NBER WORKING PAPER SERIES

SAMPLE SELECTIVITY AND THE VALIDITY OF INTERNATIONAL STUDENT ACHIEVEMENT TESTS IN ECONOMIC RESEARCH

Eric A. Hanushek
Ludger Woessmann

Working Paper 15867
http://www.nber.org/papers/w15867

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2010

Woessmann gratefully acknowledges the hospitality and support provided by the W. Glenn Campbell and Rita Ricardo-Campbell National Fellowship of the Hoover Institution, Stanford University. Support has also come from the Pact for Research and Innovation of the Leibniz Association. Hanushek has been supported by the Packard Humanities Institute. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Sample Selectivity and the Validity of International Student Achievement Tests in Economic Research
Eric A. Hanushek and Ludger Woessmann
NBER Working Paper No. 15867
April 2010
JEL No. C83,H4,I20,O40

ABSTRACT

Critics of international student comparisons argue that results may be influenced by differences in the extent to which countries adequately sample their entire student populations. In this research note, we show that larger exclusion and non-response rates are related to better country average scores on international tests, as are larger enrollment rates for the relevant age group. However, accounting for sample selectivity does not alter existing research findings that tested academic achievement can account for a majority of international differences in economic growth and that institutional features of school systems have important effects on international differences in student achievement.

Eric A. Hanushek
Hoover Institution
Stanford University
Stanford, CA 94305-6010
and NBER
hanushek@stanford.edu

Ludger Woessmann
University of Munich
Ifo Institute for Economic Research and CESifo
Poschingerstr. 5
81679 Munich, Germany
woessmann@ifo.de

1. Introduction

Economic research has made increasing use of international student achievement data,[1] but critics suggest that underlying sampling issues might compromise any comparability across countries. Non-random differences in patterns of school enrollment, sample exclusions, and non-response are clearly able to influence rankings of countries on international league tables of average student achievement. The extent, however, to which such sample selection also affects results of analyses that use the international test score data is currently unknown. This research note draws on detailed information on sampling quality to estimate whether international differences in sample selection affect the outcomes of typical economic analyses.

We find that countries having more schools and students excluded from the targeted sample, having schools and students who are less likely to participate in the test, and having higher overall school enrollment at the relevant age level tend to perform better on the international tests. However, none of these sampling patterns affect the results of typical growth regressions and education production functions, implying that they are unrelated to the associations of interest in economic analyses.

In the political debate, poor performance on international achievement tests famously motivated the National Commission on Excellence in Education (1983) to declare the United States "*A Nation at Risk*". Others have suggested, however, that this is a myth created by biased samples in other countries (Berliner and Biddle (1995)). To such critics, "The basic problem is student selectivity: The fewer the students who take the test, the higher the average score. That score … simply reflects the fact that the students represented in the test comparisons have been much more highly selected in some countries than in others" (Rotberg (1995), p. 1446; see also Bracey (1996); Prais (2003)).[2] But others disagree with the view that sample selection is a major source of bias in international achievement comparisons (e.g., Baker (1997); Adams (2003)).

[1] See Hanushek and Woessmann (forthcoming) for a review of the extensive economic literature on international educational achievement. Studies using international test score data as determinants of economic growth include Hanushek and Kimko (2000), Barro (2001), Bosworth and Collins (2003), Ciccone and Papaioannou (2009), and Hanushek and Woessmann (2008, 2009). Studies using international test score data as outcomes of education production functions include Bishop (1997), Lee and Barro (2001), Woessmann (2003), Bedard and Dhuey (2006), Brunello and Checchi (2007), Guiso, Monte, Sapienza, and Zingales (2008), Ammermueller and Pischke (2009), and West and Woessmann (forthcoming).

[2] The tests included in our analyses have been devised in an international cooperative process between all participating countries with the intent of making the assessments independent of the culture or curriculum in any particular country. Yet, another criticism that is sometimes raised against international comparisons of student

Simple calculations indicate that in fact sampling bias certainly has the potential to move country mean test scores substantially. For example, if the exclusion propensity and student achievement are bivariate normally distributed and correlated at 0.5, exclusion rates of 10 percent – not uncommon in some countries – lead to an upward bias in the resulting country mean score of 10 percent of a standard deviation (see Organisation for Economic Co-operation and Development (2007)).[3] Of course, the extent to which exclusion and performance are correlated is unknown. If exclusion is random, it does not bias results at all. But the calculation suggests that differential sampling quality may well affect overall country rankings despite the stringent technical standards and extensive efforts of quality assurance by the international testing organizations (e.g., Organisation for Economic Co-operation and Development (2009)).

