Journal of Economic Perspectives—Volume 30, Number 3—Summer 2016—Pages 3–32

The Importance of School Systems: Evidence from International Differences in Student Achievement

Ludger Woessmann

Average achievement levels of students differ markedly across countries. On the most recent international achievement tests in math and science, the average 15 year-old student in Singapore, Hong Kong, Korea, Japan, and Taiwan is more than half a standard deviation ahead of the average student of the same age in the United States (Hanushek and Woessmann 2015b). Following the rule of thumb that average student learning in a year is equal to about one-quarter to one-third of a standard deviation, these differences are roughly equivalent to what students learn during 1.5–2 years of schooling. Similarly, the average student in Finland and Estonia is 40 percent of a standard deviation ahead of the United States, and the average Canadian student is about one-third of a standard deviation ahead. On the other hand, the average student in Peru and Indonesia is more than 1.1 standard deviations behind the United States, and achievement in Ghana, South Africa, and Honduras lags more than 1.5 standard deviations behind the United States. Overall, average achievement levels among 15 year-olds between the top- and bottom-performing countries easily differ by more than two standard deviations, or the equivalent of 6–8 years of learning.

We will present evidence that the considerable differences in student achievement across countries are systematically related to differences in the organization and governance of school systems. For example, students in many high-performing countries such as Korea and Finland, as well as in some provinces of Canada, face external exit exams at the end of high school. Most schools in Hong Kong and the United Kingdom

■ *Ludger Woessmann is Director of the Center for the Economics of Education, ifo Institute for Economic Research, and Professor of Economics, University of Munich, both in Munich, Germany. His email address is woessmann@ifo.de.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.3.3

doi=10.1257/jep.30.3.3

have considerable autonomy in deciding which courses to offer and which teachers to hire, whereas virtually no schools have this autonomy in Greece. More than half the students in the Netherlands, Belgium, Ireland, and Korea attend privately operated schools, while hardly any students in Norway and Poland do so. Students in Austria and Germany are tracked into different-ability schools at age 10, while two-thirds of OECD countries have comprehensive school systems at least until age 15.

Educational institutions such as national accountability systems or tracking regimes often vary only slightly or not at all within countries, but an international perspective provides an opportunity for comparisons. As the director of the first pilot project comparing student achievement across countries remarked, "If custom and law define what is educationally allowable within a nation, the educational systems beyond one's national boundaries suggest what is educationally possible" (Foshay 1962). Studies done within or across schools of a particular country can also suffer from the risk of selection bias, if students with specific backgrounds are more likely to attend certain schools or to become involved in certain programs, while studying national-level variation circumvents some of these selection biases.

However, these advantages of the cross-country comparative approach come with some built-in limitations. Identification of causal effects raises particular challenges in an international setting. Countries may differ in a variety of hard-to-observe ways such as cultural traits, valuation of achievement, and other preferences that are associated with both institutional choices and achievement levels. Such unobserved country heterogeneity gives rise to omitted variable bias in cross-country analyses. Moreover, only a limited number of country-level observations are available in the test data.

This essay first describes the size and cross-test consistency of international differences in student achievement. It then uses the framework of an education production function to describe how different factors of the school system, as well as factors beyond the school system, are associated with cross-country achievement differences. The discussion then focuses on research that attempts to go beyond conditional correlations by addressing some sources of potential bias in cross-country analysis. This discussion suggests that the role of resource inputs seems limited; indeed, all of the above-mentioned high-achieving countries spend considerably *less* on schools per student than the United States (OECD 2013). But differences in instruction time and teacher quality do matter. In addition, institutional features including external exams, school autonomy, private competition, and tracking affect the level and distribution of student achievement across countries and account for a substantial part of the cross-country achievement variation. The conclusion points out some major implications of educational achievement for the prosperity of individuals and nations.

How Large and Consistent Are International Differences in Student Achievement?

Large-scale international testing of student achievement has more than half a century of history, and many studies provide evidence on international differences in student achievement and how they have evolved over time.

International Rankings and the Size of Cross-Country Differences

A crucial role in the emergence and continuation of comparative testing has been played by the International Association for the Evaluation of Educational Achievement (IEA), an independent cooperative of national research institutions and government agencies (IEA 2016; Mullis, Martin, Foy, and Arora 2012). Following a pilot project in 1959–61, the IEA conducted its first international math study of eleven countries in 1964. The first science and reading studies occurred with 12–16 countries in the early 1970s, and a second round in each subject was performed in the 1980s and early 1990s. Since 1995, the Trends in International Mathematics and Science Study (TIMSS) has tested math and science achievement mostly in fourth and eighth grade every four years in between 38 and 52 voluntarily participating countries. In addition, the Progress in International Reading Literacy Study (PIRLS) has tested fourth-grade reading achievement every five years since 2001, with 48 countries participating in the most recent wave.

In 2000, the Organisation for Economic Co-operation and Development (OECD) entered international testing as a second major player. Since then, its Programme for International Student Assessment (PISA) tests representative samples of 15 year-old students in math, science, and reading every three years. In both 2009 and 2012, 65 countries participated, and 71 countries have signed up to participate in the most recent PISA installment in 2015.[1]

All these tests draw random samples of students to ensure representativeness for the national target populations. In particular, the three ongoing studies have a two-stage sampling design. At a first stage, they draw a random sample of schools in each country. Within those schools, they then randomly draw one classroom per grade (TIMSS, PIRLS) or a random sample of 15 year-old students (PISA), respectively. Each of these tests uses a common set of questions in all participating countries based on a particular effort to achieve cross-country comparability. PISA, TIMSS, and PIRLS each link their own tests over time, too. But there is no direct link between the scales of the three testing regimes or across time with the older tests.

Table 1 shows the performance of the 81 countries that have participated in the most recent installments of the PISA (2012) and TIMSS (2011) international math and science tests. Achievement is expressed on the PISA scale, which is standardized to have a mean of 500 and a standard deviation of 100 among all students in OECD countries. This standardization was done in 2003 in math and in 2006 in science. On average across OECD countries, the actual within-country standard deviation is 92 in math and 93 in science (OECD 2013). The transformation of the TIMSS 2011 data to the PISA scale follows the method suggested in Hanushek and Woessmann (2015b, Annex B).

[1] In addition, there are a couple of separate international tests whose items are aligned to the US school curriculum (which may limit international comparability), a number of regional tests in Latin America and sub-Saharan Africa, and adult literacy tests (for discussion, see Hanushek and Woessmann 2011a, table 2; Hanushek and Woessmann 2015b, chapter 4; Hanushek, Schwerdt, Wiederhold, and Woessmann 2015). The International Association for the Evaluation of Educational Achievement has also conducted studies in other subjects such as foreign languages, civic education, and computer literacy.

Table 1
Performance on Recent International Student Achievement Tests, 2011–2012

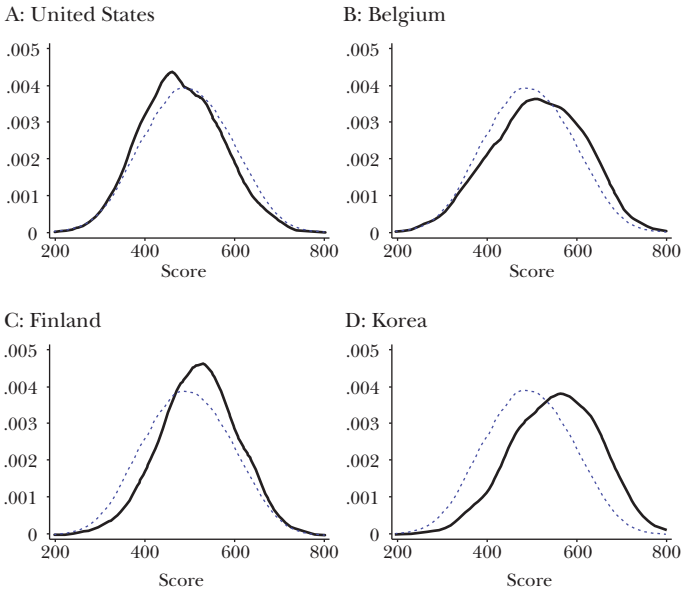
Country	*Score*	*Country*	*Score*	*Country*	*Score*
Shanghai–China	596	Norway	492	Malaysia	420
Singapore	562	Luxembourg	491	Costa Rica	418
Hong Kong–China	558	Spain	490	Mexico	414
Korea	546	Italy	489	Uruguay	413
Japan	542	**United States**	**489**	Montenegro	410
Chinese Taipei	542	Portugal	488	Bahrain	408
Finland	532	Lithuania	487	Lebanon	403
Estonia	531	Hungary	486	Georgia	401
Liechtenstein	530	Iceland	485	Brazil	398
Macao–China	529	Russian Federation	484	Jordan	397
Switzerland	523	Sweden	482	Argentina	397
Netherlands	523	Croatia	481	Albania	396
Canada	522	Slovak Republic	476	Tunisia	393
Poland	522	Ukraine	468	Macedonia	392
Vietnam	520	Israel	468	Saudi Arabia	391
Germany	519	Greece	460	Palestine	388
Australia	513	Turkey	456	Colombia	388
Ireland	512	Serbia	447	Qatar	380
Belgium	510	Bulgaria	443	Syria	379
New Zealand	508	Romania	442	Indonesia	379
Slovenia	508	United Arab Emirates	441	Botswana	376
Austria	506	Cyprus	439	Peru	371
United Kingdom	504	Thailand	435	Oman	369
Czech Republic	504	Chile	434	Morocco	348
Denmark	499	Kazakhstan	428	Honduras	328
France	497	Armenia	428	South Africa	315
Latvia	496	Iran	422	Ghana	291

Notes: The table gives the average of the scores on international math and science tests. Black: PISA 2012, 15-year-olds. Grey: TIMSS 2011, 8th grade, transformed to PISA scale as in Hanushek and Woessmann (2015b).

