible effect heterogeneity between countries or levels (e.g., within versus between countries).

Finally, we take the country means of each of these indices in each subject, as well as of student performance in each subject, properly weighting the student observations by their sampling probabilities within each country. This allows us to perform regressions at the country level, on the basis of 31 country observations. These regressions allow us to derive measures of the contribution of each of the five categories of variables to the between-country variation in test scores.

Note that this whole exercise is *not* set up to maximize explanatory power at the between-country level, but rather to fully replicate the simple cross-country model that we estimate above. That is, we do not allow for country heterogeneity. We do not distinguish between estimates based on their precision. We exclude the imputation dummies because as statistical controls they obviously do not really represent a part of the theoretical model, although they would increase the apparent explanatory power of our model. The whole point of this exercise is to estimate the very same model as before, at the student level, and then see how much variation is accounted for at the country level. In many respects, the exercise is set up in a way that loads the dice against a strong finding at the between-country level.

But despite this setup, our models can account for nearly all of the variation that exists between countries. As reported in Table 5, models regressing the country means of student performance on the five indices yield an explanatory power of between 84.6 and 87.3% of the total cross-country variation in test scores. This is reassuring for our model specification, which leaves little room for substantial unobserved country-specific heterogeneity. Most of the unexplained variation in student-level test scores, which ranges from 67.8 to 74.2% in the regressions of Table 4, thus seems to be due to unobserved within-country student-level ability differences and not due to a country-level component.

To assess the contribution of the five indices individually, we perform two analyses. First, we enter each index individually and look at the R^2 of that regression, which would attribute any variation that is joint with the other indices to this index if they are positively correlated. Second, we look at the change in the R^2 of the model that results from adding each specific index to a model that already contains the other four indices. Note that the latter procedure will result in a smaller ΔR^2 than the former if the additional index is positively correlated with the other indices, and a larger ΔR^2 if they are negatively correlated.

Table 5 Contribution to explanatory power (ΔR^2) at the country level

	Entered individually			Entered after the remaining four categories		
	Math	Science	Reading	Math	Science	Reading
Student characteristics	0.198	0.263	0.245	0.024	0.043	0.040
Family background	0.491	0.409	0.384	0.012	0.114	0.273
Home incentives and inputs	0.004	0.013	0.048	0.001	0.002	0.001
Resources and teachers	0.240	0.283	0.126	0.285	0.106	0.049
Institutions and their interactions	0.339	0.242	0.194	0.266	0.292	0.312
Full model	0.873	0.846	0.853			

Notes: Dependent variable: Country means of PISA international test scores. See text for details

When each of the five indices is entered individually, student characteristics can account for 19.8 to 26.3% of the country-level variation in test scores, family background for 38.4 to 49.1%, home incentives and inputs up to 4.8%, resources and teacher characteristics for 12.6 to 28.3% and institutions for 19.4 to 33.9%. When entered after the other four indices, the contributions of the first three indices drop considerably. More interestingly, the ΔR^2 of entering resources and teacher characteristics drops to 10.6% in science and to 4.9% in reading (it stays at 28.5% in math).[13] Institutions account for a ΔR^2 of 26.6% in math, 29.9% in science and 31.2% in reading. This shows that both resources and institutions contribute considerably to the international variation in student performance, but that the importance of institutions for the cross-country variation in test scores seems to be greater than that of resources.

7 Conclusion

The international education production functions estimated in this paper can account for most of the between-country variation in student performance in math, science and reading. Student characteristics, family backgrounds, home inputs, resources and teachers, and institutions all contribute significantly to differences in students' educational achievement. The PISA study used in this paper distinguishes itself from previous international student achievement studies through its focus on reading literacy, real-life rather than curriculum-based questions, age rather than grade as target population and more detailed data on family backgrounds and institutions. The PISA-based results of this paper corroborate and extend findings based on previous international studies. In particular, the institutional structure of the education system is again found to be strongly associated with how much students learn in different countries, consistently across the three subjects. Institutions account for roughly one quarter of the international variation in student performance.