The basic notion of measurement error in econometric analyses tells us that it is another matter whether and how such mismeasurement of country mean performance biases results of econometric analyses of relationships. First, any bias depends on whether sample selectivity is idiosyncratic or persistent over time – i.e., whether some countries have systematically more selective samples than others or not. If it is idiosyncratic, sample selectivity introduces classical measurement error that works against finding statistically significant associations: It attenuates the estimated coefficient on test scores for errors in an explanatory variable and reduces statistical power, increasing standard errors, for errors in the dependent variable. But, in applications that use averages of performance across several tests (as in most economic growth applications), the importance of any idiosyncratic measurement error will be lessened since the error variance is reduced by averaging. When sample selectivity is persistent across time, the second issue is whether it is correlated with the error term of the estimation equation. If it is orthogonal to the (conditional) variable whose association with test scores is of interest, even systematic sample selectivity simply works against finding statistically significant results. Only

achievement is that test items may be culturally biased or inappropriate for specific participating countries (e.g., Hopmann, Brinek, and Retzl (2007)). Adams, Berezner, and Jakubowski (2010) show that overall country rankings are remarkably consistent when countries are compared using just those PISA-2006 items that representatives of each specific country had initially expressed to be of highest priority for inclusion, and presumably most appropriate for their own school system. From the opposite perspective, one set international comparison (not employed here) was built on tests directly taken from the assessments used in the United States, but the results from these comparisons did not alter the low ranking of U.S. students (see Lapointe, Mead, and Phillips (1989)).

[3] This statement refers to standard deviations at the student level. While varying across specific tests, this is roughly equivalent to twice the standard deviation in country mean scores.

if it is correlated with the error term of the equation of interest does systematic sample selectivity introduce bias to econometric analyses.[4]

The next section investigates the correlation of sample selectivity with test scores. The subsequent sections provide evidence whether accounting for sample selectivity affects results of typical growth regressions and international education production functions, respectively.

2. Sample Selection and Average Test Scores

2.1 The Three Sources of Sample Selection in International Tests

It is useful to distinguish three main sources of discrepancies between the sample of students tested in a country and its total population of children at the age of interest. First, testing is always focused on students in school. Part of the children in the tested age range may no longer be in school, which eliminates them from the official target population of the international tests. This first problem is not associated with the testing so much as with the character of schooling in each country. Second, to a limited extent, national testing authorities are allowed to exclude certain schools and students from their national target population, mostly excluding small remote schools, schools serving students with disabilities, and individual students with disabilities or limited proficiency in the test language. Third, once the national sampling frame is set, non-responses may reduce the testing of students. Some of the sampled schools may not participate in the test, and some of the sampled students may be absent on the testing day. We will separately deal with each of these sources of sample selectivity, because each may have very different impacts on the validity of the testing and the importance of statistical bias.

Our empirical analysis focuses on the five international tests in mathematics and science conducted at the lower secondary level between 1995 and 2003. For consistency with the most recent economic growth research, we do not consider tests beyond 2003. We further restrict attention to tests in math and science, which are most readily comparable across countries. While documentation on the quality of sampling is mostly missing on the early international student achievement tests, since the mid-1990s the organizations responsible for the major international testing cycles – the International Association for the Evaluation of Educational

[4] Studies such as Hanushek and Woessmann (2009) that include country fixed effects deal with possible bias from systematic sampling errors by removing time-invariant factors for each country.

Achievement (IEA) and the Organisation for Economic Co-operation and Development (OECD) – provide detailed documentation of the extent to which each participating country covered the underlying student population in its sampling. In 1995, 1999, and 2003, the IEA conducted the Trends in International Mathematics and Science Study (TIMSS), whose common target population is students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-old students. In 2000 and 2003, the OECD conducted the Programme for International Student Assessment (PISA), whose target population is 15-year-old students.

Both tests allow exclusions for small geographically remote schools, for schools focused on students with intellectual or functional disabilities, and for individual students in the latter group within schools. Excluding students from the target sample is generally permissible for students who are unable to follow the general instruction of the test, but not simply because of poor academic performance or normal disciplinary problems. To limit such exclusions, the tests generally require participating countries to keep exclusion rates below 5 percent (see Mullis, Martin, Gonzalez, and Chrostowski (2004) and Organisation for Economic Co-operation and Development (2004) for details).