The cross-country differences in knowledge among same-aged students are in some cases extremely large. Remember, as a rule of thumb, learning gains on most national and international tests during one year are equal to between one-quarter and one-third of a standard deviation, which is 25–30 points on the PISA scale. Thus, the achievement difference between the average 15 year-old in the United States and in the PISA top performers—Singapore, Hong Kong, Korea, Japan, Taiwan, Finland, and Estonia—is roughly twice what students usually learn during one year. At the other end, the average difference of US achievement to the PISA bottom performers (Peru and Indonesia) amounts to the equivalent of three to four years of learning, and to five to six years to the TIMSS bottom performers (Ghana and South Africa).

In looking at lists like Table 1, it is important to focus on scores, not just ranks. For example, in the PISA 2012 math test, the achievement levels of most countries are not statistically significantly different from their closest 1–3 neighbors above and below. Where the scores are closely bunched, Portugal's achievement at rank 31 does not differ significantly from ranks 25–37 in the PISA 2012 math test (OECD 2013).

Figure 1
Distribution of Student Achievement in Selected Countries on PISA Math Test, Compared to All OECD Countries



Notes: Kernel densities of student achievement on the PISA 2012 math test. Bold solid line: specified country. Dotted line: OECD countries.

The means in Table 1 may hide important differences in the shape of the overall distribution of achievement in a country. Figure 1 displays the achievement distribution on the PISA 2012 math test for the United States and three selected countries with relatively high performance, comparing each to the overall distribution in OECD countries. The US distribution is shifted to the left and slightly more left-steep compared to the OCED distribution, but it does not have a particularly strong left or right tail. As the three example countries show, it is possible to achieve above-average mean performance with a relatively equitable distribution (Finland), with a distribution that is mostly just shifted to the right of the OECD distribution (Korea), or with a relatively unequal distribution (Belgium).

The relatively low performance of the United States compared to many OECD countries cannot be attributed to the particularly poor performance of a small group of students or of students from disadvantaged backgrounds. For example, the 25th, 50th, and 75th percentiles of the US distribution on the PISA 2012 math test are all between 13 and 15 points below the OECD average of the respective percentiles. In Hanushek, Peterson, and Woessmann (2013), my coathors and I document that both the proportion of students who achieve at a basic proficient level and the proportion of students who achieve at an advanced level in the United States are comparatively low in an international perspective. In addition, in Hanushek, Peterson, and Woessmann (2014), we show that the ranking of US students from better-educated families when compared to students from better-educated families in other countries is not

much different from the ranking of US students from less-well-educated families when compared to students from less-well-educated families in other countries.

Consistency across Different Tests

The measurement of educational achievement is subject to many psychometric and measurement choices. For example, the target population of the TIMSS test is eighth graders. Also, TIMSS has a strong curricular focus and is based on an assessment framework developed in a collaborative process with participating countries, with a test-curriculum matching analysis describing how the test matches each participating country's school curriculum. On the other hand, the target population of the PISA test is 15 year-olds, and PISA aims to assess the knowledge and skills essential for full participation in modern society, including the extrapolation and application of learned knowledge to new real-life situations.

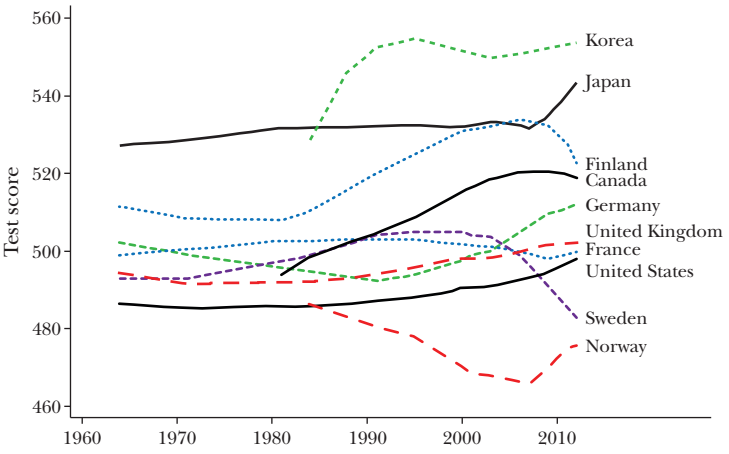
How sensitive are international comparisons to specific measurement choices? We can compare the achievement of the 28 countries that participated in the most recent installments of both tests: PISA 2012 and TIMSS 2011. Despite the differences in timing, target populations, and conceptual approaches, the correlation across the 28 countries participating in both tests is 0.944 in math and 0.930 in science (Hanushek and Woessmann 2015b). This high correlation suggests that when it comes to international comparisons, specific test designs are of secondary importance.

Another potential issue with international achievement tests is cross-country differences in sample selectivity due to different rates of enrollment, exclusion, and nonresponse. While sampling was devised to be representative of the student population in each participating country, some countries do not have universal enrollment at age 15, when students are tested in PISA. In addition, nonrandom differences in patterns of sample exclusions (for example, for handicapped children) and nonresponse can compromise comparability across countries. However, the working paper version of Hanushek and Woessmann (2011c) shows that although these factors are related to average country scores, controlling for these rates does not affect the qualitative results on institutional effects in international education production functions presented later in this paper. The variation in the extent to which countries adequately sample their entire student populations appears orthogonal to the associations analyzed here.

Changes over Time

While an assessment of countries at a point in time is reasonably straightforward, assessing changes in country performance over time is harder. The early international tests, in particular, constitute separate testing incidents without links across different tests. In Hanushek and Woessmann (2012, 2015a), we use an empirical calibration method to put all international tests from 1964 to 2003 on a common standardized scale. Our analysis shows that 73 percent of the variance across the 693 separate test observations in 50 countries occurs between countries. The remaining 27 percent combines true changes over time in countries' scores and any measurement error in the testing. That is, most of the variation in the available panel data of countries over time is across rather than within countries, implying that a large share of the country differences are consistent over time.

Figure 2
Long-Run Test Score Trends in Selected Countries, 1964–2012



Source: Extended from Hanushek and Woessmann (2015a).
Notes: Stylized depiction of standardized data from international tests 1964–2012. The figure is based on age-group- and subject-specific standardized scores from all international tests in 1964–2003 extended with the subsequently available TIMSS, PIRLS, and PISA data to 2012. It takes out age-group- and subject-specific trends in each country, smooths available test observations with locally weighted regressions, and linearly interpolates between available test observations; see Hanushek and Woessmann (2015a) for details.

Still, several countries do show either significant improvements or declines over time. Figure 2 depicts achievement trends observed in selected example countries from 1964 to 2012. The more limited variation in early decades likely reflects the lower frequency of testing before 2000. The figure shows substantial cross-sectional differences across countries. But some countries do show noteworthy changes over time. Ripley (2013) acknowledges that a previous version of this figure motivated her work on the widely acclaimed New York Times bestseller *The Smartest Kids In The World—And How They Got That Way*. While the United States was rather typical compared to most other countries, she wrote there were a few countries where "virtually *all* kids were learning critical thinking skills in math, science, and reading" (p. 4). While some countries such as Canada and Finland over the 1980s and 1990s and Germany and Japan more recently did improve substantially over time, other well-off countries deteriorated, such as Norway during the 1990s, Sweden during the 2000s, and Finland in recent years. Educational achievement levels of countries seem generally consistent over time, but they are not set in stone and can be mutable.

Descriptive Patterns Using an Education Production Function

This section uses the framework of an international education production function to document the extent to which, on a purely descriptive basis, differences in family background, school resources, and institutions can account for cross-country

differences in student achievement. These inputs are probably not exogenous to student achievement, so correlations between the inputs and test scores are very likely to be biased by omitted variables, selection, and reverse causation. While these descriptive patterns must be interpreted cautiously, they can serve as a useful guide to the more explicit discussions of causality that follow.

International Education Production Functions

An education production function models the output of education as a function of different inputs (for example, Hanushek 1986, 2002). We combine the input factors into three groups: family background factors, school resources, and institutional structures of school systems. The first group is mostly outside the control of school systems. The other two groups of factors reflect the quantity of resource inputs in the systems and the institutional structures. The basic model can be extended to include interactions between input factors.

A substantial literature has estimated such international education production functions using cross-sectional data (for an extensive review, see Hanushek and Woessmann 2011a). Early studies used aggregate country-level data to study the country-level variation in achievement scores (for example, Bishop 1997; Hanushek and Kimko 2000; Lee and Barro 2001). More recent studies also use country-level data to study, for example, the correlates of gender equality in achievement (Guiso, Monte, Sapienza, and Zingales 2008; Fryer and Levitt 2010).

However, starting with Woessmann (2003b), a number of studies have used the data from international achievement tests at the student level to estimate cross-country education production functions. Examples include Woessmann (2005b), Fuchs and Woessmann (2007), Brunello and Checchi (2007), Woessmann, Luedemann, Schuetz, and West (2009), Schneeweis (2011), and Ammermueller (2013). Because these studies use data on individual students, they can hold constant a large set of observable factors usually unavailable in national datasets. In effect, they can compare "observationally equivalent" students across countries.