[13] Note that part of the variation attributed to resources stems from the counterintuitive positive coefficient on class size.

The main findings are as follows. Our results confirm previous evidence that external exit exams are statistically significantly positively associated with student performance in math, and marginally so in science. The positive association in reading lacks statistical significance, which may however be due to poor data quality on the existence of external exit exams in this subject. As an alternative measure of external examination, regular standardized testing shows a statistically significant positive association with performance in all three subjects. Consistent with theory as well as previous evidence, superior student performance is associated with school autonomy in personnel-management and process decisions such as deciding budget allocations within schools, textbook choice and hiring of teachers (the latter only in math). By contrast, superior performance is associated with centralized decision-making in areas with scope for decentralized opportunistic behavior, such as formulating the overall school budget. The performance effects of school autonomy tend to be more positive in systems where external exit exams are in place, emphasizing the role of external exams as "currencies" of school systems. Finally, students in publicly managed schools perform worse than students in privately managed schools. However, holding the mode of private versus public management constant, the same is not true for students in schools that receive a larger share of private funding. The findings on institutional associations are mostly consistent across the three subject areas.

In terms of control variables, as in previous studies students' family background is consistently strongly associated with their educational performance. We find that the estimated effects of family background as measured by parental education, parental occupation or the number of books at home are considerably stronger in reading than in math and science. Furthermore, while boys outperform girls in math and science, the opposite is true in reading. While smaller classes do not go hand in hand with superior student performance, better equipment with instructional material and better-educated teachers do.

Obviously, this research opens numerous directions for future research. We close by naming three of them. First, this paper has focused on the average productivity of school systems. As a complement, it would be informative to analyze how the different inputs affect the equity of educational outcomes. Second, related to educational equity, possible peer and composition effects have been neglected in this paper. Given that they play a dominant role in many theoretical models of educational production, an empirical analysis of their importance in a cross-country setting would be informative, but is obviously limited by the well-known problems of their empirical identification. Finally, an empirical conceptualization of additional important institutional and other features of the school systems could contribute to an advancement of our knowledge. For example, an encompassing empirical manifestation of such factors as the incentive-intensity of teacher and school contracts, other performance incentives and teacher quality is still missing in the literature.

Acknowledgments Financial support by the Volkswagen Foundation is gratefully acknowledged. We would like to thank John Bishop, George Psacharopoulos, Andreas Schleicher, Petra Todd, Manfred Weiß, Barbara Wolfe, the editor and three anonymous referees, as well as participants at the annual conferences of the American Economic Association in Philadelphia, PA, the Australasian Meeting of the Econometric Society in Melbourne, the European Association of Labour Economists in Lisbon, the International Institute of Public Finance in Milan, the German Economic Association in Dresden, the Human Capital Workshop at Maastricht University and at the Ifo seminar in Munich for helpful comments.

Appendix: Data imputation method

To obtain a complete dataset for all students for whom performance data are available, we impute missing values of explanatory variables using a set of "fundamental" explanatory variables F that are available for all students. These fundamental variables F include gender, age, six grade dummies, three dummies on which parent the students live with, six dummies for the number of books at home, GDP per capita as a measure of the country's level of economic development and the country's average educational expenditure per student in secondary education.[14]

Missing values for student i of the variable M are imputed by first regressing F on available values (s) of M:

$$M_s = F_s\theta + \varepsilon_s. \tag{4}$$

Then, the imputed value of M for i is predicted using student i's values of the F variables and the coefficient vector θ obtained in regression (4):

$$M_i = F_i\theta. \tag{5}$$

The imputation method for implied variables is a WLS estimation for metric discrete variables, an ordered probit model for ordinal variables and a probit model for dichotomous variables. For metric discrete variables, predicted values are filled in for missing data. For ordinal and dichotomous variables, in each category the respective predicted probability is filled in for missing data.