Sampled schools in many nations are not required to participate. Moreover, individual students may be absent on the day of the assessment. Again, to limit the extent of such non-participation, response rates are generally deemed acceptable only if they reach 85 percent both at the school level and at the student level (80 percent at the student level in PISA). Substantial breaches of these sampling requirements led the Netherlands and the United Kingdom to be excluded from PISA reporting in 2000 and 2003, respectively, and several countries to be annotated as not meeting sampling guidelines in the TIMSS results tables.

Given the nature of the permissible exclusions – small, remote schools and students with special needs or language deficiencies – higher exclusion rates are likely to introduce positive selection bias into estimates of national mean performance. The direction of selection bias is not as obvious for non-response rates, but if weaker performing schools and students are less likely to participate in the test, it would go in the same direction as for exclusion rates.

Even less clear is the direction of bias for enrollment rates in tested ages. Given our focus on tests in lower secondary school, virtually all developed countries have close to universal enrollment. As a consequence, sampling differences mostly come into play when comparing developed to less-developed countries. It is generally the case that students with higher ability or other background features supportive of higher achievement are more likely to be enrolled in school, introducing bias similar to exclusion rates. But at the country level, this bias is likely to be overwhelmed by the fact that low enrollment rates in lower secondary education are a sign for a generally underdeveloped or dysfunctional education system. On net, both biases are likely to be at work, giving rise to the possibility of a positive association between enrollment rates and test performance.

The first two columns of Table 1 report descriptive statistics of the data on sample coverage for the 196 country observations on the five international tests.[5] Average school enrollment at testing age is 91.8 percent. With the exceptions of Mexico and Turkey, though, all OECD countries come close to universal enrollment at the age range of the underlying tests. Other countries with relatively low enrollment rates include Albania, Brazil, Ghana, Macedonia, Morocco, and Peru. The average exclusion rate (from elimination of schools by the central testing authorities) is 3.1 percent. The exclusion rate is higher than 10 percent on three occasions – Israel in TIMSS 1999 and 2003 (16 and 22.5 percent) and Macedonia in TIMSS 2003 (12.5). On an additional nine occasions exclusion rates fall between 7 and 10 percent, covering nine different countries and all five tests (except TIMSS 1999). The average non-response rate, which arises at the school level, is 11.6 percent. The non-response is higher than 30 percent on eight occasions: Israel in TIMSS 1995 (54.9 percent), the United States in PISA 2000 and 2003 (40.2 and 43.6), the United Kingdom in PISA 2000 and TIMSS 2003 (33.4 and 39), and the Netherlands (40.2), South Africa (37.9), and Bulgaria (36.4) in TIMSS 1995.

2.2 The Correlation of Sample Coverage with National Mean Test Scores

Table 1 (column 3) also reports the correlations of the components of sample selection with reported mean test performance of countries across the five international tests. The correlations reveal that exclusion rates and non-response rates are as expected significantly positively associated with reported test scores: The larger the share of schools and students excluded by the

[5] The sources for the data on population coverage and participation rates in the different TIMSS and PISA tests are Beaton et al. (1996), Mullis et al. (2000), Mullis, Martin, Gonzalez, and Chrostowski (2004), and Organisation for Economic Co-operation and Development (2003, 2004). Because the TIMSS tests did not report school enrollment rates, we draw on data on gross enrollment rates in lower secondary education available from the World Bank (2010) to measure enrollment rates relevant for the TIMSS tests in countries where we do not have enrollment information from PISA. We predict comparable enrollment rates for countries not participating in PISA based on a regression of enrollment rates reported by PISA on the gross enrollment rates (capped at full enrollment) for the 37 countries with both measures available.

national testing authority and the larger the share of schools and students sampled but not participating, the higher the reported country mean test score.[6] At the same time, enrollment rates are also positively correlated with test scores, suggesting that there is no simple upward bias in the test scores of countries where a substantial share of the age group is not enrolled in school.[7]

These overall results are quite robust. The significant correlation of the three measures of sample coverage with test scores is robust to controlling for fixed effects for the five underlying tests. The reported correlations are similar when test scores in math and science are used separately. Looking at correlations within each of the five international tests, enrollment rates are always positively significantly correlated with test scores. Correlations with exclusion rates are significant in PISA 2003, marginally significant in PISA 2000 and TIMSS 2003, and not otherwise. Correlations with non-response rates are significant in the PISA tests but not in the TIMSS tests.[8] As the last two columns of Table 1 show, exclusion rates and non-response rates are significantly correlated with enrollment rates but not with each other. When all three are entered in a regression to predict test scores, only enrollment rates remain significant.