For concreteness, Table 2 provides an example of a basic cross-sectional estimation of an international education production function.[2] The table shows the categories of data that are available. The dependent variable is the score from the PISA 2003 math test, with the sample restricted to the 29 participating OECD countries to provide greater comparability. The model includes a large number of explanatory variables in the three groups of input factors: family background, school resources, and institutions. The individual-level measures of family background

[2]This is a simplified version of the model used in Woessmann et al. (2009) and Hanushek and Woessmann (2011a). To allow for a more meaningful accounting analysis below, it drops the GDP per capita of the country (which is correlated with educational spending at 0.93 and yields a counterintuitive negative estimate), class size (which has a counterintuitive positive estimate), and the imputation dummies and their interactions with the main variables contained in those models. Qualitative results are similar with those variables included. Qualitative results are also unaffected when adding the country-average value of the Index of Economic, Social and Cultural Status (ESCS), the average share of students with an immigrant background in a country, or continental fixed effects to the model. Country-average ESCS in fact enters marginally significantly negatively and the migrant share insignificantly. Reported standard errors are clustered at the country level, which may be overly conservative for variables that vary at the school or student level.

are taken from student background questionnaires that students complete in the PISA study; the measures of school resources and institutions are mostly taken from school background questionnaires that the principals of participating schools complete; these measures are combined with country-level data on expenditure per student and external exit exams that come from outside sources (for details, see Appendix A of Woessmann et al. 2009). Descriptively, this model accounts for 34 percent of the achievement variance at the individual student level.

Factors beyond the School System: Family, Socioeconomic, and Cultural Background

Some of the personal characteristics that have meaningful and statistically significant magnitudes in Table 2 include student characteristics such as age, gender, and participation in early childhood education, along with indicators for family status, parental education, parental work status and occupation, the number of books at home, immigration background, and the language spoken at home. For example, the achievement difference between students in the highest category of more than 200 books at home versus the lowest category of fewer than 10 books at home—a proxy for aspects of educational, social, and economic background—amounts to more than half a standard deviation in the PISA test score.

There are two main types of analysis in the literature analyzing socioeconomic backgrounds in the international tests. The first type looks at how much socioeconomic background contributes to country-level differences in educational outcomes. The second type of analysis compares the within-country association of socioeconomic factors with student achievement, sometimes referred to as socioeconomic gradients, across countries. For example, in Schütz, Ursprung, and Woessmann (2008), we estimate the associations of family background with student achievement—interpreted as measures of the inequality of educational opportunity—in different countries using TIMSS data and relate them to measures of institutions of the school systems. We show that family background effects are systematically larger in countries with early tracking and less-extensive pre-primary education systems.[3] Jerrim and Micklewright (2014) use PISA and PIRLS data to analyze the extent to which cross-country comparisons of socioconomic gradients are affected by differences in reporting errors.

Several studies have focused on the achievement of children with an immigration background, looking at both socioeconomic and institutional characteristics. For example, Dustmann, Frattini, and Lanzara (2012) show that in many countries, observed differences in parental background (including parental education and occupation and the language spoken at home) can account for much of the lower PISA achievement of children of immigrants compared to native children. They also find that children of Turkish immigrants perform better in most host countries than Turkish children in Turkey. Also using PISA data, Cobb-Clark, Sinning, and Stillman (2012) show that the migrant–native achievement gap is significantly associated with institutional features of the host countries such as school starting age, ability tracking, private

[3]Applying a similar approach to outcomes beyond school age, Brunello and Checchi (2007) find that early tracking is related to larger effects of family background on educational attainment and earnings in the labor market, but not on on-the-job training and adult literacy.

Table 2
A Simple International Education Production Function: A Least-Squares Regression
(dependent variable is student's mathematics test score)

	Coefficient	*Standard error*
Family Background		
Age (years)	17.825***	(3.160)
Female	–14.733***	(1.639)
Preprimary education (more than 1 year)	6.832***	(2.428)
School starting age	–3.869*	(2.030)
Grade repetition in primary school	–54.579***	(4.734)
Grade repetition in secondary school	–33.726***	(6.702)
Grade		
7th grade	–47.003***	(10.051)
8th grade	–19.213*	(10.242)
9th grade	–6.772	(6.896)
11th grade	–3.275	(5.236)
12th grade	11.949*	(6.398)
Living with		
Single mother or father	20.045***	(3.949)
Patchwork family	22.678***	(4.286)
Both parents	29.524***	(3.956)
Parents' working status		
Both full-time	–2.071	(2.911)
One full-time, one half-time	8.820***	(2.327)
At least one full time	15.926***	(2.891)
At least one half time	10.531***	(2.278)
Parents' job		
Blue collar, high skilled	1.481	(2.365)
White collar, low skilled	3.743*	(1.870)
White collar, high skilled	8.189**	(3.144)
Books at home		
11–25 books	6.760***	(2.290)
26–100 books	24.749***	(2.789)
101–200 books	34.232***	(3.161)
201–500 books	54.400***	(3.238)
More than 500 books	54.166***	(3.703)
Immigration background		
First-generation student	–11.447**	(4.442)
Nonnative student	–13.776**	(5.375)
Language spoken at home		
Other national dialect or language	–17.689**	(7.064)
Foreign language	–7.887***	(2.677)
Index of Economic, Social and Cultural Status (ESCS)	19.926***	(2.153)
Community location[8]		
Town (3,000–100,000)	9.101**	(3.323)
City (100,000–1,000,000)	16.951***	(3.989)
Large city with > 1 million people	13.939***	(4.929)

Continued on next page

Table 2 (continued)

	Coefficient	*Standard error*
School Resources		
Cumulative educational expenditure per student (1,000 $)[c]	0.270**	(0.103)
Shortage of instructional materials[s]		
Large shortage	–8.737**	(3.514)
No shortage	8.678***	(2.015)
Instruction time (minutes per week)	0.044***	(0.015)
Teacher education (*share at school*)[s]		
Fully certified teachers	7.699	(8.588)
Tertiary degree in pedagogy	10.211	(6.547)
Institutions		
Competition[c]		
Private operation (country share)	56.941***	(9.758)
Government funding (country share)	57.847***	(19.486)
Accountability		
External exit exams[c]	9.433	(9.055)
Assessments used for student retention/promotion[s]	11.744**	(4.320)
Monitoring of teacher lessons by principal[s]	6.785*	(3.442)
Monitoring of teacher lessons by external inspectors[s]	4.842*	(2.816)
Assessments used to compare school to district/nation[s]	4.188	(2.870)
Assessments used to group students[s]	–8.261**	(3.021)
Autonomy and its interaction with external exit exams[s]		
Autonomy in establishing starting salaries	–15.769***	(5.229)
External exit exams × Autonomy in establishing starting salaries	14.550*	(8.104)
Autonomy in formulating budget	–9.624	(6.901)
External exit exams × Autonomy in formulating budget	7.882	(8.478)
Autonomy in determining course content	–2.053	(5.435)
External exit exams × Autonomy in determining course content	11.504	(7.262)
Autonomy in hiring teachers	18.349*	(10.436)
External exit exams × Autonomy in hiring teachers	–24.723**	(11.796)
Constant	116.126**	(51.774)
Students	219,794	
Schools	8,245	
Countries	29	
R^2 (at student level)	0.340	

Source: Own calculations on the basis of Woessmann et al. (2009) using data from the Programme for International Student Assessment (PISA) 2003; the sample is OECD countries.

Notes: The table presents results from a least-squares regression weighted by students' sampling probability. The dependent variable is student's mathematics test score. Measures vary at the student level unless noted otherwise. Robust standard errors adjusted for clustering at the country level in parentheses.

[s] Observed at school level.

[c] Observed at country level.

***, **, and * represent significance levels of 1, 5, and 10 percent, respectively.

schools, and teacher evaluation in a cross-sectional model. In a country-level analysis of the PISA data, Brunello and Rocco (2013) find that an increased share of immigrant students has a small negative effect on the achievement level of native students.

Overall, socioeconomic factors contribute substantially to the cross-country variation in test scores.[4] These factors, however, are largely outside the influence of school systems—although not necessarily beyond the effects of other family, social, and redistributive policies.

Factors of the School System: Inputs and Institutions

Measures of school resources often fail to achieve economic and statistical significance in international education production functions, and sometimes even show counterintuitive coefficients. In Table 2, the point estimate on school spending is very small: An increase in cumulative educational expenditure per student until age 15 by $25,000, or one standard deviation, is associated with an increase in student achievement of less than 7 percent of a standard deviation. If class size as observed at the individual student level is added to the model, it has a counterintuitive positive coefficient—purportedly indicating that students achieve at higher levels in larger classes. Other variables have a more intuitive interpretation: for example, students perform worse in schools whose principal reports that the school's capacity to provide instruction is hindered by a shortage or inadequacy of instructional materials such as textbooks. Both weekly instruction time and measures of teacher education are positively associated with student achievement. Evidence from TIMSS, which provides more detailed teacher information from individual teacher background questionnaires, shows similar results (Woessmann 2003b). To the extent that schools with more resources in the tested grade also tended to have more resources in earlier grades, the coefficient estimates on resources capture not just the contemporaneous effect of resources in the specific grade, but the cumulative effect of resources over the previous grades.

In contrast, institutional features of school systems are strongly associated with student achievement in studies of this sort. Table 2 offers some examples, as do Woessmann (2003b, 2005b), Fuchs and Woessmann (2007), and Woessmann et al. (2009). In particular, measures of the extent of private school operation, government funding of schools, and different features of school accountability such as external exit exams, the use of assessments, and monitoring of lessons are positively related to student outcomes.[5] In addition, there is a tendency for school autonomy in different decision-making areas to be negatively related to student achievement in systems without external exit exams but to be unrelated or positively related in systems where external exit exams promote accountability (Woessmann 2005b). In

[4] Additional factors analyzed with international achievement data include gender differences (for example, Guiso et al. 2008; Fryer and Levitt 2010), relative age at school entry (for example, Bedard and Dhuey 2006), and peer effects (for example, Ammermueller and Pischke 2009).