Because this imputation method predicts the expected values of missing data in a deterministic way, standard errors would be biased downward when estimated in the standard way. To overcome the downward bias, we employ the procedure for correction of standard errors suggested by Schafer and Schenker (2000), which provides the following corrected standard error for the coefficient estimate β of any given variable:

$$\sigma_\beta^{\text{corr}} = \sqrt{\hat{\sigma}_\beta^2 + 2r^2\hat{\sigma}_\beta^2\frac{\hat{\sigma}_{\beta_{\text{mis}}}^2}{\hat{\sigma}_\beta^2} + \hat{\sigma}_\beta^2\frac{\hat{\sigma}_{\beta_{\text{imp}}}^2}{\hat{\sigma}_\beta^2}}. \tag{6}$$

[14] The small amount of missing data on the variables in F is imputed by median imputation at the lowest available level (school or country).

The corrected standard error consists of three parts. The first part is the standard error estimated in the standard way, which disregards the imputation of some of the data. The second part is a correction term which takes into account the share r of missing values among the total number of observations for this variable, together with the ratio of the variance of the imputed data $\hat{\sigma}^2_{\beta_{\text{mis}}}$ over the variance of the full sample $\hat{\sigma}^2_{\beta}$. The third part is a correction term which takes into account the uncertainty of the imputation model, i.e. the variance of the residuals of the imputation model $\hat{\sigma}^2_{\beta_{\text{imp}}}$ (see Schafer and Schenker 2000 for details). To account for the complex survey design with intra-cluster correlations, all variance arguments in (6)—$\hat{\sigma}^2_{\beta}, \hat{\sigma}^2_{\beta_{\text{mis}}}$ and $\hat{\sigma}^2_{\beta_{\text{imp}}}$—are estimated using a clustered variance structure with sampling weights (see Schafer and Schenker 1997, Sect. 5.4, for details on the extension to weighted estimates with clustering).

References

Adams R, Wu M (eds) (2002) PISA 2000 technical report. Organisation for Economic Co-operation and Development (OECD), Paris

Akerhielm K (1995) Does class size matter? Econ Educ Rev 14:229–241

Betts JR (1998) The impact of educational standards on the level and distribution of earnings. Am Econ Rev 88:266–275

Bishop JH (1997) The effect of national standards and curriculum-based exams on achievement. Am Econ Rev 87:260–264

Bishop JH (2006) Drinking from the fountain of knowledge: student incentive to study and learn. In: Hanushek EA, Welch F (eds) Handbook of the economics of education. (forthcoming) North-Holland, Amsterdam

Bishop JH, Wößmann L (2004) Institutional effects in a simple model of educational production. Educ Econ 12:17–38

Costrell RM (1994) A simple model of educational standards. Am Econ Rev 84:956–971

Dronkers J, Robert P (2003) The effectiveness of public and private schools from a comparative perspective. EUI Working Paper SPS 2003–13. European University Institute, Florence

DuMouchel WH, Duncan GJ (1983) Using sample survey weights in multiple regression analyses of stratified samples. J Am Statist Assoc 78:535–543

Epple D, Romano RE (1998) Competition between private and public schools, vouchers, and peer-group effects. Am Econ Rev 88:33–62

Fertig M (2003a) Who's to blame? The determinants of German students' achievement in the PISA 2000 study. IZA Discussion Paper 739. Institute for the Study of Labor, Bonn

Fertig M (2003b) Educational production, endogenous peer group formation and class composition: evidence from the PISA 2000 study. IZA Discussion Paper 714. Institute for the Study of Labor, Bonn

Fertig M, Schmidt CM (2002) The role of background factors for reading literacy: straight national scores in the PISA 2000 study. IZA Discussion Paper 545. Institute for the Study of Labor, Bonn