To test whether some countries systematically sample smaller shares of the population than others, Table 2 reports correlations of exclusion rates and non-response rates across tests. (Of course, enrollment rates are relatively constant over the short time period and are not reported in the table).[9] Non-response rates are positively correlated across the five tests. By contrast, exclusion rates are significantly correlated in only three of the ten pairs of tests. Thus, sample selectivity is only to a limited degree systematic over time and has a substantial idiosyncratic component, particularly in terms of exclusion decisions made by national testing authorities.

[6] When subdividing exclusion and non-response rates into a school-level and a within-school student-level component each, both components of the non-response rate are positively correlated with test scores, whereas only the student-level component of the exclusion rate is significantly correlated with test scores.

[7] Combining exclusion and non-response rates into one non-participation rate per country also yields a positive correlation with test scores. Combining all three measures of sample coverage into one measure of total non-participation yields a negative correlation with test scores, i.e., the total is dominated by the negative correlation of non-enrollment with test scores.

[8] In PISA 2003, subcategories of student exclusions are reported for students with functional disability, intellectual disability, limited assessment language proficiency, and other. Exclusions due to functional disability are most closely correlated with test scores, exclusions due to intellectual disability and limited language proficiency only in some subjects, and the residual other category not.

[9] Note, however, that Hanushek and Woessmann (2009) find that changes in enrollment rates of over longer periods of time are uncorrelated with trends in test scores.

3. Sample Selection and the Results of Growth Regressions

Economists have focused on two uses of international test scores: modeling cross-country growth differences and modeling how educational institutions affect student achievement (see Hanushek and Woessmann (forthcoming)). It is possible to illustrate the impact of sample selection on results in both areas by introducing measures of test participation rates into representative published models of each type. In this section, we analyze the effect of potentially biased testing on the analysis of long-term economic growth.

We employ the basic growth regression framework of Hanushek and Woessmann (2008), where the average annual growth rate in real GDP per capita over 1960-2000 is expressed as a function of initial GDP per capita, initial years of schooling, and a test score measure that combines performance on all international student achievement tests from primary through upper secondary school between 1964 and 2003. The first column of Table 3 replicates the basic model of Hanushek and Woessmann (2008). The second column reports the same model for the sample of 45 countries for which we have information on sampling quality. Test scores have a significant positive effect on economic growth, with a one standard deviation increase in test scores associated with 1.74-1.98 percentage points of additional average annual growth.[10]

Column (3) adds our three measures of sample coverage – enrollment, exclusion, and non-response rates – to the growth model. They enter statistically insignificantly, individually or jointly, and do not significantly affect the coefficient on test scores. That is, the variation in the extent to which sampling is selective across countries is orthogonal to the variation in conditional economic growth. Thus, the positive association between test scores and economic growth cannot be explained by international differences in sample selectivity.[11]

[10] Concerns about identification of causal impacts frequently arise in such growth models. While not conclusive, instrumental-variable, first-differenced, and differences-in-differences models are developed in Hanushek and Woessmann (2009) to rule out commonly hypothesized threats to the identification of causal effects of test scores on economic growth.

[11] The same results hold if exclusion rates and non-response rates are summed up to a joint non-participation measure, and if all three measures of sample coverage are combined into a measure of total non-participation. When entering measures of school-level and within-school student-level non-response separately, neither enters significantly or affects qualitative results. When entering measures of school-level and within-school exclusions separately, school-level exclusions tend to enter marginally significantly negatively, without affecting the coefficient on test scores. None of the exclusion subcategories available in PISA 2003 – functional disability, intellectual disability, limited language proficiency, and other – captures statistical significance or affects the test score result. Controlling for limited coverage of national populations due to exclusion of certain regions or non-test-language schools from the national desired population, as is the case in a few countries in the TIMSS tests, also does not affect the qualitative results.