[5] External exit exams reach statistical significance in a specification of the model of Table 2 that excludes the interactions with school autonomy. Results on the country-level variables in Table 2 are qualitatively the same in a two-step specification that first estimates Table 2 with country fixed effects and then regresses the coefficients captured on these fixed effects on the country-level variables.

a study of a variable not included in Table 2, Edwards and Garcia Marin (2015) find no significant association of country-aggregate student achievement in the PISA test with whether the right to education is included in a country's political constitution.

The results on instruction time, teacher education, and institutional effects provide a *prima facie* case for the relevance of school systems. Another piece of evidence for this relevance arises from adding school fixed effects to the estimation of an international education production function. Using PISA data, Freeman and Viarengo (2014) show that estimated school fixed effects are associated with observable school policies and teaching practices as well as with socioeconomic gradients. While they do not rule out nonrandom selection into schools as playing a role here, they interpret these results as indications of the potential importance of what schools do, as opposed to national or individual traits.

While most of the international achievement datasets are cross-sectional, Singh (2015) uses a longitudinal dataset that observes individual students at ages 5 and 8 in four developing countries. The findings show that the large cross-country learning gaps between low-performing Peru and high-performing Vietnam (apparent earlier in Table 1) are virtually nonexistent at school-entry age. They emerge over the first few school years in a way that is most consistent with large cross-country differences in the productivity of a school year (estimated from discontinuities in completed grades emerging from birth months in combination with enrollment thresholds), rather than with observed differences in socioeconomic background and time use. Again, these findings suggest that school systems have important effects.

Accounting for the Cross-Country Variation in Test Scores

As indicated, the model in Table 2 accounts for about one-third of the total student-level variation in the international model. This variation includes within-country variation as well as cross-country variation. The former is likely to include a component of random measurement error because of idiosyncrasies in individual performance on the testing day, a component that would cancel out at the national level.

So to what extent can family background factors, school resources, and institutions account for differences in student achievement across countries? To answer this question, we have to combine the large number of explanatory variables into a smaller number of factors. The student-level estimation of Table 2 provides one coefficient per variable: that is, it effectively forces the between-country associations of student achievement with the input factors to be the same as the within-country associations. We use these coefficient estimates on the individual variables in the model of Table 2 to combine the family background variables into one factor. That is, we simply calculate a linear combination that is the sum of the products of the individual variables times their respective coefficient estimates. We do the same for the school resource variables and the institutional variables. We then collapse the three combined input factors to the level of the 29 OECD country observations to obtain three aggregate country-level variables.

For descriptive purposes, we regress aggregate academic achievement on these three composite inputs for the 29 country-level observations. The share of the cross-country variance in achievement accounted for by the three input factors is 83 percent. That is, using the student-level model to additively and linearly combine

Table 3
Accounting for the Achievement Variance at the Country Level

	Family background	*School resources*	*Institutions*	*All three factors*
Accounted variance when only this factor is included in the model	0.504	0.181	0.533	0.834
Change in accounted variance when this factor is added to a model that already includes the other two factors	0.208	0.045	0.259	

Source: Author using data from the PISA 2003.
Notes: The table shows the share of the country-level variance in PISA 2003 mathematics test scores accounted for by the respective factor. Each factor represents a linear combination of individual variables using coefficient estimates from the student-level regression shown in Table 2, collapsed to the country level.

the input variables into three factors that can be collapsed to the country level, our simple international education production function descriptively accounts for more than four-fifths of the total cross-country variation in student achievement.

Table 3 breaks this explained variance in the country-level model down into components accounted for by the three groups of input factors. As in any regression analysis, the contribution of each factor depends on the other variables in the model. However, the role of family background factors appears substantial, contributing between 21 and 50 percent to the total cross-country variance in student achievement. By contrast, the contribution of school resources is much smaller, at 4 to 18 percent. Institutional differences again contribute importantly to the cross-country achievement variation, at 26 to 53 percent.[6]

Details of the extent to which the simple model accounts for the achievement of individual countries are shown in Table 4. For each country, the table shows how much of the country's difference from the international mean can be accounted for by each set of input factors.[7] For 14 of the 29 countries, the unaccounted-for residual achievement is less than 10 percent of a standard deviation. But for some examples, the model does not perform very well: in top-performing Finland, only 12.9 of the 44.5 percentage points of superior achievement (in standard deviations) are accounted for by the model. Differences from the international mean in family

[6]Compared to the models in Woessmann et al. (2009) and Hanushek and Woessmann (2011a), the model here excludes GDP per capita and class size, whose counterintuitive coefficients would hamper the interpretation of the accounting analysis. Including them would, in fact, reduce the separate contributions accounted for by the family background and school resource factors at the country level. Results are similar when including the imputation dummies contained in those models. It is debatable whether the model should include grade levels, individual grade repetition, and school starting age; however, results are similar when excluding these variables. The family background factor includes both individual student characteristics and genuine family factors; when separating the two, most of the country-level contribution goes to the genuine family factors and little to the student characteristics.

[7]To estimate the contribution of each input factor, we first run the country-level model on demeaned variables and then multiply the respective coefficient estimates with each country's value of the respective input factor. The contributions of the three input factors then sum to the predicted value (shown as "accounted difference" in Table 4) in this model.

Table 4
Accounting for Each Country's Difference from the International Mean

				Of which: accounted for by		
	Observed difference (1)	*Unaccounted difference* (2)	*Accounted difference* (3)	*Family background* (4)	*School resources* (5)	*Institutions* (6)
Finland	44.5	31.7	12.9	2.7	–1.3	11.5
Korea	42.0	14.3	27.7	13.0	5.6	9.1
Netherlands	38.4	–8.0	46.4	–3.4	–0.3	50.1
Japan	34.0	4.4	29.6	17.5	2.9	9.2
Canada	33.0	17.4	15.6	15.9	3.2	–3.5
Belgium	29.5	–11.8	41.3	–1.2	1.4	41.0
Switzerland	26.5	27.3	–0.8	–13.2	9.5	2.9
Australia	24.5	2.1	22.4	14.0	6.6	1.7
New Zealand	24.5	17.8	6.7	16.2	–3.0	–6.4
Czech Republic	16.4	2.1	14.3	16.1	–9.0	7.2
Iceland	15.1	–11.6	26.7	29.7	4.9	–7.9
Denmark	14.1	6.0	8.1	0.4	6.5	1.2
Sweden	10.0	5.5	4.5	5.9	–1.0	–0.4
United Kingdom	8.4	–9.1	17.5	13.0	2.7	1.8
Austria	5.5	5.7	–0.2	2.1	6.1	–8.5
Ireland	3.9	–15.0	18.8	–3.3	1.6	20.5
Germany	3.5	5.4	–1.9	–4.0	–0.8	2.8
Slovak Republic	–1.0	6.3	–7.3	4.2	–18.0	6.5
Norway	–4.3	–26.4	22.1	22.1	2.1	–2.1
Luxembourg	–6.3	–10.7	4.4	–25.5	19.3	10.6
Hungary	–9.3	–18.7	9.4	4.5	–5.4	10.4
Poland	–9.5	2.5	–12.0	–11.5	–8.1	7.6
Spain	–14.1	–2.7	–11.4	–4.8	–5.4	–1.2
United States	–16.1	–14.7	–1.4	2.3	9.1	–12.9
Portugal	–33.5	23.0	–56.5	–27.0	–2.8	–26.7
Italy	–33.9	–5.5	–28.3	2.7	3.6	–34.7
Greece	–55.1	–22.1	–33.0	–4.1	–3.0	–26.0
Turkey	–75.8	–4.4	–71.5	–31.7	–17.5	–22.3
Mexico	–114.8	–10.6	–104.2	–52.7	–9.9	–41.6

Notes: Each entry shows the country's test score difference from the international mean on the PISA 2003 mathematics test, expressed in student-level standard deviations. Column 1: actual difference. Column 2: difference not accounted for by a country-level regression of the actual test score difference on the three combined input factors (family background, school resources, institutions), each of which is measured as a linear combination of individual variables using coefficient estimates from the student-level regression of Table 2, collapsed to the country level. Column 3: difference accounted for by this country-level regression. Columns 4–6: difference accounted for by family background, school resources, and institutions, respectively. By constructions, columns 2 and 3 sum to column 1, and columns 4–6 sum to column 3.

background and school resources hardly contribute to this, but 11.5 percentage points are contributed by differences in the institutional setting, which in Finland include the existence of external exams, almost universal use of assessments for student retention, and widespread school autonomy over course content. For Korea, about two-thirds of the high relative achievement is accounted for by the model, and all three groups of input factors contribute to this, including a large share of privately operated schools, external exams, widespread monitoring of

teacher lessons, and universal course-content autonomy. For third-achieving Netherlands, the model in fact over-predicts its high achievement, and all of this is due to superior institutions—in particular, the largest share of privately operated schools, external exams, widespread course-content autonomy, and use of assessments for retention. At the lower end, most of the poor performance of Mexico and Turkey is accounted for by the model, in particular detrimental family background and institutions. The model does not do well at predicting US performance; institutions such as salary autonomy without external exit exams would predict the lower-than-average achievement level, but better family background and, in particular, abundant school resources would point the other way.

