The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

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

Ludger Woessmann

CESIFO WORKING PAPER NO. 5951
CATEGORY 5: ECONOMICS OF EDUCATION
JUNE 2016

ISSN 2364-1428

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 5951*

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

Abstract

Students in some countries do far better on international achievement tests than students in other countries. Is this all due to differences in what students bring with them to school – socio-economic background, cultural factors, and the like? Or do school systems make a difference? This essay argues that differences in features of countries' school systems, and in particular their institutional structures, account for a substantial part of the cross-country variation in student achievement. It first documents the size and cross-test consistency of international differences in student achievement. Next, it uses the framework of an education production function to provide descriptive analysis of the extent to which different factors of the school system, as well as factors beyond the school system, account for cross-country achievement differences. Finally, it covers research that goes beyond descriptive associations by addressing leading concerns of bias in cross-country analysis. The available evidence suggests that differences in expenditures and class size play a limited role in explaining cross-country achievement differences, but that differences in teacher quality and instruction time do matter. This suggests that what matters is not so much the amount of inputs that school systems are endowed with, but rather how they use them. Correspondingly, international differences in institutional structures of school systems such as external exams, school autonomy, private competition, and tracking have been found to be important sources of international differences in student achievement.

JEL-Codes: I210, H520, L380, J240, D020.

Keywords: student achievement, international comparison, education production function, schools, education, institutions, external exams, autonomy, competition, private schools, tracking, educational expenditure, teachers, instruction time, TIMSS, PISA.

Ludger Woessmann
Ifo Institute – Leibniz Institute for
Economic Research
at the University of Munich
Poschingerstrasse 5
Germany – 81679 Munich
woessmann@ifo.de

June 7, 2016
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.

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. When put on the same metric, 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. Why do students in different countries achieve at such vastly different levels? Apart from differences in socio-economic and cultural backgrounds, do differences in the organization and governance of school systems play a role?

This essay presents evidence from research investigating the international differences in student achievement that indicates that school systems are indeed important. Given the importance of family inputs in education production, people often wonder what school systems can accomplish. In fact, all of the above-mentioned high-achieving countries spend considerably *less* on schools per student than the United States (OECD (2013)), questioning the importance of school inputs. But school systems are not just about how much is spent, but also about how resources are used, which is importantly shaped by their institutional framework.

Different countries have certainly adopted very different institutional structures for their school systems. For example, in contrast to the United States and several other countries, students in many 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 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. We will present evidence that all these institutional differences are systematically related to differences in student achievement across countries.

A simple interpretation is that the institutional framework of a school system – including the extent of accountability, autonomy, and competition – provides varying incentives that affect student outcomes. If a country's institutions ensure that stakeholders in the system have incentives to focus on improving student outcomes, the system may be steered toward efficient and effective investments and operations.

The international perspective provides a unique opportunity to estimate such institutional effects. Institutional features of school systems differ much more markedly across than within countries. In fact, institutions such as national accountability systems or tracking regimes often vary only slightly or not at all within countries. As the director of the first pilot project of 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)). In addition, international data allow us to estimate whether specific effects differ systematically across countries or whether they reflect general results. System-level analyses will also capture general equilibrium effects that emerge, for example, if the presence of private alternatives changes what public schools do or if the introduction of a country-wide accountability system changes the composition of the teaching force. Such effects usually elude school-level, and in particular short-term experimental, analyses within a school system. To the extent that the selection of students with specific backgrounds into particular schools cancels out at the national level, aggregating indicators of institutional features to the country level also circumvents important selection biases.

However, these advantages of the cross-country comparative approach come with some built-in limitations. The biggest issue is that identification of causal effects raises particular challenges in an international setting. Most importantly, countries may differ from one another 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. As a consequence, identification is difficult in international data, restricting inferences to specific

topics. 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 provide descriptive analysis of the extent to which different factors of the school system, as well as factors beyond the school system, are associated with cross-country achievement differences. In the final part, it focuses on research that attempts to go beyond conditional correlations by addressing some sources of potential bias in cross-country analysis. It concludes by pointing out the major implications of educational achievement for the prosperity of individuals and nations.

The main theme that emerges from this discussion is that school systems are indeed important. The role of resource inputs seems limited, but differences in instruction time and teacher quality do matter, suggesting that what school systems do is indeed relevant. 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.

How Large and Consistent Are International Differences in Student Achievement?

By now, large-scale international testing of student achievement has more than half a century of history, and a large number of 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 et al. (2012)). Following a pilot project in 1959-61, the IEA conducted its first international math study of eleven countries in 1964 (Figure 1). The first science and reading studies occurred in the early 1970s, and a second round in each subject was performed in the 1980s and early 1990s. Subsequently, the IEA moved its testing in math, science, and reading to regular testing cycles: 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 psychometrically over time, too. But there is no direct link between the scales of the three testing regimes or across time with the older tests.

Figure 2 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.[2] The transformation of the TIMSS data to the PISA scale follows the empirical calibration method suggested by Hanushek and Woessmann (2015b, Annex B).

The cross-country differences in knowledge among same-aged students are in some cases extremely large. Remember, as a general rule of thumb, average learning gains on most national and international tests during one year are equal to between one-quarter and one-third of a

[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 (see Hanushek and Woessmann (2011a), Table 2; Hanushek and Woessmann (2015b), chapter 4; Hanushek et al. (2015)). The IEA has also conducted studies in other subjects such as foreign languages, civic education, and computer literacy.

[2] This standardization was done in 2003 in math and in 2006 in science. On average across OECD countries, the within-country standard deviation is 92 in math and 93 in science (OECD (2013)).

standard deviation, which would be a difference in score of 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 42-73 percent of a standard deviation, or roughly between one-and-a-quarter to more than two times what students usually learn during one year.[3] At the other end, the average difference of U.S. 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 Figure 2, it is important to focus on overall scores, not on ranks. The achievement differences between similarly ranked countries are often quite small, at a few percentage points of a standard deviation. In fact, many of these close differences do not reach statistical significance. For example, in the PISA 2012 math test, the standard error of the achievement level in most countries is 2-3 percent of a standard deviation, with extremes ranging from 0.8 to 4.8 standard deviations. As a consequence, the achievement levels of most participating countries are usually not statistically significantly different from their closest 1-3 neighbors to the top and the bottom; at the extreme, Portugal's achievement at rank 31 does not differ significantly from ranks 25-37 in the PISA 2012 math test (OECD (2013)). Therefore, what matters is not so much the specific rank of a country, but the absolute achievement differences – which turn out to be immense in the full set of participating countries.

The presented mean differences sometimes hide important differences in the shape of the overall distribution of achievement in a country. Figure 3 displays the achievement distribution on the PISA 2012 math test for the United States and three selected countries with relatively high performance. The U.S. distribution is shifted to the left and slightly more left-steep compared to the overall student distribution in OECD countries, 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 (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

[3] Performance was even higher in the Shanghai region of China, which performed more than one standard deviation above the United States.

students from disadvantaged backgrounds. For example, the 25th, 50th, and 75th percentiles of the U.S. distribution on the PISA 2012 math test are all between 13 and 15 points below the OECD average of the respective percentiles. Hanushek, Peterson, and Woessmann (2013) 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, Hanushek, Peterson, and Woessmann (2014) show that the ranking of U.S. students from better-educated families when compared to students from better-educated families in other countries is not much different from the ranking of U.S. students from less-well-educated families when compared to students from less-well-educated families abroad.

Consistency across Different Tests

The measurement of educational achievement is subject to many psychometric and measurement choices, particularly in an internationally comparative context. In fact, the two major testing cycles at the secondary school level have somewhat different foci. For example, the target population of the TIMSS test is 8th-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. Such different emphases raise the question of how consistent measured achievement is across different testing approaches.