To this point, the test score measure refers to all international achievement tests, whereas our sampling information refers only to the five international tests conducted since 1995. In column (4), we therefore use a test score measure created from just the five tests at the lower secondary level in 1995-2003 for which we have sampling information. While the point estimate on this test score measure is slightly (but not significantly) smaller – presumably because of attenuation when using a measure based on fewer test information – qualitative results on the effect of including sampling information are the same.[12]

To ensure that the latter specification does not just capture test score variation that emerged towards the end (1995-2003) of the growth period of our analysis (1960-2000), column (5) uses the average test score of all international tests (1964-2003) as an instrument for the recent tests. Qualitative results are unchanged in this two-stage least-squares regression. In column (6), we restrict the analysis to only that part of the variation in recent test scores that is related to test score variation on the early tests (1964-1985), ensuring that only test score variation that can be traced back to the early tests is used in the analysis. While this reduces the sample to the 20 countries that participated in the early tests, the qualitative result on the effect of test scores on economic growth is unaffected. The same is true if we use only growth rates from 1980-2000 in this final specification (coefficient on test score equals 1.707). This final specification uses only test score variation in the identification that mostly pre-dates the growth rates while at the same time using only variation related to tests for which we have the relevant sampling information as control variables.

In additional analyses, we tested whether results are affected by how often countries participated in international tests, which might be another source of differential reliability of international test information across countries. Qualitative results are unaffected by controlling for how often countries participated and for indicators of participation in early or recent tests, by looking at sub-samples of countries participating fewer than or at least five times and participating in the early tests or not, and by weighting regressions by the number of test participations per country.[13]

[12] Hanushek and Woessmann (2009) present extensive sensitivity tests on the use of varying specifications, different assessments of performance, and different time periods for tests and growth.

[13] Detailed results are available from the authors on request.

4. Sample Selection and the Results of Education Production Functions

A second use of international test data is focused on how institutional features of national school systems affect student outcomes, a central question in the analysis of educational production functions using international data. This estimation has systematically found that institutional features of school systems capturing choice, accountability, and autonomy account for a substantial part of the cross-country variation in student achievement, whereas measures of school resources generally do not (see Hanushek and Woessmann (forthcoming) for a review). In this context, sample selectivity may be a particular issue. For example, evidence from Florida suggests that schools may respond to high-stakes test-based school accountability by excluding low-performing students from counting on the accountability test through reclassifying them as disabled (Figlio and Getzler (2006)). On the other hand, the tests that provide the internationally comparable achievement data are not the tests underlying the accountability systems, mitigating worries that incentives for sample selection affect the international testing.

The first column of Table 4 replicates a basic set of estimates of international education production functions based on Woessmann, Luedemann, Schuetz, and West (2009)). These estimates employ PISA-2003 math data at the student level and pool all OECD countries with available data. Apart from the institutional measures reported here – two measures of choice, six measures of accountability, and four measures of autonomy and their interaction with external exit exams – the model includes 15 student control variables such as age, gender, and immigrant status; 17 family-background controls such as family status, parental occupation, and the number of books at home; and 10 school-input controls such as educational expenditure, class size, shortage of materials, and teacher education (not shown here). The main pattern of results on the institutional effects is a positive association of student achievement with the share of privately operated schools, government funding, external exit exams, and school-level accountability measures. Several school-level measures of school autonomy are negatively related to achievement in systems without accountability, but positively in those with accountability.

Column (2) adds our three measures of sample coverage. Enrollment, exclusion, and non-response rates are jointly insignificant, and the pattern of results remains unaffected.[14] Again,

[14] The non-response rate is actually marginally significant, but negative, i.e., countries with higher non-response rates perform worse, rather than better, after controlling for the components of the production function. Analyses using the school-level and within-school student-level components of exclusion and non-response rates

the results suggest that sample selectivity is orthogonal to the associations of interest in international education production functions and thus does not affect their results.

In line with the results reported above, column (3) shows that enrollment rates are in fact positively related to student achievement in PISA 2003 as long as the components of the production function are not controlled for. (Exclusion rates and non-response rates are not significantly related to test scores in this OECD country sample, also when entered individually.) However, as column (4) shows, this association is driven solely by the two countries with enrollment rates below 90 percent (Mexico and Turkey).

5. Conclusions

Enrollment, exclusion, and non-response rates are positively correlated with reported country mean scores on international student achievement tests. But the sample selectivity indicated by these measures does not affect the results of typical research on economic growth and educational production. The international variation in selectivity of student samples is orthogonal to the associations of interest in these economic literatures.

separately reveal that the negative coefficient on the non-response rate is solely due to its school-level component. Qualitative results remain unaffected when including the components separately. Combined versions of the three sample coverage measures do not enter the model significantly and do not affect the main results about the importance of institutional features.