Inputs to the School System: Explorations into Causal Effects

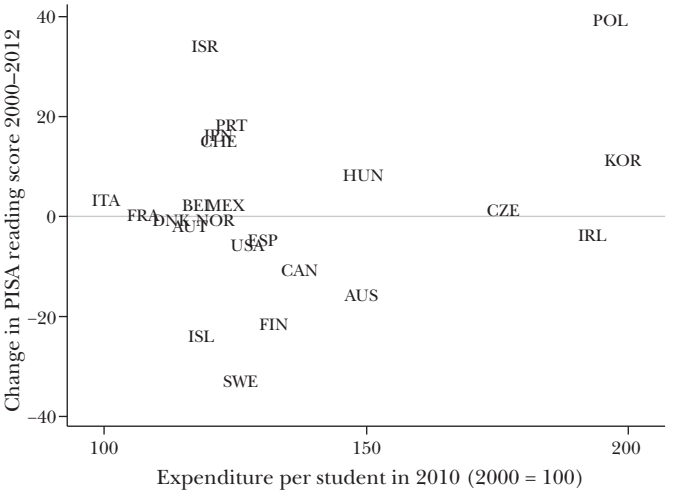
Inputs are clearly not exogenous to the education process. There may be reverse causation, for example, if educational systems assign additional resources to schools that serve low-achieving students, or if schools with poor student outcomes are induced to implement specific reforms. There may be bias from selection in that parents from low-achieving (or high-achieving) students tend to select into schools that offer specific resources for their children, or if high-performing schools have some ability to select high-achieving students. There may be omitted variables correlated with both inputs and outcomes, including country-level factors such as culture and valuation of education that may drive both inputs and learning effort, and also differences in preferences for high-quality education among parents or differences in motivation or ability of students. The direction of the net bias from such factors is not always obvious.

As a straightforward first step to exclude certain sources of bias when analyzing the possible effect of expenditure per student, one can ignore differences in the *levels* of expenditure and only use *changes* in average country expenditure over time as an explanatory variable in first-difference or differences-in-differences panel-type models. To the extent that sources of bias such as countries' cultures and parental background do not change significantly over time, they will no longer bias estimates based on changes in expenditure. In this spirit, in Gundlach, Woessmann, and Gmelin (2001), we calculated changes in expenditure and changes in test performance in several OECD countries over a 25-year period (1970–1994), finding that even substantial increases in real expenditure per student did not go hand in hand with improvements in student achievement.

More recently, the linking of the PISA tests over time allows for a direct comparison of spending changes to changes in achievement. As is directly obvious from Figure 3, changes in PISA performance from 2000 to 2012 are not systematically related to concurrent changes in expenditure per student. Countries with large spending increases do not show different achievement trends from countries that spend only little more.[8] While this analysis may be attenuated by the fact that changes

[8]The coefficient estimate on expenditure in the simple underlying first-differenced regression is insignificant. Similarly, using data from the first three PISA waves, the working-paper version of Hanushek

Figure 3
Changes in Educational Spending and in Student Achievement across Countries



Source: Hanushek and Woessmann (2015a) based on OECD data.
Notes: Scatter plot of the expenditure per student in 2010 relative to 2000 (constant prices, 2000 = 100) against change in PISA reading score, 2000–2012.

in expenditure may take some time to translate into actual inputs and then to affect student outcomes, the 25-year time horizon of the previous analysis should be able to reflect major effects. Of course, if other factors changed in a way correlated with both spending and test outcomes, looking for correlations between them—whether in levels or in differences—would still suffer from bias in these aggregate analyses.

Several studies have sought to use arguably exogenous variation in particular inputs by applying more elaborate identification methods. Here, we will discuss some of the evidence suggesting that smaller class sizes do not make much difference to educational outcomes but that more instruction time and higher teacher quality do make a difference.

Class Size

Most of the efforts that seek to uncover a causal effect of class size on test outcomes using international data turn to within-country variation. For example, in each school, natural cohort fluctuations in enrollment give rise to random class-size variation between adjacent grades (Hoxby 2000). In Woessmann and West (2006), we combine school fixed effects—which seek to eliminate between-school variation—with an approach that uses average class size in the school's grade as an instrumental variable, thus eliminating bias from sorting within a grade in a school. Applying this identification strategy to TIMSS data in 18 countries, we find significant beneficial

and Woessmann (2011b) reports insignificant negative coefficient estimates on expenditure per student in first-differenced and fixed-effects models.

effects of smaller classes in only two countries and can rule out large class-size effects in the majority of countries. Our estimates using this approach suggest that conventional cross-sectional estimates of class-size effects are substantially biased.[9]

These results are in line with results from a second quasi-experimental identification strategy suggested by Angrist and Lavy (1999) that exploits the existence of maximum class-size rules in many countries. Say that the maximum class size is 40, and that a certain grade has 120 students divided into three classes of 40 students each. If the grade rises to 121 students, the group is then divided into four classes—three of 30 students and one of 31 students. In this way, the rules give rise to discontinuous jumps in average class sizes whenever the enrollment in a grade in a school passes multiples of the maximum class size. Exploiting the induced class-size variation for ten European countries in a regression discontinuity design using TIMSS data, the results in Woessmann (2005a) rule out large causal class-size effects in all countries, with statistically significant but small effects in only two countries. Furthermore, the cross-country variation in estimated class-size effects in both studies is consistent with an interpretation that smaller classes have beneficial effects only in countries with relatively low teacher quality, as measured by relative teacher salary and teacher education.

The latter result is also confirmed in Altinok and Kingdon (2012), who apply yet another identification strategy to estimating class-size effects. To avoid bias from nonrandom sorting of students into schools and from unobserved student and family characteristics, they exploit the fact that the same students are tested in different subjects in TIMSS—math and science (sometimes in several specific domains). Using student fixed effects, they identify class-size effects from variation in class size between the two subjects for the same students (in countries where such variation exists). They find significant class-size effects in only 14 of 47 countries and even these are mostly small, confirming that class sizes play a limited role at best in understanding achievement differences in the international data.

Instruction Time

The length of school instruction time seems to play a more important role in educational achievement. For example, in an attempt to address omitted variable bias from unobserved individual subject-invariant characteristics such as underlying ability, motivation, or parental support, Lavy (2015) applies the within-student between-subject identification approach to estimate the effect of instruction time in the PISA 2006 data. The approach exploits the fact that different students have different instruction times in math, language, and science. He finds that instruction time has a significant positive effect on student achievement that is modest to large, suggesting that increasing instruction time by one hour per week would increase achievement by 6 percent of a standard deviation in OECD and Eastern European

[9] Applying the same instrumental variable strategy combined with school fixed effects—as well as an identification strategy based on restrictions placed on higher moments of the error distribution—to the PISA math data for the United States and the United Kingdom, Denny and Oppedisano (2013) find positive effects of larger classes, significant in the United Kingdom.

countries. This effect is only about half as large in developing countries. Furthermore, the effect of instruction time is larger in schools that have accountability measures such as using achievement data for evaluation, as well as in schools that have budgetary and personnel autonomy.

Rivkin and Schiman (2015) replicate the main finding of positive effects of instruction time in the within-student between-subject approach using the PISA 2009 data and confirm it in a model that uses within-subject variation in instructional time across grades within schools for identification. Furthermore, their results indicate that there are diminishing returns to instruction time and its effect is larger in classrooms with better environments as indicated by survey responses on questions about disruption, bullying, attendance, and other indicators of the quality of classroom environments.

Positive effects of instruction time are also confirmed in the setting of a specific education reform in Germany. The reform, which was implemented across German states at different times in the 2000s, reduced the length of the academic-track high school from nine to eight years. The reform did not change the curriculum requirements or the minimum required instruction time, so that the weekly instruction time increased in each grade. Pooling the 9th-grade samples of the extended PISA test in Germany from 2000 to 2009, Andrietti (2015) estimates the effects of the reform in a differences-in-differences framework that exploits the differing implementation years across states. Results suggest that an increase in weekly instruction time by one hour in both 8th and 9th grade increases achievement in the different subjects by between 2 and 3 percent of a standard deviation. Results are also confirmed in a "triple-difference" model that includes students in school types not affected by the reform as an additional control group.

A couple of studies have also shown that additional instruction time is related to smaller achievement gaps between different socioeconomic groups. Pooling several waves of TIMSS and PISA data, Schneeweis (2011) finds that instruction time is positively associated with the integration of immigrant students, with some models including country fixed effects so that effects are effectively identified from within-country changes over time. Pooling data from PISA and PIRLS for a differences-in-differences estimation with country fixed effects, Ammermueller (2013) finds that the achievement difference between students with different numbers of books at home is lower when instruction time is longer. There is also descriptive evidence that enrollment in early childhood education—that is, additional time before school—is related to reduced socioeconomic gradients and to better integration of migrant children (Schütz, Ursprung, and Woessmann 2008; Schneeweis 2011). Taken together, the results indicate that school instruction time can increase educational opportunities for students from disadvantaged backgrounds.

Teacher Quality

Teacher quality can be measured in a variety of ways. For example, Hanushek, Piopiunik, and Wiederhold (2014) use occupation-specific data on adult skills from the Programme for the International Assessment of Adult Competencies (PIAAC) to measure teacher skills in numeracy and literacy in 23 countries. Combining these aggregate measures of teacher skills with student-level PISA data, they estimate the

effect of teacher cognitive skills on international differences in student achievement, controlling among other factors for PIAAC-based estimates of parents' cognitive skills. Models with student fixed effects that exploit within-country variation between subjects suggest that teacher skills increase student achievement. Constructing a pseudo-panel from the PIAAC data using teachers' year of birth, they also exploit cross-country differences in how alternative job opportunities for women over time have attracted people with different skills into teaching. Bietenbeck, Piopiunik, and Wiederhold (2015) apply a within-student between-subject approach to a regional achievement test of 13 sub-Saharan African countries that includes subject-specific tests of teachers. They find a significant positive effect of teacher subject knowledge on student achievement that is complementary to access to subject-specific textbooks.