To gauge the sensitivity of 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. Figure 4 plots the average math achievement in the two tests against one another. Despite the differences in timing, target populations, and conceptual approaches, all countries align very closely around the 45-degree line. In fact, 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 the tests do not just capture noise and that issues of particularities of 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 non-response. While

sampling was devised to be representative of the student population in each participating country, some countries do no longer have universal enrollment at age 15, when students are tested in PISA. In addition, non-random differences in patterns of sample exclusions (e.g., for handicapped children) and non-response 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 the cross-sectional assessment of countries at a point in time is reasonably straightforward, assessing the changes in country performance over time is harder. The early international tests, in particular, constitute separate testing incidents without links across different tests. Hanushek and Woessmann (2012, 2015a) use an empirical calibration method to put all international tests from 1964-2003 on a common standardized scale. Their 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 is consistent over time.

Still, several countries do show either significant improvements or declines over time. Figure 5 provides a stylized depiction of the achievement trends observed in selected example countries from 1964-2012, based on the available testing data. The more limited variation in early decades likely reflects the lower frequency of testing before 2000. However, the figure shows both substantial cross-sectional differences across countries, and also that some countries 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

manage to improve substantially over time, other well-off countries deteriorated, such as Norway during the 1990s, Sweden during the 2000s, and Finland in recent years. It appears that educational achievement levels of countries are generally consistent over time, but also that they are not set in stone and can be mutable.

Descriptive Patterns Using an Education Production Function

The remainder of this paper addresses the task of understanding the sources of the observed international differences in student achievement. 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. A fundamental challenge in all analyses of this kind is that these inputs are not necessarily exogenous to student achievement, and are very likely to be biased by omitted variables, selection, and reverse causation. But 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

To obtain evidence on whether a specific feature of the school system is consistently and systematically related to student outcomes, one cannot rely on simple comparisons of two countries – e.g., the United States and Finland as one of the international top performers. The production of education is a complex process that includes a multitude of input factors. As any two countries will differ in many underlying factors, it is impossible to ascertain which factor is responsible for the achievement difference. Therefore, to understand possible sources of international differences in student achievement, one has to study the achievement variation across as many countries as possible.

As a general framework for such an analysis, economists have used the concept of an education production function, which models the output of the education process as a function of different input factors (e.g., Hanushek (1986, 2002)). To study international differences in student achievement and the possible role of school systems therein, we can thus estimate international education production functions of the following form using cross-country data:

$$\textit{student achievement} = \beta_1 \textit{ family background} + \beta_2 \textit{ school resources} + \beta_3 \textit{ institutions} + \varepsilon \qquad (1)$$

To give structure to possible determinants of international achievement differences, 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 two aspects: how much resource inputs the systems use and the institutional structures in which they use them. This basic model can be straightforwardly extended to include interaction effects between different 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 (e.g., Bishop (1997); Hanushek and Kimko (2000); Lee and Barro (2001)).[4] More recent studies also use country-level data to study, for example, the correlates of gender equality in achievement (Guiso et al. (2008); Fryer and Levitt (2010)).

However, starting with Woessmann (2003b), a number of studies have used the micro data from international achievement tests at the student level to estimate extensive multivariate cross-country education production functions.[5] Because these microeconometric 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 1 provides an example of a basic cross-sectional estimation of an international education production function.[6] 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

[4] Early studies that used data from international student achievement tests to estimate education production functions within individual countries include Heyneman and Loxley (1983) and Toma (1996).

[5] Other examples using individual-level international student achievement data in cross-country regressions include Woessmann (2005b), Fuchs and Woessmann (2007), Brunello and Checchi (2007), Woessmann et al. (2009), Schneeweis (2011), and Ammermueller (2013).

[6] 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.

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 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 (see Appendix A of Woessmann et al. (2009) for details). Many factors in all three groups are statistically and economically significantly associated with student achievement. Descriptively, this simple model accounts for 34 percent of the achievement variance at the individual student level.

Factors beyond the School System: Family, Socio-economic, and Cultural Background

Some of the personal background characteristics that have meaningful and statistically significant magnitudes in Table 1 include student characteristics such as age, gender, and participation in early childhood education, along with several 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 vs. 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 socio-economic backgrounds in the international tests. The first type looks at how much socio-economic background contributes to country-level differences in educational outcomes. For example, Bishop (1997), Hanushek and Kimko (2000), and Lee and Barro (2001) all report strong cross-country relations of student achievement levels with parental income and education, usually proxied by per-capita GDP and adult education.

The second type of analysis compares the within-country association of socio-economic factors with student achievement, sometimes referred to as socio-economic gradients, across countries. For example, Schuetz, Ursprung, and Woessmann (2008) 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. They show that family background effects are systematically

larger in countries with early tracking and less extensive pre-primary education systems.[7] Jerrim and Micklewright (2014) use PISA and PIRLS data to analyze the extent to which cross-country comparisons of socio-economic gradients are affected by differences in reporting errors.

Several studies have focused on the achievement of children with an immigration background. 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.[8] They also find that children of Turkish immigrants perform better in most host countries than Turkish children in Turkey.[9] 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, socio-economic factors contribute substantially to the cross-country variation in test scores.[10] 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

International education production functions also analyze two aspects of the school system. The first aspect is how much resource inputs are available, which – in terms of the education production function – depicts movements along the production function. The second aspect is institutional features of the school system that determine how the inputs are used, equivalent to shifts in the production function.

On the first aspect, measures of school resources often fail to achieve economic and statistical significance in international education production functions or even show counterintuitive coefficients (e.g., Hanushek and Kimko (2000); Woessmann (2003b)), although

[7] 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.

[8] This work builds on prior work by Entorf and Minoiu (2005), Schnepf (2007), and Schneeweis (2011). For analysis of immigrant achievement in international tests in specific countries see, for example, Lüdemann and Schwerdt (2013) and Cattaneo and Wolter (2015).

[9] 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 schools, and teacher evaluation in a cross-sectional model.

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

there are exceptions (e.g., Lee and Barro (2001)). In the model of Table 1, the point estimate on school spending is very small (even if potentially inflated by the omission of a country's GDP per capita): 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 by 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.

Apart from material inputs, the model also includes variables capturing more qualitative aspects of the teaching process such as instruction time and proxies for teacher quality such as teacher education and experience. In our model using PISA data, 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.

On the second aspect, several institutional features of school systems are strongly associated with student achievement in the model of Table 1 (see also Woessmann (2003b); Fuchs and Woessmann (2007); 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 (see also Bishop (1997); Woessmann (2005b)), the use of assessments, and monitoring of lessons are positively related to student outcomes.[11] 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 unrelated or positively

[11] External exit exams reach statistical significance in a specification of the model of Table 1 that excludes the interactions with school autonomy. Results on the country-level variables in Table 1 are qualitatively the same in a two-step specification that first estimates Table 1 with country fixed effects and then regresses the coefficients captured on these fixed effects on the country-level variables.

related in systems where external exit exams promote accountability (see also Woessmann (2005b)).[12]

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 socio-economic gradients. While they do not rule out non-random selection into schools as playing a role, 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 specific 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 Figure 2) 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 socio-economic background and time use. Again, these findings suggest that school systems have important effects.

Accounting for the Cross-Country Variation in Test Scores

To what extent can family background factors, school resources, and institutions account for differences in student achievement *across countries*? As indicated, the model in Table 1 accounts for about one-third of the total student-level variation in the international model. However, 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.

[12] Using country-aggregate PISA data, Edwards and Garcia Marin (2015) find no significant association of student achievement with whether the right to education is included in a country's political constitution.

To estimate the contribution of the model to the country-level variation in test scores, we have to combine the large number of explanatory variables into a small number of factors. The student-level estimation of Table 1 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 the coefficient estimates on the individual variables in the model of Table 1 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 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-fifth of the total cross-country variation in student achievement.

Table 2 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. Regardless of order, 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.[13]

The fact that the simple model accounts for most of the variation at the country level is visualized by the close alignment of countries around the regression line in Figure 6, which plots

[13] 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.

the actual country-average test scores against the test scores predicted by the model. There is no high-achieving country for which the model would predict low achievement, or vice versa. For each country, the vertical distance to the regression line depicts the residual achievement not explained by the model. For example, Finland, Switzerland, and Portugal perform better than predicted by the model, whereas Norway, Greece, and Hungary perform worse.