References

Adams, Raymond J. 2003. "Response to 'Cautions on OECD's recent educational survey (PISA)'." *Oxford Review of Education* 29, no. 3: 377-389.

Adams, Raymond J., Alla Berezner, and Maciej Jakubowski. 2010. "Analysis of PISA 2006 preferred items ranking using the percentage correct method." OECD Education Working Paper 46, Paris: Organisation for Economic Co-operation and Development.

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, no. 3: 315-348.

Baker, David P. 1997. "Surviving TIMSS: Or, everything you blissfully forgot about international comparisons." *Phi Delta Kappan* 79, no. 4 (December): 295-300.

Barro, Robert J. 2001. "Human capital and growth." *American Economic Review* 91, no. 2: 12-17.

Beaton, Albert E., Ina V. S. Mullis, Michael O. Martin, Eugenio J. Gonzalez, Dana L. Kelly, and Teresa A. Smith. 1996. *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Bedard, Kelly, and Elizabeth Dhuey. 2006. "The persistence of early childhood maturity: International evidence of long-run age effects." *Quarterly Journal of Economics* 121, no. 4: 1437-1472.

Berliner, David C., and Bruce J. Biddle. 1995. *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Boston: Addison-Wesley Publishing Company.

Bishop, John H. 1997. "The effect of national standards and curriculum-based examinations on achievement." *American Economic Review* 87, no. 2: 260-264.

Bosworth, Barry P., and Susan M. Collins. 2003. "The empirics of growth: An update." *Brookings Papers on Economic Activity* 2003, no. 2: 113-206.

Bracey, Gerald W. 1996. "International comparisons and the condition of American education." *Educational Researcher* 25, no. 1 (January-February): 5-11.

Brunello, Giorgio, and Daniele Checchi. 2007. "Does school tracking affect equality of opportunity? New international evidence." *Economic Policy* 22, no. 52: 781-861.

Ciccone, Antonio, and Elias Papaioannou. 2009. "Human capital, the structure of production, and growth." *Review of Economics and Statistics* 91, no. 1: 66-82.

Figlio, David N., and Lawrence S. Getzler. 2006. "Accountability, ability and disability: Gaming the system?" In *Advances in Applied Microeconomics, volume 14: Improving school accountability: Check-ups or choice*, edited by Timothy J. Gronberg and Dennis W. Jansen. Amsterdam: Elsevier: 35-49.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, math, and gender." *Science* 320, no. 5880: 1164-1165.

Hanushek, Eric A., and Dennis D. Kimko. 2000. "Schooling, labor force quality, and the growth of nations." *American Economic Review* 90, no. 5 (December): 1184-1208.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The role of cognitive skills in economic development." *Journal of Economic Literature* 46, no. 3 (September): 607-668.

———. 2009. "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation." NBER Working Paper W14633, Cambridge, MA: National Bureau of Economic Research.

———. forthcoming. "The economics of international differences in educational achievement." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin and Ludger Woessmann. Amsterdam: North Holland.

Hopmann, Stefan Thomas, Gertrude Brinek, and Martin Retzl, eds. 2007. *PISA zufolge PISA: Hält PISA, was es verspricht? / PISA according to PISA: Does PISA keep what it promises?* Vienna: LIT Verlag.

Lapointe, Archie E., Nancy A. Mead, and Gary W. Phillips. 1989. *A world of differences: An international assessment of mathematics and science*. Princeton, NJ: Educational Testing Service.

Lee, Jong-Wha, and Robert J. Barro. 2001. "Schooling quality in a cross-section of countries." *Economica* 68, no. 272: 465-488.

Mullis, Ina V. S., Michael O. Martin, Eugenio J. Gonzalez, and Steven J. Chrostowski. 2004. *TIMSS 2003 international mathematics report: Finding for IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.

Mullis, Ina V. S., Michael O. Martin, Eugenio J. Gonzalez, Kelvin D. Gregory, Robert A. Garden, Kathleen M. O’Connor, Steven J. Chrostowski, and Teresa A. Smith. 2000. *TIMSS 1999 international mathematics report: Findings from IEA’s repeat of the Third International Mathematics and Science Study at the eighth grade*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

National Commission on Excellence in Education. 1983. *A nation at risk: The imperative for educational reform*. Washington, D.C.: U.S. Government Printing Office.

