Testing

Annika B. Bergbauer, Eric A. Hanushek, and Ludger Woessmann

Online Appendix

September 28, 2021

Data Appendix: Sources and Construction of Testing Measures

We derive a series of measures of different forms student testing over the period 2000-2015 from the PISA school background questionnaires and other sources. Information on testing is classified into four categories with varying strength of generated incentives: SCOMP, SINT, IRPT, and TMON (see Appendix A.1-A.4). We aggregate each assessment measure to the country-by-wave level. In Appendix A.5, we discuss how we combine the different indicators into an aggregate measure for each of the four testing categories. Details on the precise underlying survey questions and any changes in question wording over time are found in Appendix Table A3.

A.1 Standardized Testing with External Comparison (SCOMP)

Drawing on four different sources, we combine four separate indicators of standardized testing designed to allow for external comparisons.

First, from the PISA school background questionnaires, we measure the share of schools in each participating country that is subject to assessments for external comparison. In particular, school principals respond to the question, “In your school, are assessments of 15-year-old students used to compare the school to district or national performance?” Figure A2 provides a depiction of the evolution of this measure from 2000 to 2015 for each country.

Second, in the 2015 version of its Education at a Glance (EAG) publication, the OECD (2015) published an indicator of the existence of national/central examinations at the lower secondary level together with the year that is was first established. The data were collected by experts and institutions working within the framework of the OECD Indicators of Education Systems (INES) program in a 2014 OECD-INES Survey on Evaluation and Assessment. National examinations are defined as “standardized student tests that have a formal consequence for students, such as an impact on a student's eligibility to progress to a higher level of education or to complete an officially-recognized degree" (OECD 2015, p. 483). According to this measure, five of the 37 countries with available data have introduced national standardized exams in lower secondary school between 2000 and 2015.[1]

Third, following a very similar concept, the Eurydice unit of the Education, Audiovisual and Culture Executive Agency (EACEA) of the European Commission provides information on the year of first full implementation of national testing in a historical overview of national testing of students in Europe (Eurydice 2009; see also Braga, Checchi, and Meschi 2013). In particular, they classify national tests for taking decisions about the school career of individual students, including tests for the award of certificates, promotion at the end of a school year, or streaming at the end of primary or lower secondary school. We extend their measure to the year 2015 mostly based on information provided in the Eurydice (2017) online platform. During our period of observation, eight of the 18 European countries introduced national tests for career decisions and two abolished them.

Fourth, Leschnig, Schwerdt, and Zigova (2017) compile a dataset of the existence of central exit examinations at the end of secondary school over time for the 31 countries participating in the Programme for the International Assessment of Adult Competencies (PIAAC). They define central exit exams as "a written test at the end of secondary school, administered by a central authority, providing centrally developed and curriculum based test questions and covering core subjects." Following Bishop (1997), they do not include commercially prepared tests or university entrance exams that do not have direct consequences for students passing them. Central exit exams "can be organized either on a national level or on a regional level and must be

[1] In federal countries, all system-level indicator measures are weighted by population shares in 2000.

mandatory for all or at least the majority of a cohort of upper secondary school." We extend their time period, which usually ends in 2012, to 2015. Five of the 30 countries in our sample introduced central exit exams over our 15-year period, whereas two countries abandoned them.

A.2 Standardized Testing for Internal Comparison (SINT)

Beyond externally comparative testing, the PISA school background questionnaire also provides three additional measures of standardized testing that allow for different types of monitoring but do not readily provide for external comparison.

First, school principals answer the question, "Generally, in your school, how often are 15-year-old students assessed using standardized tests?" Answer categories start with "never" and then range from "1-2 times a year" ("yearly" in 2000) to more regular uses. We code a variable that represents the share of schools in a country that use standardized testing at all (i.e., at least once a year).

Second, school principals provide indicators on the following battery of items: "During the last year, have any of the following methods been used to monitor the practice of teachers at your school?" Apart from a number of non-test-based methods of teacher practice monitoring, one of the items included in the battery is "tests or assessments of student achievement." We use this to code the share of schools in a country that monitors teacher practice by assessments.

Third, school principals are asked, "In your school, are achievement data used in any of the following accountability procedures?" One consistently recorded item is whether "achievement data are tracked over time by an administrative authority," which allows us to construct a measure of the share of schools in a country for which an administrative authority tracks achievement data. The reference to over-time tracking by administrations indicates that the achievement data are standardized to be comparable over time.