Details of the extent to which the simple model accounts for the achievement of individual countries are shown in Table 3. For each country, the table shows how much each set of input factors accounts for in terms of its difference from the international mean.[14] For 14 of the 29 countries, the unaccounted residual achievement is less than 10 percent of a standard deviation. But for some specific examples such as top-performing Finland, the model does not perform very well in accounting for the high achievement. 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 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, and 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 U.S. 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 in fact point the other way.

[14] 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 3) in this model.

Inputs to the School System: Explorations into Causal Effects

To go beyond descriptive analysis, this section and the following one turn to evidence that aims to address some of the major sources of bias in cross-sectional estimates of the effects of specific inputs and institutions, respectively.

Resource Inputs

The fundamental challenge to the interpretation of the descriptive evidence is that inputs are clearly not exogenous to the education process. All sorts of standard sources of endogeneity may affect the associations. There may be reverse causation, for example, if educational systems assign additional resources to schools that serve low-achieving students in an attempt to remediate low achievement, 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 particularly 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, but also differences in preferences for high-quality education among parents or differences in motivation or ability of students. The direction of the bias from these factors and others 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, Gundlach, Woessmann, and Gmelin (2001) 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 on psychometrically linked tests. As is directly

obvious from Figure 7, 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. The coefficient estimate on expenditure in the simple underlying first-differenced regression is insignificant, and without the apparent outlier Poland, the point estimate is negative.[15] While this analysis may be attenuated by the fact that changes 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 any 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.

Thus, several studies have sought to use arguably exogenous variation in a particular resource input, class size, by applying more elaborate identification methods. Most of the efforts that seek to uncover a causal effect of class size on test outcomes using international data turn to some kind of within-country micro-level variation. For example, in each school, natural cohort fluctuations in enrollment give rise to random class-size variation between adjacent grades (Hoxby (2000)). To identify such variation, Woessmann and West (2006) combine school fixed effects – which eliminate any between-school variation – with an instrumental-variable approach that instruments actual class size by the average class size in the grade in the school, thus eliminating bias from sorting within a grade in a school. Applying this identification strategy to TIMSS data in 18 countries, they find significant beneficial effects of smaller classes in only two countries, and can rule out large class-size effects in the majority of countries. Their estimates using this approach suggest that conventional cross-sectional estimates of class-size effects are substantially biased.[16]

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

[15] Similarly, using data from the first three PISA waves, the working-paper version of Hanushek and Woessmann (2011b) reports insignificant negative coefficient estimates on expenditure per student in first-differenced and fixed-effects models.

[16] 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 surprising positive effects of larger classes, significant in the United Kingdom.

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 non-random sorting of students into schools and from (subject-invariant) 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 the result that class sizes play a limited role at best in understanding achievement differences in the international data.

There is less research on the impact of other resource inputs in the international data. Falck, Mang, and Woessmann (2015) use the within-student between-grade identification with student fixed effects to estimate the effect of classroom computer use in the TIMSS data. They find a null effect on average, but this combines differential effects of different types of computer use: positive effects of using computers to look up information and negative effects of using computers to practice skills. Across countries, the effects are mostly confined to developed countries and not strongly visible in developing countries.

Instruction Time

At times, the disappointing evidence on resource inputs has led observers to question the role that schools play in determining student achievement. Are the international differences in student achievement all due to differences in what students bring with them to school, or do

schools and school systems matter after all? An answer can come from testing whether variation in the length of school instruction time matter for student outcomes.

In an attempt to address omitted variable bias from unobserved subject-invariant individual characteristics such as underlying ability, motivation, or parental support, Lavy (2015) applies the within-student between-subject identification approach to estimating 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 countries. However, 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 across grades within schools for identification. This result weakens concerns about possible remaining biases from subject-specific unobserved factors that might correlate with instruction time and with achievement. 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 in different 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 total curriculum requirements or the total minimum required instruction time, so that the weekly instruction time increased in each grade. Pooling the 9^{th}-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 8^{th} and 9^{th} 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 socio-economic 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. Taken together, the results indicate that instruction time can increase educational opportunities for students from disadvantaged backgrounds.[17]

Teacher Quality

The pattern of results that resource inputs have little impact on student achievement, but that instruction time in school does have an impact, suggests that the quality of instruction and of teachers more generally may play an important role. This is indeed corroborated by a couple of studies that use different measures of teacher quality.

First, 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.[18]

[17] There is also descriptive evidence that enrollment in early childhood education – i.e., additional time before school – is related to reduced socio-economic gradients and to better integration of migrant children (Schuetz, Ursprung, and Woessmann (2008); Schneeweis (2011)).

[18] Bietenbeck, Piopiunik, and Wiederhold (2015) apply within-student between-subject identification 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 which is complementary to access to subject-specific textbooks.

Second, Dolton and Marcenaro-Gutierrez (2011) provide an analysis of teacher salaries and student achievement across countries. Using aggregate country-level data for 26 OECD countries from several TIMSS and PISA waves, they find that both absolute teacher salary and teachers' relative salary position in a country's income distribution are related to better student achievement. 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 U.S. 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 (specific dimensions of) teacher quality do seem to matter for student achievement. More broadly, these findings suggest ways in which what school systems do is 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 which may frame how resources are used.

Institutional Structures of School Systems: Explorations into Causal Effects

An emerging literature has started to address identification issues in estimating the role of the institutional framework in international student achievement. This is where the international comparative approach promises to be particularly fruitful, 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

An important institutional feature of exam systems that differentiates countries is whether at the end of high school, students' learning outcomes are assessed by external exams. Such external exams can act as an accountability device that reveals the overall outcome of the efforts of students and schools. Key aspects of curriculum-based external exit exam systems are that they produce achievement signals that have real consequences for students, define achievement not just relative to other students in the school but to an external standard, and signal not just a pass-fail signal but multiple levels of achievement in a subject (Bishop (1997)). External exams have been argued to increase external rewards for learning, decrease peer pressure against learning, improve student-teacher relationships, and enhance monitoring of teachers and schools (Bishop and Woessmann (2004); Schwerdt and Woessmann (2015)).

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 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, Woessmann (2003a) finds 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.[19] 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, Woessmann (2011) finds 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

[19] Using a longitudinal component of the German PISA 2003 test which tested 9th-grade students again one year later, Jürges et al. (2012) confirm that external exit exams have a positive effect on the value-added in math achievement in the last year before the exit exams in non-academic tracks, focused in achievement areas that are part of the curriculum.

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

Conceptually, decentralizing decision-making authority to schools can have mixed effects. On the one hand, local decision-makers may have better knowledge of local circumstances and locally optimal teaching methods, so that increased school autonomy could improve student outcomes. On the other hand, autonomous decision-making may be hindered by limited local decision-making capacity and allow schools to act opportunistically in pursuit of conflicting goals, in particular in settings with limited external standards and accountability. As a result, school autonomy may be conducive to student achievement in school systems with strong surrounding structures that ensure high common standards, whereas school-based decision-making may in fact hurt student achievement in low-performing systems that lack basic standards and local capacity. Consistently, cross-sectional evidence from international achievement tests concerning school autonomy has been quite 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, Hanushek, Link, and Woessmann (2013) 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, they 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, they 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, they aggregate their 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, they 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.[20] 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 that allows schools to pursue opposing interests without being monitored, these negative aspects should be eased in school systems where external exams provide comparative information on ultimate performance, thereby constraining opportunistic behavior. Indeed, Hanushek, Link, and Woessmann (2013) 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 et al. (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, in their database 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 is another institutional feature that 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

[20] Using country-level data from different international achievement tests between 1980 and 2000, Falch and Fischer (2012) find that decentralization of general public spending is positively related to student achievement in 24 early OECD countries.

government funding of Dutch privately operated schools is the same (at 95 percent) as in public schools, a feature going back to a constitutional regulation of government funding being independent of the school operator. Private school operation may be related to the extent of school autonomy discussed above, but again these are conceptually different issues; public schools can have substantial autonomy, and private schools can have limited autonomy.