Organisation for Economic Co-operation and Development. 2003. *Literacy skills for the world of tomorrow: Further results from PISA 2000*. Paris: OECD.

———. 2004. *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

———. 2007. *PISA 2006: Science competencies for tomorrow's world*. Vol. 1 – Analysis. Paris: OECD.

———. 2009. *PISA 2006 technical report*. Paris: OECD.

Prais, Sig J. 2003. "Cautions on OECD's recent educational survey (PISA)." *Oxford Review of Education* 29, no. 2 (June): 139-163.

Rotberg, Iris C. 1995. "Myths about test score comparisons." *Science* 270, no. 5241 (December 1): 1446-1448.

West, Martin R., and Ludger Woessmann. forthcoming. "'Every Catholic child in a Catholic school': Historical resistance to state schooling, contemporary school competition, and student achievement." *Economic Journal*.

Woessmann, Ludger. 2003. "Schooling resources, educational institutions, and student performance: The international evidence." *Oxford Bulletin of Economics and Statistics* 65, no. 2: 117-170.

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

World Bank. 2010. *EdStats: Education statistics version 5.3* 2010 [cited March 18 2010]. Available from http://www.worldbank.org/education/edstats.

Table 1: Sample coverage – descriptive statistics and correlation with test scores

	Mean	Min	Correlation with		
Source of sample	*(Std. dev.)*	Max	Test score	Enrollment rate	Exclusion rate
selection problems	(1)	(2)	(3)	(4)	(5)
Enrollment rate	91.8	42.7	0.571***	1.000	
	(11.3)	103.0	*(0.000)*		
Exclusion rate	3.1	0.0	0.133*	0.127*	1.000
	(2.8)	22.5	*(0.063)*	*(0.076)*	
Non-response rate	11.6	0.0	0.198***	0.207***	0.097
	(9.4)	54.9	*(0.005)*	*(0.004)*	*(0.177)*

Notes: 196 country-level observations: all participants in the five international tests (TIMSS 1995, 1999, 2003; PISA 2000, 2003). Test score is average of math and science on the Hanushek and Woessmann (2009) comparable scale. Correlations: *p*-values in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 2: Sample coverage – correlation across tests

	Exclusion rate				Non-response rate			
	TIMSS			PISA	TIMSS			PISA
	1995	1999	2003	2000	1995	1999	2003	2000
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TIMSS 1999	0.132				0.514***			
	(0.519)				*(0.007)*			
TIMSS 2003	-0.036	0.670***			0.336	0.790***		
	(0.866)	*(0.000)*			*(0.100)*	*(0.000)*		
PISA 2000	-0.266	0.250	-0.041		0.531***	0.738***	0.740***	
	(0.163)	*(0.263)*	*(0.862)*		*(0.003)*	*(0.000)*	*(0.000)*	
PISA 2003	0.036	0.500**	0.274	0.384**	0.577***	0.708***	0.893***	0.756***
	(0.856)	*(0.021)*	*(0.257)*	*(0.023)*	*(0.001)*	*(0.000)*	*(0.000)*	*(0.000)*

Notes: Columns (1)-(4): correlations among exclusion rates across tests. Columns (5)-(8): correlations among non-response rates across tests. *p*-values in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3: Sample coverage and the role of test scores in growth regressions

Test-score measure:	All grades and years (AA)			Lower secondary, 1995-2003 (LR)	LR instrumented by AA	LR instrumented by tests before 1985
	(1)	(2)[a]	(3)	(4)	(5)[b]	(6)[b]
Test score	1.980***	1.741***	1.690***	1.338***	1.396***	1.651***
	(0.217)	*(0.228)*	*(0.278)*	*(0.214)*	*(0.227)*	*(0.429)*
Years of schooling 1960	0.026	0.041	0.028	0.068	0.060	0.114
	(0.078)	*(0.074)*	*(0.079)*	*(0.074)*	*(0.075)*	*(0.111)*
GDP per capita 1960	-0.302***	-0.294***	-0.310***	-0.320***	-0.320***	-0.362***
	(0.055)	*(0.051)*	*(0.052)*	*(0.052)*	*(0.052)*	*(0.085)*
Enrollment rate			0.009	0.011	0.010	-0.007
			(0.011)	*(0.010)*	*(0.010)*	*(0.041)*
Exclusion rate			-0.055	-0.050	-0.049	-0.019
			(0.058)	*(0.057)*	*(0.057)*	*(0.075)*
Non-response rate			0.016	0.012	0.013	0.003
			(0.015)	*(0.015)*	*(0.015)*	*(0.020)*
Constant	-4.737***	-3.788***	-4.255***	-2.954***	-3.071***	-2.741
	(0.855)	*(0.863)*	*(0.962)*	*(0.818)*	*(0.832)*	*(2.996)*
No. of countries	50	45	45	45	45	20
R^2 (adj.)	0.728	0.685	0.680	0.689	0.688	0.777
F-test (3 coverage rates)			0.79	0.74	0.68	0.03
p-value			*(0.505)*	*(0.533)*	*(0.571)*	*(0.993)*
F-test (instr. in 1st stage)					311.92	32.14