Measuring teacher quality by both absolute teacher salary and teachers' relative salary position in a country's income distribution, Dolton and Marcenaro-Gutierrez (2011) find that higher teacher quality is related to better student achievement using data from several TIMSS and PISA waves. The results are consistent with positive effects of recruiting higher ability individuals into teaching. Results are confirmed when adding country fixed effects, so that estimates are identified from (relatively short-term) fluctuations in teacher pay within countries.

Apart from studies of direct measures of teacher quality, recent evidence also indicates the relevance of teaching practices. Again applying within-student between-subject identification to circumvent bias from unobserved student characteristics, Schwerdt and Wuppermann (2011) show in the US TIMSS sample that for given levels of teaching methods, traditional lecture-style teaching is related to better student achievement compared to classroom problem-solving. Using the same estimation strategy on TIMSS data for the United States and nine advanced countries, Bietenbeck (2014) finds that traditional teaching practices are related to better overall skills, factual knowledge, and solving of routine problems, whereas modern teaching practices are related to better reasoning skills. After showing cross-country correlations of teaching practices with measures of social capital, Algan, Cahuc, and Shleifer (2013) apply a cross-sectional model with school fixed effects to TIMSS and PIRLS data to show that progressive practices of having students work in groups are positively related to student beliefs about cooperation and to student self-confidence.

Despite the result that resource inputs overall play a limited role, instruction time and certain dimensions of teacher quality do seem to matter for student achievement. More broadly, these findings suggest ways in which what school systems do are relevant for educational achievement. Moreover, looking into determinants of instruction time and teacher quality leads naturally to questions about the institutional framework of school systems that may frame how resources are used.

Institutional Structures of School Systems: Explorations into Causal Effects

An international comparative approach promises to be fruitful in studying the effects of educational institutions because institutional structures often do not vary

nearly as much within countries as they do across countries. Specific institutional features that have been found to matter for cross-country differences in student achievement include external exams, school autonomy, private competition, and tracking.

External Exams

In some countries, learning outcomes are assessed by curriculum-based external exit exams that have real consequences for students (Bishop 1997). A large literature has shown consistent positive associations between external exams and student achievement (Hanushek and Woessmann 2011a). However, such cross-country associations may be biased by unobserved country characteristics such as specific cultures. For example, a society that favors high educational achievement might both introduce external exams and also make efforts to induce students to study, and a positive correlation between external exams and student achievement does not show that the former has a causal effect on the latter.

There are several ways to explore whether these cultural effects are important in explaining the connection from exit exams to test scores. One approach is to look at variation in test scores and exams only within continents. If the international variation in test scores would have been biased by features more relevant in some continents than in others—say, if countries in Asia place a higher value on educational success than countries in other regions—then the coefficient on external exams will decline in such a model. However, in Woessmann (2003a), I find that the association between external exams and student achievement in the first two TIMSS waves is robust to the inclusion of continental fixed effects. Another approach looks at evidence across states within Germany and compares this with other OECD countries. German states differ in whether they have external exams or not, but are otherwise much more similar than OECD countries. However, in this mixture of PISA data on German states and other countries, students in systems with external exams have around 20 percent of a standard deviation higher achievement, and this association is statistically indistinguishable between the OECD country sample and the German state sample (Woessmann 2010). This result corroborates that the international association is unlikely to be driven by fundamental differences in culture, language, or other institutional settings that do not vary within Germany.

In yet another approach, Jürges, Schneider, and Büchel (2005) use the German TIMSS 1995 data in a differences-in-differences approach that exploits variation across subjects: specifically, in the relevant school tracks, most German states that have external exams have them in math but not in science. The identifying assumption of this model is that cross-state achievement differences would not differ between subjects in the absence of the external exam treatment. While smaller than their cross-sectional estimates, their differences-in-differences estimates are significant and substantial at between 13 and 26 percent of a standard deviation. If there are spillovers between subjects—for example, improved math knowledge due to external exams also facilitates students' learning in science—these estimates provide a lower bound for the full effect of external exams. Until the early 2000s, only seven of the 16 German states had external exams, but all but one have introduced them over the course of the 2000s. Lüdemann (2011) exploits the different

timing of the introduction of external exams across states and school types in a differences-in-differences approach using the German extended PISA waves from 2000 to 2006. The identifying assumption is that there would have been common trends in the absence of the external exam treatment. Results indicate significant positive effects of the introduction of central exit exams even in the short run.

While external exams direct incentives particularly on students, a way to incentivize teachers to focus on student outcomes is performance-related pay. Apart from showing a positive association of teacher pay with student achievement in PISA data, in Woessmann (2011), I find that teacher salary adjustments for outstanding performance are positively associated with student achievement across countries. The use of a country-level measure of teacher performance pay avoids bias from within-country selection, and results are robust to including continental fixed effects and to controlling for other forms of teacher salary adjustments that are not based on performance. An advantage of the cross-country approach is that it captures general-equilibrium effects such as sorting into the teaching profession and other long-run incentive effects, whereas short-term merit pay experiments capture only incentive effects, not selection effects.

School Autonomy

On the one hand, school autonomy may be conducive to student achievement in school systems with strong surrounding structures that ensure high common standards; on the other hand, school-based decision-making could hurt student achievement in low-performing systems that lack basic standards and local capacity. Cross-sectional evidence from international achievement tests concerning school autonomy has been mixed (Hanushek and Woessmann 2011a), but these studies may also be particularly plagued by identification issues.

To avoid bias from unobserved cross-country differences such as those arising from culture and other government institutions, in Hanushek, Link, and Woessmann (2013), we introduce the analytical approach of country panel analysis with country fixed effects. Because many countries have reformed their school systems to become more or less autonomous over time, we can exploit country-level variation over time by including country fixed effects that control for systematic, time-invariant differences across countries. While such panel analysis does not necessarily identify random variation, we show that prior achievement and prior GDP do not predict autonomy reforms. To avoid bias from within-country selection of students into autonomous schools and of schools to become autonomous, we aggregate the school autonomy measure to the country level, reflecting the average share of autonomous schools in a country.

Pooling the individual data of over one million students in 42 countries in the four PISA waves from 2000 to 2009, we find that school autonomy has a significant effect on student achievement, but this effect varies systematically with the level of economic and educational development: The effect is strongly positive in developed and high-performing countries but strongly negative in developing and low-performing countries. The estimates suggest that going from no to full autonomy over academic content would increase student achievement by 53 percent of a standard deviation in the highest-income country (Norway) and reduce student achievement by 55 percent of a standard deviation in the lowest-income country (Indonesia).

If part of the negative effect of school autonomy stems from a lack of accountability, these negative aspects should be eased in school systems where external exams provide comparative information on ultimate performance. Indeed, in Hanushek, Link, and Woessmann (2013), we find a significant positive interaction between changes in school autonomy and (initial) external exit exams—that is, introducing autonomy is more beneficial in school systems that have accountability through external exams.

The effects of school autonomy may also be interrelated with the management capacity of schools. Collecting data on school management practices in operations, monitoring, target setting, and people management in eight countries, Bloom, Lemos, Sadun, and Van Reenan (2015) find higher management quality to be related to better student achievement. While mostly focusing on specific national achievement datasets, they also report a positive correlation with average PISA scores across Italian and German regions. Furthermore, autonomous public schools score highly in terms of management quality. Interestingly, while their previous work suggested that most of the variation in management quality in other sectors is within-country, about half of the variance in management quality in the school sector is between countries, underlining the importance of cross-country analysis of institutional environments in school systems.

Private Competition

The extent to which schools are operated by public or private entities differs markedly across countries. For example, more than three-quarters of 15 year-old students in the Netherlands attend privately operated schools and more than 60 percent in Belgium and Ireland, but this share is below 10 percent in many other countries. Private school operation is largely independent of the funding of schools; for example, the average share of government funding of Dutch privately operated schools is the same (at 95 percent) as in public schools, a feature going back to constitutional provisions. Private school operation may be related to the extent of school autonomy, but again these are conceptually different issues: public schools can have substantial autonomy, and private schools can have limited autonomy. A key point is that competition from private alternatives may improve the performance of public schools as well, which may lift the achievement level systemwide.

Cross-country evidence indeed suggests a strong association of achievement levels with the share of privately operated schools (for example, Woessmann 2009), but identification issues are again obvious in cross-country analyses: low quality of the public school system may induce a political system to encourage private alternatives or parents to choose private alternatives, and other country features related to the supply of or demand for private schools may introduce omitted variable bias.

To identify exogenous variation in the share of private schools across countries, in West and Woessmann (2010), my coauthor and I argue that historical differences in Catholic versus Protestant denominations provide a natural experiment. In late 19th century, Catholic doctrine resisted the emerging nondenominational public school systems and spurred efforts to establish private schools in many countries. These efforts were most successful in countries with substantial shares of Catholic

populations but without a Catholic state religion. Therefore, the share of Catholics in a country's population in 1900 (interacted with an indicator as to whether Catholicism was the state religion) can be used as an instrumental variable for the share of privately operated schools in the 2003 PISA data.[10] The results suggest that a 10 percentage point increase in private school shares, induced by historical Catholic resistance to state schooling, leads to an increase in math achievement by at least 9 percent of a standard deviation. Much of this effect accrues to students in public schools, suggesting that most of the overall effect reflects benefits of private competition and parental choice. In addition to increasing achievement, private competition is also estimated to reduce total educational expenditure per student.

Tracking

Countries vary in the extent to which students are tracked into different school types by ability. No country has differing-ability schools in the early grades of primary school. Some countries such as Austria and Germany track students into different-ability schools as early as age 10. Many other countries have a comprehensive school system (although perhaps with some streaming within schools) through the end of high school. A common concern is that early tracking may increase inequality as lower-achieving groups are tracked into lower-ability schools, perhaps because of peer effects.