The interest here is the extent to which differences in the share of privately operated schools contribute to differences in student achievement levels across countries. While differences in school-level performance between public and private schools represent one potential mechanism for such a relationship,[21] a key aspect is the general-equilibrium effect introduced by the competition from private schools. As in the analysis of competitive vs. monopolistic providers in any other market, the basic idea is that when parents can choose between different providers that offer real alternatives, providers have a stronger incentive to offer high quality at limited cost. Otherwise, market forces would lead to a loss of customers. Therefore, the existence of private alternatives may lift the performance of public schools, as well, potentially eroding any public-private differences at the school level while at the same time lifting the achievement level system-wide.

Cross-country evidence indeed suggests a strong association of achievement levels with the share of privately operated schools (e.g., 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, West and Woessmann (2010) argue that historical differences in Catholic vs. Protestant denomination provide a natural experiment. In late 19th century, Catholic doctrine strongly resisted the emerging non-denominational 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 that Catholicism was

[21] See Toma (1996) and Vandenberghe and Robin (2004) for within-country analyses of public and private schools in different international achievement tests.

not the state religion) can be used as an instrumental variable for the share of privately operated schools in the 2003 PISA data. To strengthen the identifying assumption that the historical Catholic share is not directly related to current student achievement, the model controls for current differences in Catholic shares.[22]

Results suggest that the share of privately operated schools has a strong positive effect on student achievement across countries. A ten 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, rather than merely differences in effectiveness between privately and publicly operated schools. In addition to increasing achievement, private competition is also estimated to reduce total educational expenditure per student.

Tracking

Another institutional feature of school systems that has been studied in an international context is the age at which students are tracked into different school types serving students of different ability. 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. While predictions on the effect of early tracking on achievement levels diverge and strongly depend on the type of peer effects assumed, it has generally been argued that early tracking may increase inequality by systematically disadvantaging lower-achieving groups. Apart from regionally staggered reforms in some countries, tracking regimes usually do not vary within countries, rendering cross-country identification a promising approach.

To avoid bias from differences in unobserved country characteristics, Hanushek and Woessmann (2006) suggest a differences-in-differences model that exploits variation across grade levels within countries. Building on the idea that no country has differing-ability schools in the early grades of primary school, their identification strategy compares achievement changes from primary to later schooling across tracked and untracked countries. Using country-level data

[22] In addition, there is ample evidence that historically, Catholics have placed less emphasis on education than Protestants (e.g., 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.

for several pairs of PIRLS, TIMSS, and PISA achievement tests administered at the primary and secondary school levels, they find that early tracking significantly increases the inequality in countries' achievement outcomes (measured by standard deviations or percentile differences in achievement scores). They 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 differences-in-differences identification 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.[23] Piopiunik (2014) exploits a school reform in Bavaria that lowered the age of tracking between the two lowest-ability school types to estimate a triple-differences 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-fifth 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. There are differences in

[23] Using the German extension of the 4th-grade PIRLS test, Lüdemann and Schwerdt (2013) show that even conditioning on PIRLS achievement and on a measure of general intelligence, second-generation immigrant students get worse grades and are less likely to be recommended by their teachers for higher-track school types. This additional disadvantage is mostly accounted for by immigrant students' less favorable socio-economic background.

the extent to which this simple descriptive model can account for the achievement levels of specific countries. For example, while the high performance within the OECD of Korea, Japan, and the Netherlands – as well as the low performance of Mexico and Turkey – are predicted reasonably well, this is not true for the high performance of Finland.

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. Approaches such as instrumental-variable models exploiting historical incidents or natural fluctuations, regression discontinuity designs exploiting specific assignment rules, differences-in-differences models exploiting within-student variation across subjects or within-country variation across grade levels, and panel models with country fixed effects using country-level variation over time aim to address concerns with cross-sectional methods for specific aspects of school systems in the observational data.

Some stylized facts 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, indicating the relevance of schools. Third, a number of institutional features of school systems seem to contribute to the cross-country differences in student achievement. 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. For example, high-performing countries such as Finland, Korea, and the Netherlands all combine external exit exams with school autonomy over course content, and the latter two countries also have large sectors of privately operated schools. 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. While the expansion over the past decade of research in this area has contributed to our understanding of the importance of school systems, many important questions remain open. 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 in an emerging panel structure.

Regional variation within some countries may add additional dimensions of analysis. A useful direction for international testing efforts would be to conduct studies in many countries that are longitudinal at the student level. Recently, research has made some headway towards causal identification, but much of this is still in its infancy and concerns about possible bias in cross-country analysis remain. Existing causal identification strategies will be sharpened and new approaches developed. In addition, the evidence on differential effects of school autonomy, as well as evidence on smaller effects of instruction time in developing countries, indicates that results from developed countries do not necessarily generalize to developing countries, or vice versa. Deeper investigation of the extent to which specific results generalize in different settings or not could help to increase our understanding of underlying models of educational production. 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. School systems also differ in many other ways such as specific accountability devises or teacher policies that promise fruitful investigation.

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, Hanushek and Woessmann (2012, 2015a) 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)).[24]

[24] For additional work on student achievement and economic growth, see Hanushek and Kimko (2000), Barro (2001), Ciccone and Papaioannou (2009), and Kaarsen (2014); see Hanushek and Woessmann (2008, 2011a) for reviews.

Hanushek and Woessmann (2012, 2015a) 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, they show that achievement tests before 1985 predict subsequent growth. To address potential bias from omitted factors such as differing economic institutions or cultures, they 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 U.S. labor market, whereas immigrants from the same home countries but schooled in the U.S. do not. Within the United States, Hanushek, Ruhose, and Woessmann (2015) confirm an important role for educational achievement in explaining differences in GDP per capita across U.S. 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 et al. (2015).[25]

But 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 (e.g., Black and Devereux (2011)). Education affects the education and health of children, own health, crime, and citizenship (e.g., Lochner (2011)). More broadly, a "capabilities approach" to welfare analysis in the style of Sen and Nussbaum 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.

[25] For additional evidence from the United States, see, e.g., Murnane, Willett, and Levy (1995), Mulligan (1999), and Chetty et al. (2011).

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): 203-234.

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): 315-348.

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 Maimondides' rule to estimate the effect of class size on scholastic achievement." *Quarterly Journal of Economics* 114 (2): 533-575.

Barro, Robert J. 2001. "Human capital and growth." *American Economic Review* 91 (2): 12-17.

Becker, Sascha O., and Ludger Woessmann. 2009. "Was Weber wrong? A human capital theory of Protestant economic history." *Quarterly Journal of Economics* 124 (2): 531-596.

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): 1437-1472.

Bietenbeck, Jan. 2014. "Teaching practices and cognitive skills." *Labour Economics* 30: 143-153.

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. Munich: CESifo.

Bishop, John H. 1997. "The effect of national standards and curriculum-based examinations on achievement." *American Economic Review* 87 (2): 260-264.

Bishop, John H., and Ludger Woessmann. 2004. "Institutional effects in a simple model of educational production." *Education Economics* 12 (1): 17-38.

Black, Sandra E., and Paul J. Devereux. 2011. "Recent developments in intergenerational mobility." In *Handbook of Labor Economics, Vol. 4, Part B*, 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): 647-674.

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): 234-246.

Cattaneo, Maria A., and Stefan C. Wolter. 2015. "Better migrants, better PISA results: Findings from a natural experiment." *IZA Journal of Migration* 4: 18.

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.

Ciccone, Antonio, and Elias Papaioannou. 2009. "Human capital, the structure of production, and growth." *Review of Economics and Statistics* 91 (1): 66-82.

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): 143-185.

Edwards, Sebastian, and Alvaro Garcia Marin. 2015. "Constitutional rights and education: An international comparative study." *Journal of Comparative Economics* 43 (4): 938-955.

Entorf, Horst, and Nicoleta Minoiu. 2005. "What a difference immigration policy makes: A comparison of PISA scores in Europe and traditional countries of immigration." *German Economic Review* 6 (3): 355-376.

Falch, Torberg, and Justina A.V. Fischer. 2012. "Public sector decentralization and school performance: International evidence." *Economics Letters* 114 (3): 276-279.