Fuchs T, Wößmann L (2004a) What accounts for international differences in student performance? A re-examination using PISA data. CESifo Working Paper 1235. CESifo, Munich

Fuchs T, Wößmann L (2004b) Computers and student learning: Bivariate and multivariate evidence on the availability and use of computers at home and at school. Brussels Econ Rev 47:359–385

Gundlach E, Wößmann L, Gmelin J (2001) The decline of schooling productivity in OECD countries. Econ J 111:C135–C147

Hanushek EA (2002) Publicly provided education. In: Auerbach AJ, Feldstein M (eds) Handbook of public economics, Vol 4. North Holland, Amsterdam, pp 2045–2141

Hanushek EA et al (1994) Making schools work: improving performance and controlling costs. Brookings Institution Press, Washington
Hoxby CM (1999) The productivity of schools and other local public goods producers. J Public Econ 74:1–30
Hoxby CM (2001) All school finance equalizations are not created equal. Q J Econ 116:1189–1231
Jürges H, Schneider K, Büchel F (2005) The effect of central exit examinations on student achievement: quasi-experimental evidence from TIMSS Germany. J Eur Econ Assoc 3:1134–1155
Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York
Moulton BR (1986) Random group effects and the precision of regression estimates. J Econ 32:385–397
Nechyba TJ (2000) Mobility, targeting, and private-school vouchers. Am Econ Rev 90:130–146
Nechyba TJ (2003) Centralization, fiscal federalism, and private school attendance. Int Econ Rev 44:179–204
Organisation for Economic Co-operation and Development (OECD) (2000) Measuring student knowledge and skills: the PISA 2000 assessment of reading, mathematical and scientific literacy. OECD, Paris
Organisation for Economic Co-operation and Development (OECD) (2001) Knowledge and skills for life: first results from the OECD Programme for International Student Assessment (PISA) 2000. OECD, Paris
Organisation for Economic Co-operation and Development (OECD) (2002) Manual for the PISA 2000 database. OECD, Paris
Organisation for Economic Co-operation and Development (OECD) (2003) Education at a glance: OECD indicators 2003. OECD, Paris
Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
Schafer JL, Schenker N (1997) Inference with imputed conditional means. Pennsylvania State University, Department of Statistics, Technical Report #97–05 (available at http://www.stat.psu.edu/reports/1997/tr9705.pdf)
Schafer JL, Schenker N (2000) Inference with imputed conditional means. J Am Statist Assoc 95:144–154
Shleifer A (1998) State versus private ownership. J Econ Perspect 12:133–150
Todd PE, Wolpin KI (2003) On the specification and estimation of the production function for cognitive achievement. Econ J 113:F3–F33
West MR, Wößmann L (2006) Which school systems sort weaker students into smaller classes? International evidence. Eur J Politi Econ (forthcoming) (available as CESifo Working Paper 1054, CESifo, Munich)
Wolter SC, Coradi Vellacott M (2003) Sibling rivalry for parental resources: a problem for equity in education? A six-country comparison with PISA data. Swiss J Sociol 29:377–398
Wooldridge JM (2001) Asymptotic properties of weighted m-estimators for standard stratified samples. Econ Theory 17: 451–470
World Bank (2003) World development indicators CD-Rom. World Bank, Washington
Wößmann L (2003a) Schooling resources, educational institutions and student performance: the international evidence. Oxford Bull Econ Statist 65:117–170
Wößmann L (2003b) Central exit exams and student achievement: international evidence. In: Peterson PE, West MR (eds) No child left behind? The politics and practice of school accountability. Brookings Institution Press, Washington, pp 292–323
Wößmann L (2003c) Central exams as the "currency" of school systems: international evidence on the complementarity of school autonomy and central exams. DICE Report – J Inst Comp 1:46–56
Wößmann L (2005) Educational production in Europe. Econ Policy 20:445–504
Wößmann L, West MR (2006) Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. Eur Econ Rev 50:695–736