In Hanushek and Woessmann (2006), we suggest an identification strategy that compares achievement changes from primary to later schooling across tracked and untracked countries. Using country-level data for several pairs of PIRLS, TIMSS, and PISA achievement tests administered at the primary and secondary school levels in the context of a differences-in-differences model, we find that early tracking significantly increases the inequality in countries' achievement outcomes. We do not find a consistent effect of early tracking on the level of achievement, although most estimates tend to be negative. Interestingly, simple cross-sectional estimations do not indicate an association of tracking with educational inequality.

A variety of other results suggest that earlier tracking tends to raise the inequality of educational outcomes. Applying the same estimation strategy across grades to student-level PIRLS and PISA data, Ammermueller (2013) finds that early tracking and the number of tracked school types increase the effect of parental education on student achievement. Again using the same identification strategy to estimate the effect of tracking on the migrant–native achievement gap in a pooled micro dataset of all PIRLS, TIMSS, and PISA waves from 1995 to 2012, Ruhose and Schwerdt (2016) do not find that early tracking affects native and migrant students differently in general. However, they find a detrimental effect of early tracking on the relative achievement of first-generation migrants and the presumably less-integrated subgroup of second-generation migrant students who do not speak the host-country language at home. Piopiunik (2014) exploits a school reform in Bavaria that

[10] There is ample evidence that historically, Catholics have placed less emphasis on education than Protestants (for example, Becker and Woessmann 2009), which would bias the instrumental-variable model against finding beneficial effects of competition. Indeed, the current share of Catholics enters negatively in the second-stage model.

lowered the age of tracking between the two lowest-ability school types to estimate a "triple-difference" model using variation across three German PISA waves that allow a comparison of outcomes in the reformed system to pre-reform outcomes, to other German states, and to the non-treated highest-ability school type. Results suggest that earlier tracking reduced achievement in both low- and middle-track schools.

Conclusions

What explains the large international differences in student achievement? On a descriptive basis, a simple model of three combined factors of family background, school resources, and institutions is able to account for more than four-fifths of the total cross-country variation in student achievement. Family background and institutions contribute roughly equally to this exercise, whereas the contribution of school resources is quite limited—although the predictive power of the model varies across countries. Beyond these descriptive patterns, a growing literature uses quasi-experimental methods in an attempt to identify causal effects of school systems in the international test data, as well as different types of fixed effects models that aim to avoid certain sources of bias. Some patterns emerge from this literature. First, this work tends to confirm that resource inputs such as expenditure per student or class size appear to have limited effects on student achievement. Second, instruction time and measures of teacher quality do play a role. Third, a number of institutional features of school systems seem to contribute to the cross-country differences in student achievement. For example, external exit exams and competition from privately operated schools positively affect achievement levels. School autonomy has positive effects in developed countries and where external exit exams introduce accountability, but negative effects in developing countries. Early tracking into differing-ability schools seems to increase inequality in achievement without increasing achievement levels.

Clearly, the exploitation of the potential of international differences in student achievement to improve our understanding of educational processes is work in progress. In the future, increasing numbers of participating countries and an expanding number of waves of available international achievement tests will raise the scope of possible investigations. A useful direction for international testing efforts would be to conduct studies in many countries that are longitudinal at the student level. Existing causal identification strategies will be sharpened and new approaches developed. It may be especially useful to focus on interactions between the kinds of factors examined here: for example, little is known about the particular institutional settings that may strengthen the effectiveness of resource use.

As this work proceeds, it is perhaps useful to remember what is at stake. Levels and changes in educational achievement are a powerful determinant of output levels and economic growth. It has long been common to use average years of schooling in regressions that seek to explain economic growth. But average years of schooling may be a very noisy measure of actual educational achievement as measured by test scores. Thus, in Hanushek and Woessmann (2012, 2015a), we show that a model that includes only countries' average years of schooling and

their initial level of GDP per capita as predictors accounts for one-quarter of the total cross-country variation in growth rates in GDP per capita from 1960 to 2000 (or 2009). However, adding average scores on the international achievement tests between 1964 and 2003 to the model accounts for more than three-quarters of the variation in long-term growth rates of per-capita GDP—indeed, it renders the commonly used quantitative measure of years of schooling insignificant. Differences in math and science achievement can fully account both for the stunning growth performance of the East Asian miracle countries and for the disheartening growth performance of Latin American countries (Hanushek and Woessmann 2016).

In Hanushek and Woessmann (2012, 2015a), we report several econometric analyses that provide a *prima facie* case that the close and robust association of educational achievement with countries' long-run economic growth reflects a causal effect of population skills. To preclude simple reverse causation, we show that achievement tests before 1985 predict subsequent growth. To address potential bias from omitted factors such as differing economic institutions or cultures, we present instrumental-variable models that use only part of the skill variation that can be predicted from institutional differences in school systems; show that changes in test scores predict changes in growth; perform development accounting analyses that take parameter values from the micro literature; and report differences-in-differences models showing that immigrants educated in their home countries receive returns to their home-country cognitive skills on the US labor market, whereas immigrants from the same home countries schooled in the United States do not. Additional recent work on student achievement and international income differences is given in Kaarsen (2014), and we provide reviews in Hanushek and Woessmann (2008, 2011a). Within the United States, in Hanushek, Ruhose, and Woessmann (2015), we confirm an important role for educational achievement in explaining differences in GDP per capita across US states. At the individual level, performance on adult achievement tests is strongly associated with employment and earnings in each of the 23 countries analyzed in Hanushek, Schwerdt, Wiederhold, and Woessmann (2015). Murnane, Willett, and Levy (1995) and Chetty et al. (2011), among others, provide additional evidence on individual returns to educational achievement in the United States.

Of course, the implications of improved educational achievement go well beyond individual earnings and macroeconomic growth rates. Education is important for economic inequality and the transmission of inequality across generations (for example, Black and Devereux 2011). Education affects the education and health of children, own health, crime, and citizenship (for example, Lochner 2011). More broadly, a "capabilities approach" to welfare analysis in the style of Sen and Nussbaum (for example, Nussbaum and Sen 1993) emphasizes that education is an important determinant of the ability of people to develop their own capacities and in that sense to be able to exercise autonomy and choice in all aspects of life.

■ *Helpful comments from Mark Gertler, Gordon Hanson, Eric Hanushek, Enrico Moretti, Marc Piopiunik, Jens Ruhose, Timothy Taylor, Martin West, and Simon Wiederhold are gratefully acknowledged.*

References

Algan, Yann, Pierre Cahuc, and Andrei Shleifer. 2013. "Teaching Practices and Social Capital." *American Economic Journal: Applied Economics* 5(3): 189–210.

Altinok, Nadir, and Geeta Kingdon. 2012. "New Evidence on Class Size Effects: A Pupil Fixed Effects Approach." *Oxford Bulletin of Economics and Statistics* 74(2): 203–34.

Ammermueller, Andreas. 2013. "Institutional Features of Schooling Systems and Educational Inequality: Cross-country Evidence from PIRLS and PISA." *German Economic Review* 14(2): 190–213.

Ammermueller, Andreas, and Jörn-Steffen Pischke. 2009. "Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study." *Journal of Labor Economics* 27(3): 315–48.

Andrietti, Vincenzo. 2015. "The Causal Effects of Increased Learning Intensity on Student Achievement: Evidence from a Natural Experiment." Universidad Carlos III de Madrid, Working Paper, Economic Series 15-06. Madrid: Universidad Carlos III.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2): 533–75.

Becker, Sascha O., and Ludger Woessmann. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124(2): 531–96.

Bedard, Kelly, and Elizabeth Dhuey. 2006. "The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects." *Quarterly Journal of Economics* 121(4): 1437–72.

Bietenbeck, Jan. 2014. "Teaching Practices and Cognitive Skills." *Labour Economics* 30: 143–53.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2015. "Africa's Skill Tragedy: Does Teachers' Lack of Knowledge Lead to Low Student Performance?" CESifo Working Paper 5470.

Bishop, John H. 1997. "The Effect of National Standards and Curriculum-based Examinations on Achievement." *American Economic Review* 87(2): 260–64.

Black, Sandra E., and Paul J. Devereux. 2011. "Recent Developments in Intergenerational Mobility." In *Handbook of Labor Economics*, Vol. 4B, edited by Orley Ashenfelter and David Card, 1487–1541. Amsterdam: North Holland.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen. 2015. "Does Management Matter in Schools?" *Economic Journal* 125(584): 647–74.

Brunello, Giorgio, and Daniele Checchi. 2007. "Does School Tracking Affect Equality of Opportunity? New International Evidence." *Economic Policy* 22(52): 781–861.

Brunello, Giorgio, and Lorenzo Rocco. 2013. "The Effect of Immigration on the School Performance of Natives: Cross Country Evidence Using PISA Test Scores." *Economics of Education Review* 32(1): 234–46.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593–1660.

Cobb-Clark, Deborah A., Mathias Sinning, and Steven Stillman. 2012. "Migrant Youths' Educational Achievement: The Role of Institutions." *Annals of the American Academy of Political and Social Science* 643(1): 18–45.

Denny, Kevin, and Veruska Oppedisano. 2013. "The Surprising Effect of Larger Class Sizes: Evidence Using Two Identification Strategies." *Labour Economics* 23: 57–65.

Dolton, Peter, and Oscar D. Marcenaro-Gutierrez. 2011. "If You Pay Peanuts Do You Get Monkeys? A Cross-country Analysis of Teacher Pay and Pupil Performance." *Economic Policy* 26(65): 5–55.

Dustmann, Christian, Tommaso Frattini, and Gianandrea Lanzara. 2012. "Educational Achievement of Second-Generation Immigrants: An International Comparison." *Economic Policy* 27(69): 143–85.