Falck, Oliver, Constantin Mang, and Ludger Woessmann. 2015. "Virtually no effect? Different uses of classroom computers and their effect on student achievement." CESifo Working Paper 5266. Munich: CESifo.

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., 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): 433-462.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, math, and gender." *Science* 320 (5880): 1164-1165.

Gundlach, Erich, Ludger Woessmann, and Jens Gmelin. 2001. "The decline of schooling productivity in OECD countries." *Economic Journal* 111 (471): C135-C147.

Hanushek, Eric A. 1986. "The economics of schooling: Production and efficiency in public schools." *Journal of Economic Literature* 24 (3): 1141-1177.

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: 212-232.

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 No. 20727. Cambridge, MA: National Bureau of Economic Research.

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. Cambridge, MA: National Bureau of Economic Research.

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): 607-668.

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): 427-491.

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): 344-345.

Heyneman, Stephen P., and William Loxley. 1983. "The effect of primary school quality on academic achievement across twenty-nine high and low income countries." *American Journal of Sociology* 88 (6): 1162-1194.

Hoxby, Caroline M. 2000. "The effects of class size on student achievement: New evidence from population variation." *Quarterly Journal of Economics* 115 (3): 1239-1285.

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 [accessed January 8, 2016].

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): 766-781.

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): 1134-1155.

Jürges, Hendrik, Kerstin Schneider, Martin Senkbeil, and Claus H. Carstensen. 2012. "Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy." *Economics of Education Review* 31 (1): 56-65.

Kaarsen, Nicolai. 2014. "Cross-country differences in the quality of schooling." *Journal of Development Economics* 107: 215-224.

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): 465-488.

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.

Lüdemann, Elke, and Guido Schwerdt. 2013. "Migration background and educational tracking." *Journal of Population Economics* 26 (2): 455-481.

Mulligan, Casey B. 1999. "Galton versus the human capital approach to inheritance." *Journal of Political Economy* 107 (6, pt. 2): S184-S224.

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): 251-266.

OECD. 2013. *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science (Volume I)*. 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* 52:134-154.

Schneeweis, Nicole. 2011. "Educational institutions and the integration of migrants." *Journal of Population Economics* 24 (4): 1281-1308.

Schnepf, Sylke V. 2007. "Immigrants' educational disadvantage: An examination across ten countries and three surveys." *Journal of Population Economics* 20 (3): 527-545.

Schuetz, Gabriela, Heinrich W. Ursprung, and Ludger Woessmann. 2008. "Education policy and equality of opportunity." *Kyklos* 61 (2): 279-308.

Schwerdt, Guido, and Ludger Woessmann. 2015. "The information value of central school exams." CESifo Working Paper 5404. Munich: CESifo.

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): 365-379.

Singh, Abhijeet. 2015. "Learning more with every year: School year productivity and international learning divergence." University of Oxford, Mimeo [http://www.cesifo-group.de/de/ifoHome/events/Archive/conferences/2015/09/2015-09-11-ee15-Hanushek/Programme.html].

Toma, Eugenia F. 1996. "Public funding and private schooling across countries." *Journal of Law and Economics* 39 (1): 121-148.

Vandenberghe, Vincent, and Stephane Robin. 2004. "Evaluating the effectiveness of private education across countries: A comparison of methods." *Labour Economics* 11 (4): 487-506.

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): 117-170.

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.

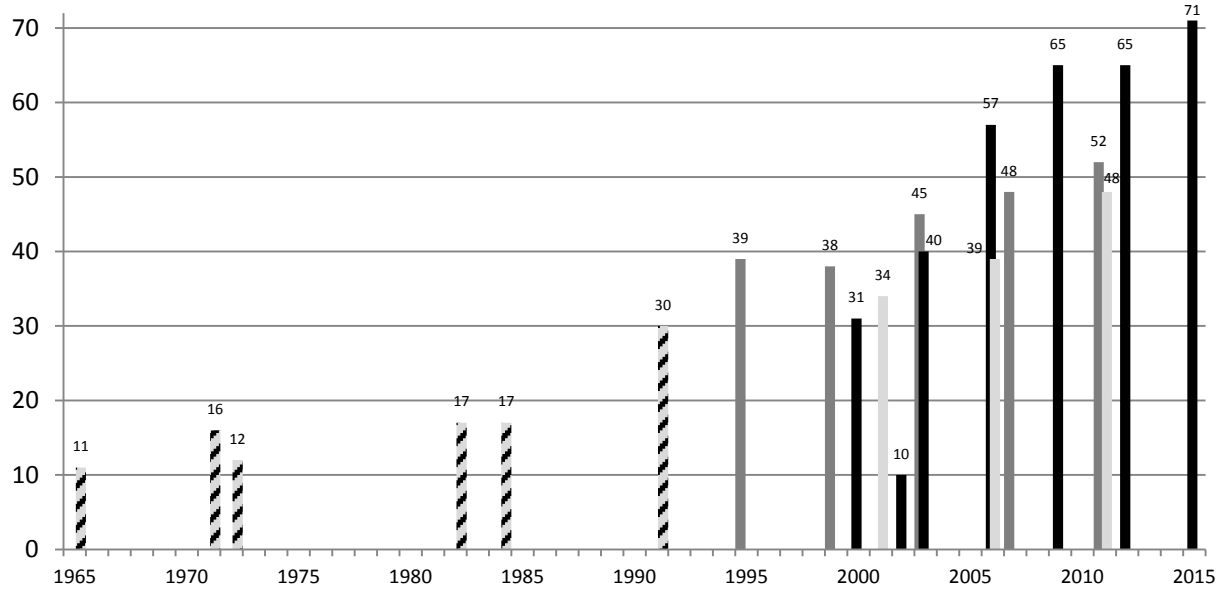
Woessmann, Ludger. 2010. "Institutional determinants of school efficiency and equity: German states as a microcosm for OECD countries." *Journal of Economics and Statistics* 230 (2): 234-270.

Woessmann, Ludger. 2011. "Cross-country evidence on teacher performance pay." *Economics of Education Review* 30 (3): 404-418.

Woessmann, Ludger, Elke Luedemann, Gabriela Schuetz, and Martin R. West. 2009. *School accountability, autonomy, and choice around the world*. Cheltenham, UK: Edward Elgar.