Notes: Dependent variable: average annual growth rate in GDP per capita, 1960-2000. Test score is average of math and science. See Hanushek and Woessmann (2009) for details on the basic specification. AA = all grades, all years. LR = lower secondary, recent years (1995-2003). Standard errors in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

a. Sample of countries with available information on measures of sample coverage.

b. Two-stage least-squares regression.

Table 4: Sample coverage and institutional effects in education production functions

	(1)	(2)	(3)	(4)[a]
Share of privately operated schools	57.585***	56.610***		
	(8.355)	*(9.239)*		
Share of government funding	81.839***	81.677***		
	(22.327)	*(25.595)*		
External exit exams (EEE)	25.338*	21.625**		
	(10.054)	*(10.283)*		
Assessments used for retention/promotion	12.185***	12.430***		
	(1.631)	*(1.663)*		
Internal monitoring of teacher lessons	4.557***	5.601***		
	(1.343)	*(1.391)*		
External monitoring of teacher lessons	3.796***	3.793***		
	(1.415)	*(1.416)*		
Assessments used for external comparisons	2.134*	3.172**		
	(1.259)	*(1.291)*		
Assessments used to group students	-6.065***	-5.344***		
	(1.301)	*(1.325)*		
Autonomy in formulating budget	-9.609***	-10.332***		
	(2.178)	*(2.215)*		
EEE x Autonomy in formulating budget	9.143***	8.746***		
	(3.119)	*(3.154)*		
Autonomy in establishing starting salaries	-8.632***	-5.478*		
	(3.251)	*(3.280)*		
EEE x Autonomy in establishing starting salaries	5.868	3.810		
	(3.980)	*(3.988)*		
Autonomy in determining course content	0.175	0.669		
	(1.907)	*(1.915)*		
EEE x Autonomy in determining course content	3.224	3.405		
	(2.858)	*(2.876)*		
Autonomy in hiring teachers	20.659***	20.896***		
	(2.249)	*(2.299)*		
EEE x Autonomy in hiring teachers	-28.935***	-27.005***		
	(3.365)	*(3.425)*		
Enrollment rate		0.143	2.424***	1.900
		(0.300)	*(0.382)*	*(1.751)*
Exclusion rate		0.577	-3.225	-2.459
		(.300)	*(2.091)*	*(2.094)*
Non-response rate		-0.523*	0.291	0.262
		(0.302)	*(0.440)*	*(0.437)*
Students	219,794	219,794	219,794	184,956
Schools	8,245	8,245	8,245	6,962
Countries	29	29	29	27
R^2	0.390	0.391	0.070	0.005
F-test (3 coverage rates)		0.98	14.42	0.87
p-value		*0.419*	*0.000*	*0.470*

Notes: Dependent variable: PISA 2003 international mathematics test score. Sample: OECD countries. Least-squares regressions weighted by students' sampling probability. The models additionally control for 15 variables of student characteristics, 17 variables of family background, 10 variables of school inputs, imputation dummies, and interaction terms between imputation dummies and the variables. See Woessmann, Luedemann, Schuetz, and West (2009) and Hanushek and Woessmann (forthcoming) for details on the basic specification. Robust standard errors adjusted for clustering at the school level in parentheses (clustering at country level for all country-level variables, which here are private operation, government funding, external exit exams, and the three measures of sample coverage). Significance level (based on clustering-robust standard errors): *** 1 percent, ** 5 percent, * 10 percent.

a. Sample of countries with enrollment rates of at least than 90 percent (excludes Mexico and Turkey).