Edwards, Sebastian, and Alvaro Garcia Marin. 2015. "Constitutional Rights and Education: An International Comparative Study." *Journal of Comparative Economics* 43(4): 938–55.

Foshay, Arthur W. 1962. "The Background and the Procedures of the Twelve-Country Study." In *Educational Achievement of Thirteen-year-olds in Twelve Countries: Results of an International Research Project, 1959–61*, edited by Arthur W. Foshay, Robert L. Thorndike, Fernand Hotyat, Douglas A. Pidgeon, and David A. Walker. Hamburg: Unesco Institute for Education.

Freeman, Richard B., and Martina Viarengo. 2014. "School and Family Effects on Educational Outcomes across Countries." *Economic Policy* 29(79): 395–446.

Fryer, Roland G., Jr., and Steven D. Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal: Applied Economics* 2(2): 210–240.

Fuchs, Thomas, and Ludger Woessmann. 2007.

"What Accounts for International Differences in Student Performance? A Re-examination Using PISA Data." *Empirical Economics* 32(2–3): 433–64.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, Math, and Gender." *Science* 320(5880): 1164–65.

Gundlach, Erich, Ludger Woessmann, and Jens Gmelin. 2001. "The Decline of Schooling Productivity in OECD Countries." *Economic Journal* 111(471): C135–C47.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24(3): 1141–77.

Hanushek, Eric A. 2002. "Publicly Provided Education." In *Handbook of Public Economics*, Vol. 4, edited by Alan J. Auerbach and Martin Feldstein, 2045–2141. Amsterdam: North Holland.

Hanushek, Eric A., and Dennis D. Kimko. 2000. "Schooling, Labor Force Quality, and the Growth of Nations." *American Economic Review* 90(5): 1184–1208.

Hanushek, Eric A., Susanne Link, and Ludger Woessmann. 2013. "Does School Autonomy Make Sense Everywhere? Panel Estimates from PISA." *Journal of Development Economics* 104: 212–32.

Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann. 2013. *Endangering Prosperity: A Global View of the American School*. Washington, DC: Brookings Institution Press.

Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann. 2014. "U.S. Students from Educated Families Lag in International Tests." *Education Next* 14(4): 8–18.

Hanushek, Eric A., Marc Piopiunik, and Simon Wiederhold. 2014. "The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance." NBER Working Paper 20727.

Hanushek, Eric A., Jens Ruhose, and Ludger Woessmann. 2015. "Human Capital Quality and Aggregate Income Differences: Development Accounting for U.S. States." NBER Working Paper 21295.

Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. "Returns to Skills around the World: Evidence from PIAAC." *European Economic Review* 73: 103–130.

Hanushek, Eric A., and Ludger Woessmann. 2006. "Does Educational Tracking Affect Performance and Inequality? Differences-in-differences Evidence across Countries." *Economic Journal* 116(510): C63–C76.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46(3): 607–68.

Hanushek, Eric A., and Ludger Woessmann. 2011a. "The Economics of International Differences in Educational Achievement." In *Handbook of the Economics of Education*, Vol. 3, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 89–200. Amsterdam: North Holland.

Hanushek, Eric A., and Ludger Woessmann. 2011b. "How Much Do Educational Outcomes Matter in OECD Countries?" *Economic Policy* 26(67): 427–91.

Hanushek, Eric A., and Ludger Woessmann. 2011c. "Sample Selectivity and the Validity of International Student Achievement Tests in Economic Research." *Economics Letters* 110(2): 79–82.

Hanushek, Eric A., and Ludger Woessmann. 2012. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17(4): 267–321.

Hanushek, Eric A., and Ludger Woessmann. 2015a. *The Knowledge Capital of Nations: Education and the Economics of Growth*. Cambridge, MA: MIT Press.

Hanushek, Eric A., and Ludger Woessmann. 2015b. *Universal Basic Skills: What Countries Stand to Gain*. Paris: Organisation for Economic Co-operation and Development.

Hanushek, Eric A., and Ludger Woessmann. 2016. "Knowledge Capital, Growth, and the East Asian Miracle." *Science* 351(6271): 344–45.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics* 115(3): 1239–85.

IEA. 2016. "Brief History of IEA: 55 Years of Educational Research." Amsterdam: International Association for the Evaluation of Educational Achievement. http://www.iea.nl/brief_history.html.

Jerrim, John, and John Micklewright. 2014. "Socio-economic Gradients in Children's Cognitive Skills: Are Cross-country Comparisons Robust to Who Reports Family Background?" *European Sociological Review* 30(6): 766–81.

Jürges, Hendrik, Kerstin Schneider, and Felix Büchel. 2005. "The Effect of Central Exit Examinations on Student Achievement: Quasi-experimental Evidence from TIMSS Germany." *Journal of the European Economic Association* 3(5): 1134–55.

Kaarsen, Nicolai. 2014. "Cross-Country Differences in the Quality of Schooling." *Journal of Development Economics* 107: 215–24.

Lavy, Victor. 2015. "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries." *Economic Journal* 125(588): F397–F424.

Lee, Jong-Wha, and Robert J. Barro. 2001. "Schooling Quality in a Cross-section of Countries." *Economica* 68(272): 465–88.

Lochner, Lance. 2011. "Nonproduction Benefits of Education: Crime, Health, and Good Citizenship." In *Handbook of the Economics of Education*, Vol. 4, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 183–282. Amsterdam: North Holland.

Lüdemann, Elke. 2011. "Intended and Unintended Short-run Effects of the Introduction of Central Exit Exams: Evidence from Germany." In Elke Lüdemann, *Schooling and the Formation of Cognitive and Non-cognitive Outcomes*. ifo Beiträge zur Wirtschaftsforschung 39. Munich: ifo Institut.

Mullis, Ina V. S., Michael O. Martin, Pierre Foy, and Alka Arora. 2012. *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77(2): 251–66.

Nussbaum, Martha C., and Amartya Sen, eds. 1993. *The Quality of Life*. Oxford University Press.

OECD. 2013. *PISA 2012 Results: What Students Know and Can Do—Student Performance in Mathematics, Reading and Science*, Vol 1. Paris: Organisation for Economic Co-operation and Development.

Piopiunik, Marc. 2014. "The Effects of Early Tracking on Student Performance: Evidence from a School Reform in Bavaria." *Economics of Education Review* 42: 12–33.

Ripley, Amanda. 2013. *The Smartest Kids in the World—And How They Got That Way*. New York: Simon & Schuster.

Rivkin, Steven G., and Jeffrey C. Schiman. 2015. "Instruction Time, Classroom Quality, and Academic Achievement." *Economic Journal* 125(588): F425–F448.

Ruhose, Jens, and Guido Schwerdt. 2016. "Does Early Educational Tracking Increase Migrant–Native Achievement Gaps? Differences-in-Differences Evidence across Countries." *Economics of Education Review* 52: 134–54.

Schneeweis, Nicole. 2011. "Educational Institutions and the Integration of Migrants." *Journal of Population Economics* 24(4): 1281–1308.

Schütz, Gabriela, Heinrich W. Ursprung, and Ludger Woessmann. 2008. "Education Policy and Equality of Opportunity." *Kyklos* 61(2): 279–308.

Schwerdt, Guido, and Amelie C. Wuppermann. 2011. "Is Traditional Teaching Really All That Bad? A Within-Student Between-Subject Approach." *Economics of Education Review* 30(2): 365–79.

Singh, Abhijeet. 2015. "Learning More with Every Year: School Year Productivity and International Learning Divergence." Presented at the CESifo Area Conference on the Economics of Education, September 11–12, 2015. Available at: http://www.cesifo-group.de/de/ifoHome/events/Archive/conferences/2015/09/2015-09-11-ee15-Hanushek/Programme.html.

West, Martin R., and Ludger Woessmann. 2010. "'Every Catholic Child in a Catholic School': Historical Resistance to State Schooling, Contemporary Private Competition and Student Achievement across Countries." *Economic Journal* 120 (546): F229–F255.

Woessmann, Ludger. 2003a. "Central Exit Exams and Student Achievement: International Evidence." In *No Child Left Behind? The Politics and Practice of School Accountability*, edited by Paul E. Peterson and Martin R. West, 292–323. Washington, D.C.: Brookings Institution Press.

Woessmann, Ludger. 2003b. "Schooling Resources, Educational Institutions, and Student Performance: The International Evidence." *Oxford Bulletin of Economics and Statistics* 65(2): 117–70.

Woessmann, Ludger. 2005a. "Educational Production in Europe." *Economic Policy* 20(43): 446–504.

Woessmann, Ludger. 2005b. "The Effect Heterogeneity of Central Exams: Evidence from TIMSS, TIMSS-Repeat and PISA." *Education Economics* 13(2): 143–169.

Woessmann, Ludger. 2009. "Public–Private Partnerships and Student Achievement: A Cross-Country Analysis." In *School Choice International: Exploring Public–Private Partnerships*, edited by Rajashri Chakrabarti and Paul E. Peterson, 13–45. Cambridge, MA: MIT Press.

Woessmann, Ludger. 2010. "Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries." *Journal of Economics and Statistics* 230(2): 234–70.

Woessmann, Ludger. 2011. "Cross-Country Evidence on Teacher Performance Pay." *Economics of Education Review* 30(3): 404–18.

Woessmann, Ludger, Elke Luedemann, Gabriela Schuetz, and Martin R. West. 2009. *School Accountability, Autonomy, and Choice around the World*. Cheltenham, UK: Edward Elgar.

Woessmann, Ludger, and Martin R. West. 2006. "Class-Size Effects in School Systems around the World: Evidence from Between-Grade Variation in TIMSS." *European Economic Review* 50(3): 695–736.