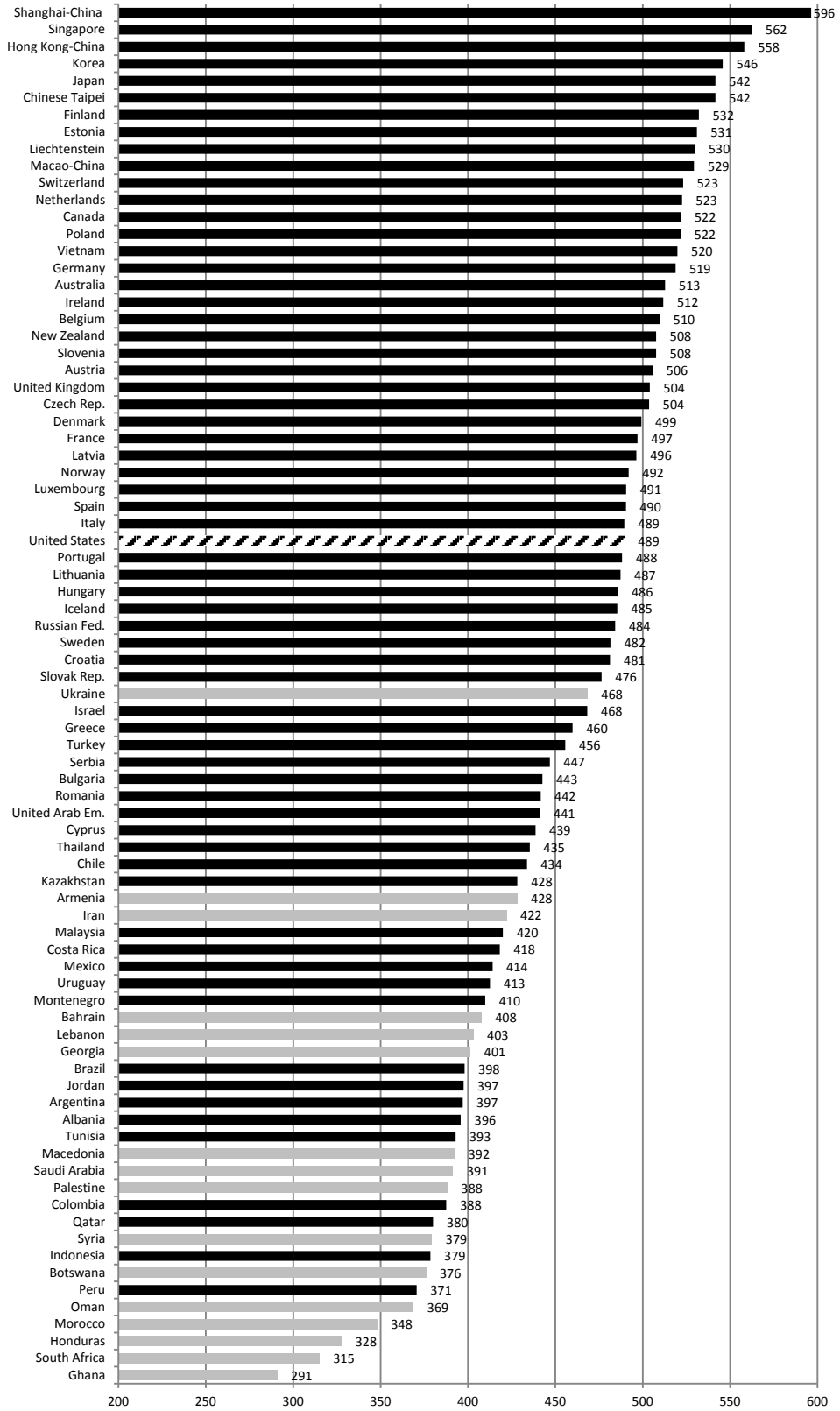
Woessmann, Ludger, and Martin R. West. 2006. "Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS." *European Economic Review* 50 (3): 695-736.

Figure 1: Countries participating in international student achievement tests, 1964-2015



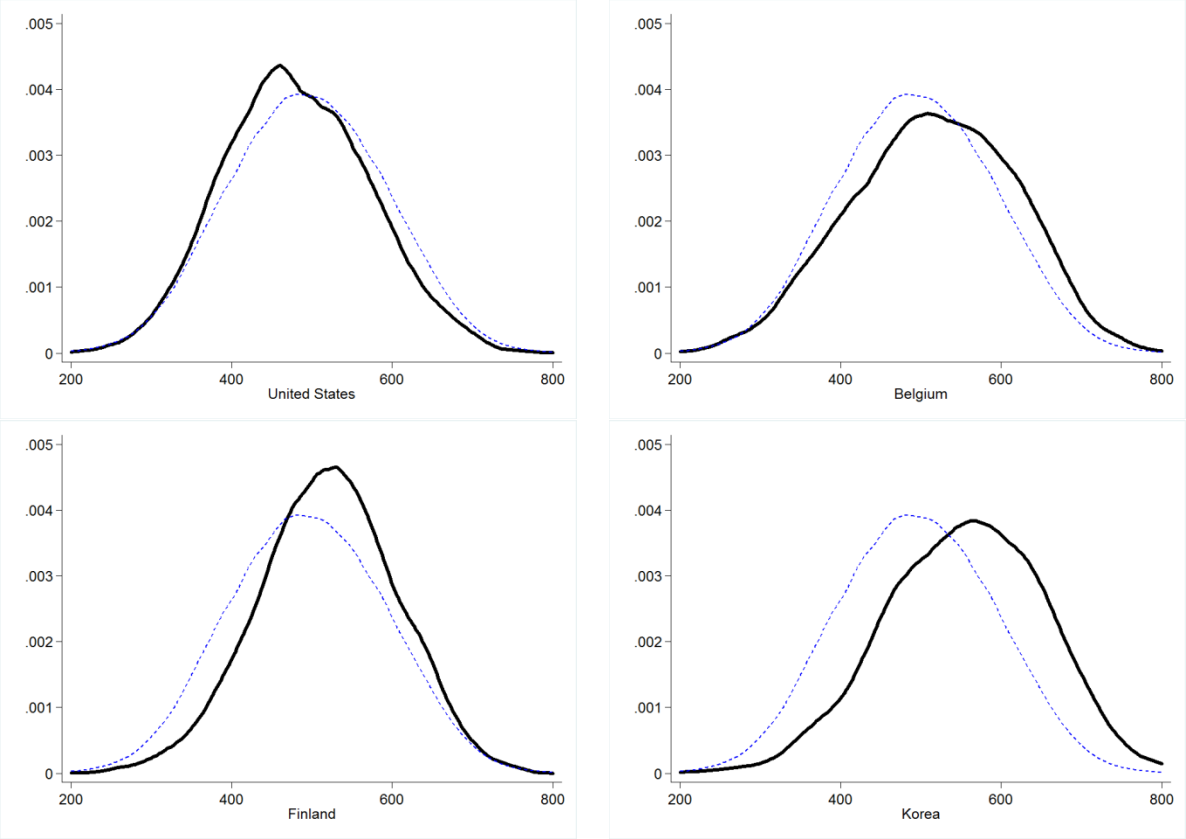
Notes: shaded: pre-1995 IEA studies (1964: 1st math; 1971: 1st science; 1972: 1st reading; 1982: 2nd math; 1984: 2nd science; 1991: 2nd reading); dark grey: IEA's TIMSS study; black: OECD's PISA study; light grey: IEA's PIRLS study.

Figure 2: Performance on recent international student achievement tests, 2011-12



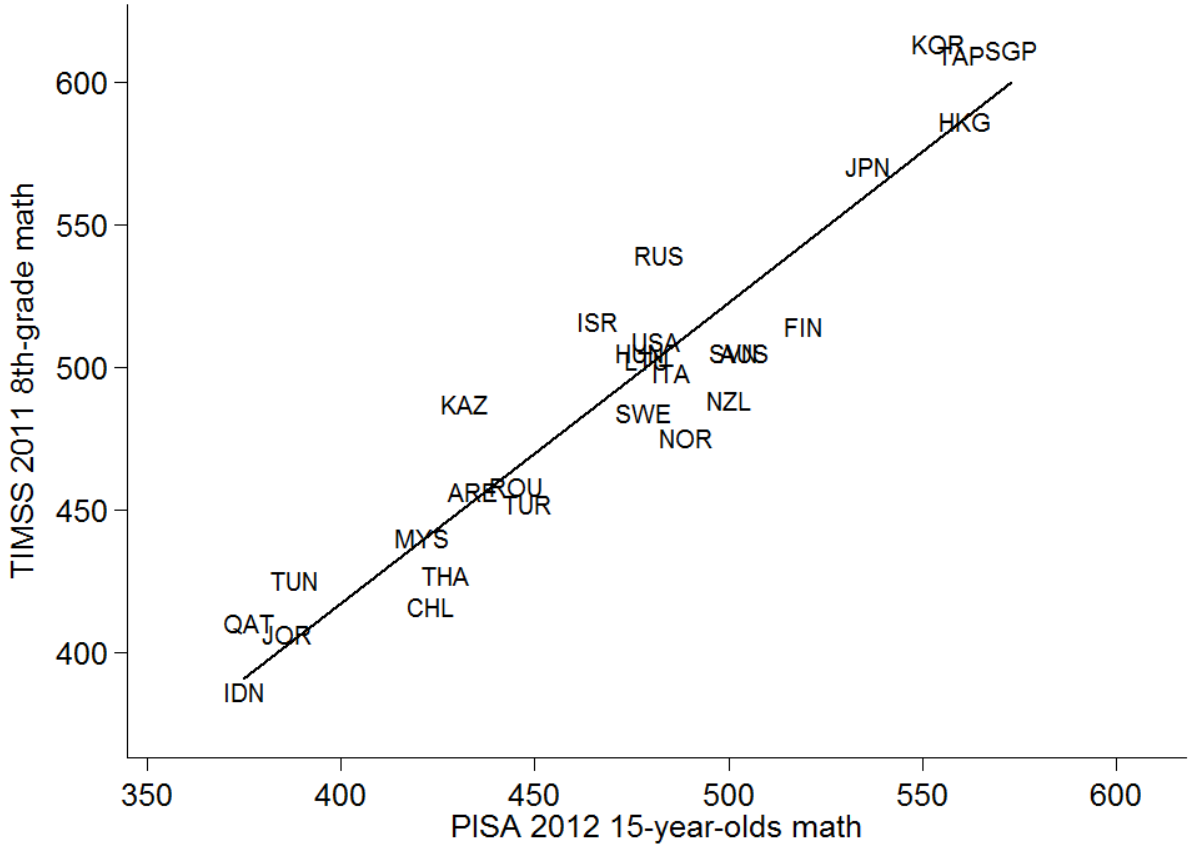
Notes: Average score on international math and science tests. Black: PISA 2012, 15-year-olds; grey: TIMSS 2011, 8th grade, transformed to PISA scale as in Hanushek and Woessmann (2015b).

