Economics Letters 110 (2011) 79–82

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Sample selectivity and the validity of international student achievement tests in economic research

Eric A. Hanushek [a,*], Ludger Woessmann [b]

[a] Hoover Institution, Stanford University, CESifo, and NBER, United States
[b] University of Munich, Ifo Institute for Economic Research, CESifo, and IZA, Germany

ARTICLE INFO

Article history:
Received 27 March 2010
Received in revised form 24 September 2010
Accepted 26 October 2010
Available online xxxx

JEL classification:
I20
O40
C83

Keywords:
Sample selection
International student achievement tests
Economic growth

ABSTRACT

Larger rates of exclusion, non-response, and age-specific enrollment are related to better country average scores on international student achievement tests. But accounting for sample selectivity does not alter existing evidence that academic achievement enters importantly in economic growth regressions.

Introduction

Economic research has made increasing use of international student achievement data, 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, implying that they are unrelated to the associations of interest in these economic analyses.

To critics of international comparisons, "The basic problem is student selectivity: … the average 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)). Simple calculations indicate that sample bias certainly has the potential to move country mean scores substantially. For example, if exclusion propensity and student achievement are bivariate normally distributed and correlated at 0.5, a 10% exclusion rate – not uncommon in some countries – leads to an upward bias in the resulting country mean score of 10% of a standard deviation (see Organisation for Economic Co-operation and Development (2007)).

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. If idiosyncratic, sample selectivity introduces classical measurement error that works against finding statistically significant associations. But, economic

* Corresponding author. Hoover Institution, Stanford University, Stanford, CA 94305-6010, United States. Tel.: +1 650 / 736 0942.
E-mail address: hanushek@stanford.edu (E.A. Hanushek).

[1] See Hanushek and Woessmann (2011) for a review of the extensive economic literature on international educational achievement, including Hanushek and Kimko (2000), Barro (2001), Bosworth and Collins (2003), Ciccone and Papaioannou (2009), and Hanushek and Woessmann (2008, 2009).

[2] Hanushek and Woessmann (2010) provide additional results, literature references and data sources, and evidence that sample selectivity also does not affect results of typical international education production functions.

0165-1765/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2010.10.022

Table 1
Sample coverage — descriptive statistics and correlation with test scores.

Source of sample selection problems	Mean (Std. dev.) (1)	Min Max (2)	Correlation with Test score (3)	Enrollment rate (4)	Exclusion rate (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%, ** 5%, * 10%.

growth applications that use averages of scores across several tests lessen the importance of any idiosyncratic measurement error since the error variance is reduced by averaging. When sample selectivity is persistent over 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 if it is correlated with the error term of the equation of interest does systematic sample selectivity introduce bias to econometric analyses.

Sample selection and average test scores

Our empirical analysis focuses on sample selectivity for 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. Since the mid-1990s the major international testing cycles provide detailed documentation of the extent to which each participating country covered the underlying student population in its sampling. The Trends in International Mathematics and Science Study (TIMSS), conducted in 1995, 1999, and 2003, has a common target population of students enrolled in the upper of the two adjacent grades that contain the largest proportion of 13-year-old students. The Programme for International Student Assessment (PISA), conducted in 2000 and 2003, has a target population of 15-year-old students.

There are three main sources of sample selectivity. Because each may have very different impacts on the validity of testing and the importance of statistical bias, we will separately deal with each of them. First, both tests allow exclusions for small geographically remote schools, for schools focused on students with intellectual or functional disabilities, and within schools for individual students in the latter group or with limited proficiency in the test language. 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%.

Second, 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% both at the school level and at the student level (80% at the student level in PISA).

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.

Third, testing is always focused on students in school. Part of the children in the tested age range may no longer be in school. This problem is not associated with the testing so much as with the character of schooling in each country. Here, however, the direction of bias is unclear. 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, leading potentially to 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. Column 3 reports the correlations of the three components of sample selection with reported mean test performance. 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 national testing authority and the larger the share of schools and students sampled but not participating, the higher the reported country mean test score. At the same time, enrollment rates are also positively correlated with test scores, suggesting no simple upward

Table 2
Sample coverage — correlation across tests.

	Exclusion rate				Non-response rate			
	TIMSS			PISA	TIMSS			PISA
	1995 (1)	1999 (2)	2003 (3)	2000 (4)	1995 (5)	1999 (6)	2003 (7)	2000 (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%, ** 5%, * 10%.

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%, ** 5%, * 10%.
a. Sample of countries with available information on measures of sample coverage.
b. Two-stage least-squares regression.

bias where a substantial share of the age group is not enrolled in school.

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. 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 understand persistence of sampling issues, 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).[3] 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.

Sample selection and the results of growth regressions

Economists have extensively used international test scores to model cross-country growth differences. The impact of sample selection on their results can be illustrated by introducing measures of test participation rates into a representative published model. We employ the basic growth regression framework of Hanushek and Woessmann (2008), replicated in the first column of Table 3, which expresses the average annual growth rate in real GDP per capita over 1960–2000 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. Column (2) 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.[4]

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.

To this point, the test score measure refers to all international achievement tests, whereas our sampling information refers only to the five tests conducted at the lower secondary level since 1995. In column (4), we therefore use as a test score measure the average of just these five tests. 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.

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 variation on the early tests (1964–1985), ensuring that only variation traceable to the early tests is used. While this reduces the sample to the 20 countries participating in the early tests, the qualitative result on the effect of test scores on economic growth is

[3] 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.

[4] 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 causal interpretation.

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). The latter specification uses only test score variation for identification that mostly pre-dates growth rates and that at the same time is related to tests for which we have the relevant sampling information as control variables.

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.

Acknowledgments

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.

References

Barro, Robert J., 2001. Human capital and growth. American Economic Review 91 (2), 12–17.

Bosworth, Barry P., Collins, Susan M., 2003. The empirics of growth: an update. Brookings Papers on Economic Activity (2), 113–206.

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

Hanushek, Eric A., Kimko, Dennis D., 2000. Schooling, labor force quality, and the growth of nations. American Economic Review 90 (5), 1184–1208 (December).

Hanushek, Eric A., Woessmann, Ludger, 2008. The role of cognitive skills in economic development. Journal of Economic Literature 46 (3), 607–668 (September).

Hanushek, Eric A., Woessmann, Ludger, 2009. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. NBER Working Paper 14633. Cambridge, MA: National Bureau of Economic Research.

Hanushek, Eric A., Woessmann, Ludger, 2010. Sample selectivity and the validity of international student achievement tests in economic research. NBER Working Paper 15867. Cambridge, M:, National Bureau of Economic Research.

Hanushek, Eric A., Woessmann, Ludger, 2011b. The economics of international differences in educational achievement. In: Hanushek, Eric A., Machin, Stephen, Woessmann, Ludger (Eds.), Handbook of the Economics of Education, Vol. 3. Amsterdam: North Holland.

Organisation for Economic Co-operation and Development, 2007. PISA 2006: science competencies for tomorrow's world, Vol. 1 — Analysis. OECD, Paris.

Rotberg, Iris C., 1995. Myths about test score comparisons. Science 270 (5241), 1446–1448 (December 1).