Figure 3: The distribution of student achievement in selected countries



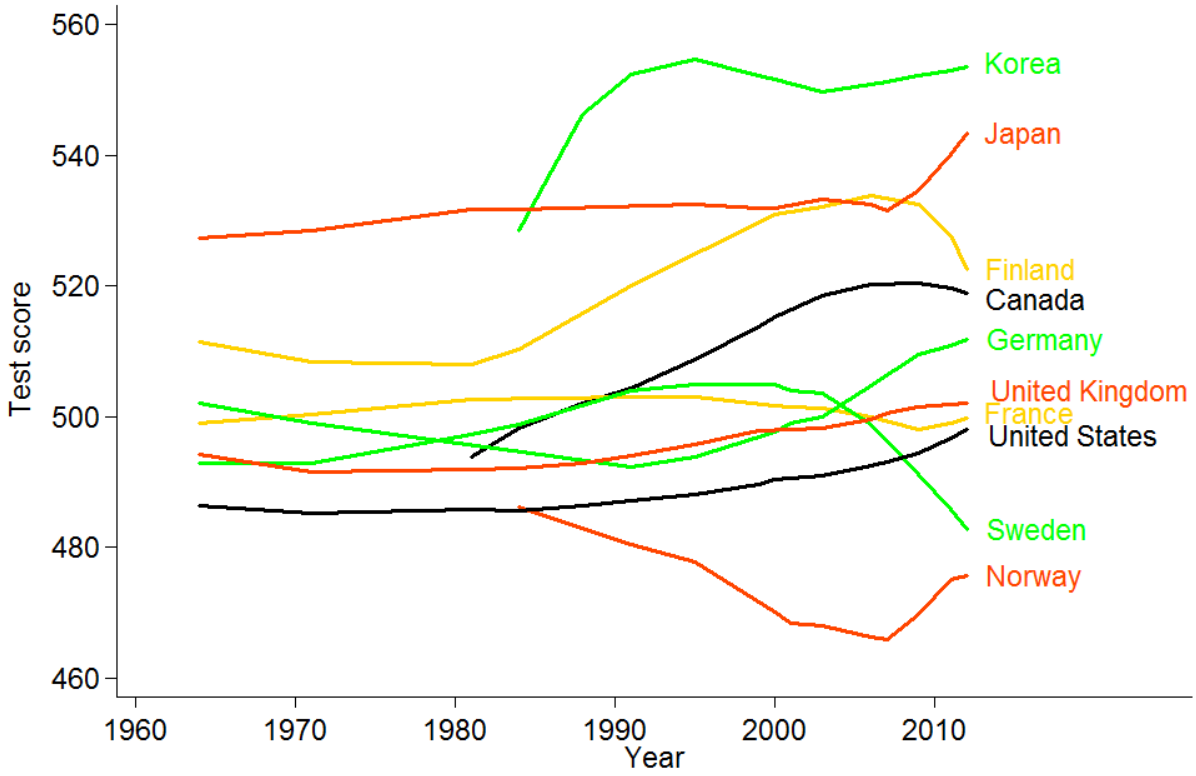
Notes: Kernel densities of student achievement on the PISA 2012 math test. Bold solid line: specified country; thin dotted line: OECD countries.

Figure 4: Student achievement in PISA and TIMSS



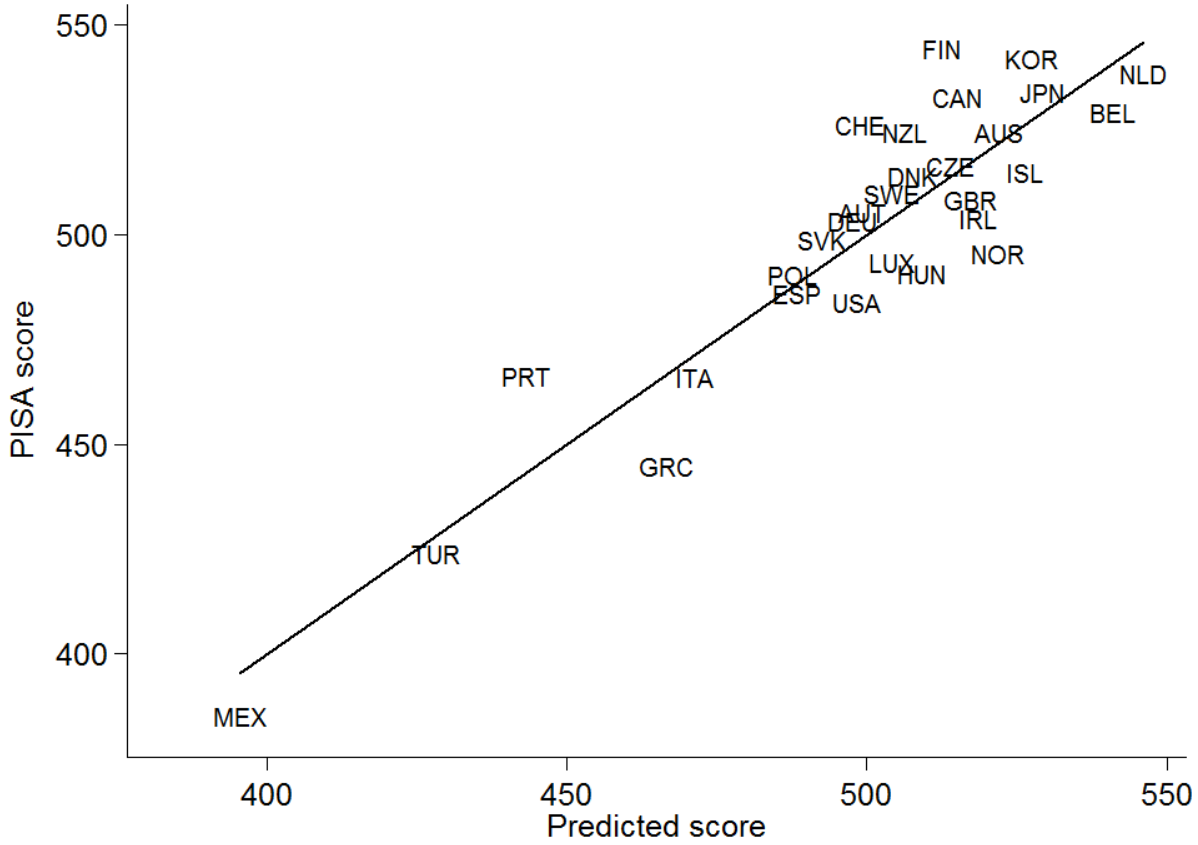
Notes: Math test scores in the PISA 2012 test of 15-year-olds and in the TIMSS 2011 test of 8th-graders.

Figure 5: Long-run test score trends in selected countries, 1964-2012



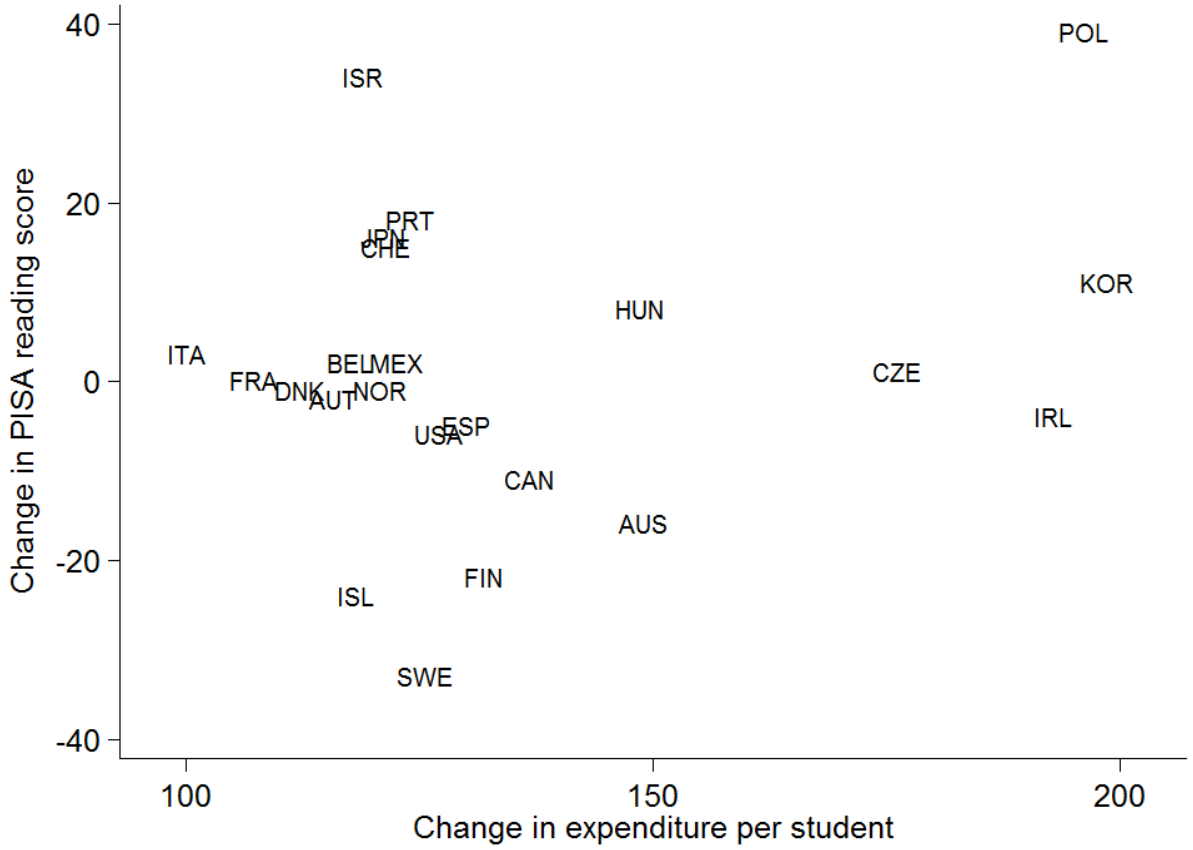
Notes: Stylized depiction of standardized data from international tests 1964-2012. The figure is based on age-group- and subject-specific standardized scores from all international tests in 1964-2003 extended with the subsequently available TIMSS, PIRLS, and PISA data to 2012. It takes out age-group- 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. Colors chosen for expositional reasons only.
Source: Extended from Hanushek and Woessmann (2015a).

Figure 6: Actual and predicted score on international PISA test



Notes: Actual score on PISA 2003 mathematics test and score predicted by input factors in a country-level regression. The three input factors are family background, school resources, and institutions, each of which is measured as a linear combination of individual variables using coefficient estimates from the student-level regression of Table 1, collapsed to the country level. See Table 3 for country letter codes.

Figure 7: Changes in educational spending and in student achievement across countries



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

Table 1: A simple international education production function

	Coef.	Std. err.
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)*
Non-native 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[$]		
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 1 (continued)

	Coef.	Std. err.
School Resources		
Cumulative educational expenditure per student (1,000 $)[c]	0.270**	*(0.103)*
Shortage of instructional materials[s]		
A lot	-8.737**	*(3.514)*
Not at all	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 to 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 x Auton. in establishing starting salaries	14.550*	*(8.104)*
Autonomy in formulating budget	-9.624	*(6.901)*
External exit exams x Autonomy in formulating budget	7.882	*(8.478)*
Autonomy in determining course content	-2.053	*(5.435)*
External exit exams x Autonomy in determining course content	11.504	*(7.262)*
Autonomy in hiring teachers	18.349*	*(10.436)*
External exit exams x 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	

Notes: Data: Programme for International Student Assessment (PISA) 2003. Sample: OECD countries. Dependent variable: students' mathematics test score. Least-squares regressions weighted by students' sampling probability. Measures vary at the student level unless noted otherwise: [s] observed at school level; [c] observed at country level. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level (based on clustering-robust standard errors): *** 1 percent, ** 5 percent, * 10 percent. Source: Own calculations on the basis of Woessmann et al. (2009).

Table 2: 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	

Notes: 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 1, collapsed to the country level.

Table 3: Accounting for each country's difference from the international mean

		Observed difference	Unaccounted difference	Accounted difference	Of which: accounted for by		
					Family background	School resources	Institutions
		(1)	(2)	(3)	(4)	(5)	(6)
FIN	Finland	44.5	31.7	12.9	2.7	-1.3	11.5
KOR	Korea	42.0	14.3	27.7	13.0	5.6	9.1
NLD	Netherlands	38.4	-8.0	46.4	-3.4	-0.3	50.1
JPN	Japan	34.0	4.4	29.6	17.5	2.9	9.2
CAN	Canada	33.0	17.4	15.6	15.9	3.2	-3.5
BEL	Belgium	29.5	-11.8	41.3	-1.2	1.4	41.0
CHE	Switzerland	26.5	27.3	-0.8	-13.2	9.5	2.9
AUS	Australia	24.5	2.1	22.4	14.0	6.6	1.7
NZL	New Zealand	24.5	17.8	6.7	16.2	-3.0	-6.4
CZE	Czech Republic	16.4	2.1	14.3	16.1	-9.0	7.2
ISL	Iceland	15.1	-11.6	26.7	29.7	4.9	-7.9
DNK	Denmark	14.1	6.0	8.1	0.4	6.5	1.2
SWE	Sweden	10.0	5.5	4.5	5.9	-1.0	-0.4
GBR	United Kingdom	8.4	-9.1	17.5	13.0	2.7	1.8
AUT	Austria	5.5	5.7	-0.2	2.1	6.1	-8.5
IRL	Ireland	3.9	-15.0	18.8	-3.3	1.6	20.5
DEU	Germany	3.5	5.4	-1.9	-4.0	-0.8	2.8
SVK	Slovak Republic	-1.0	6.3	-7.3	4.2	-18.0	6.5
NOR	Norway	-4.3	-26.4	22.1	22.1	2.1	-2.1
LUX	Luxembourg	-6.3	-10.7	4.4	-25.5	19.3	10.6
HUN	Hungary	-9.3	-18.7	9.4	4.5	-5.4	10.4
POL	Poland	-9.5	2.5	-12.0	-11.5	-8.1	7.6
ESP	Spain	-14.1	-2.7	-11.4	-4.8	-5.4	-1.2
USA	United States	-16.1	-14.7	-1.4	2.3	9.1	-12.9
PRT	Portugal	-33.5	23.0	-56.5	-27.0	-2.8	-26.7
ITA	Italy	-33.9	-5.5	-28.3	2.7	3.6	-34.7
GRC	Greece	-55.1	-22.1	-33.0	-4.1	-3.0	-26.0
TUR	Turkey	-75.8	-4.4	-71.5	-31.7	-17.5	-22.3
MEX	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 the country-level regression depicted in Figure 6. Column 3: difference accounted for by the country-level regression depicted in Figure 6. 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.