A.3 Internal Reporting (IRPT)

The PISA school background questionnaire also provides information on three testing policies where tests are not necessarily standardized and are mostly used for pedagogical management.

In particular, school principals report on the prevalence of assessments of 15-year-old students in their school for purposes other than external comparisons. Our first measure of IRPT captures whether assessments are used "to inform parents about their child's progress." The second measure covers the use of assessments "to monitor the school's progress from year to year." Each measure is coded as the share of schools in a country using the respective type of internal assessments.

The question on use of achievement data in accountability procedures referred to above also includes an item indicating that "achievement data are posted publicly (e.g. in the media)." Our third measure thus captures the share of schools in a country where achievement data are posted publicly. In the questionnaire item, the public posting is rather vaguely phrased and is likely to be understood by school principals to include such practices as posting the school mean of the grade point average of a graduating cohort, derived from teacher-defined grades rather than any standardized test, at the school's blackboard.

A.4 Teacher Monitoring (TMON)

Finally, the PISA school background questionnaire provides three additional measures of internal monitoring that are all focused on teachers.

First, again reporting on the prevalence of assessments of 15-year-old students in their school, school principals report whether assessments are used "to make judgements about teachers' effectiveness."

The battery of methods used to monitor teacher practices also includes two types of assessments based on observations of teacher practices by other persons rather than on student achievement tests. Our second measure in this area captures the share of schools where the practice of teachers is monitored through “principal or senior staff observations of lessons.” Our third measure captures whether “observation of classes by inspectors or other persons external to the school” are used to monitor the practice of teachers.

A.5 Constructing Combined Measures for the Four Testing Categories

Many of the separate testing indicators are obviously correlated with each other, in particular within each of the four groups of testing categories. For example, the correlation between the EAG measure of national standardized exams in lower secondary school and the Eurydice measure of national tests for career decisions is 0.59 in our pooled dataset (at the country-by-wave level) and 0.54 after taking out country and year fixed effects (which reflects the identifying variation in our model). Similarly, the two internal-testing measures of assessments to inform parents and assessments to monitor school progress are correlated at 0.42 in the pooled data and 0.57 after taking out country and year fixed effects (all highly significant).

While these correlations are high, there is also substantial indicator-specific variation. These differences may reflect slight differences in the concepts underlying the different indicators and different measurement error in the different indicators, but also substantive differences in the measured assessment dimensions. In our main analysis, we combine the individual indicators into one measure for each of the four testing categories, but in additional analyses we report results for each indicator separately.

Our construction of the combined measures takes into account that the different indicators are available for different sets of waves and countries, as indicated in Appendix Table A4.

Before combining the indicators, we therefore impute missing observations in the aggregate country-by-wave dataset from a linear time prediction within each country. That is, for each country with at least some observations on a given indicator, we regress the available data for the indicator on a time variable and use the predicted values of this regression to impute the missing data for this country. We then construct the combined measures of the four testing categories as the simple average of the individual imputed indicators in each category for which data are available in a country. To ensure that the imputation does not affect our results, all our regression analyses include a full set of imputation dummies that equal one for each underlying indicator that was imputed and zero otherwise.

The combined measures of the four testing categories are also correlated with each other (Table A5). In the pooled dataset of 303 country-by-wave observations, the correlations range from 0.278 between SCOMP and TMON to 0.583 between SINT and IRPT. After taking out country and year fixed effects, the correlations are lowest between SCOMP and all other categories (all below 0.2), moderate between SINT and the other categories (all below 0.3), and largest between IRPT and TMON (0.485). Because of potential multicollinearity, we run our analyses both for each aggregate assessment category separately and considering all four categories simultaneously.

Appendix References

Bishop, John H. 1997. "The Effect of National Standards and Curriculum-Based Exams on Achievement." *American Economic Review* 87(2): 260-264.

Braga, Michela, Daniele Checchi, and Elena Meschi. 2013. "Educational Policies in a Long-run Perspective." *Economic Policy* 28(73): 45-100.

Eurydice. 2009. *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. Brussels: European Commission; Education, Audiovisual and Culture Executive Agency (EACEA), Eurydice.

__________. 2017. Online Platform, ec.europa.eu/eurydice. Brussels: Education Audiovisual & Culture Executive Agency (EACEA), Eurydice Unit.

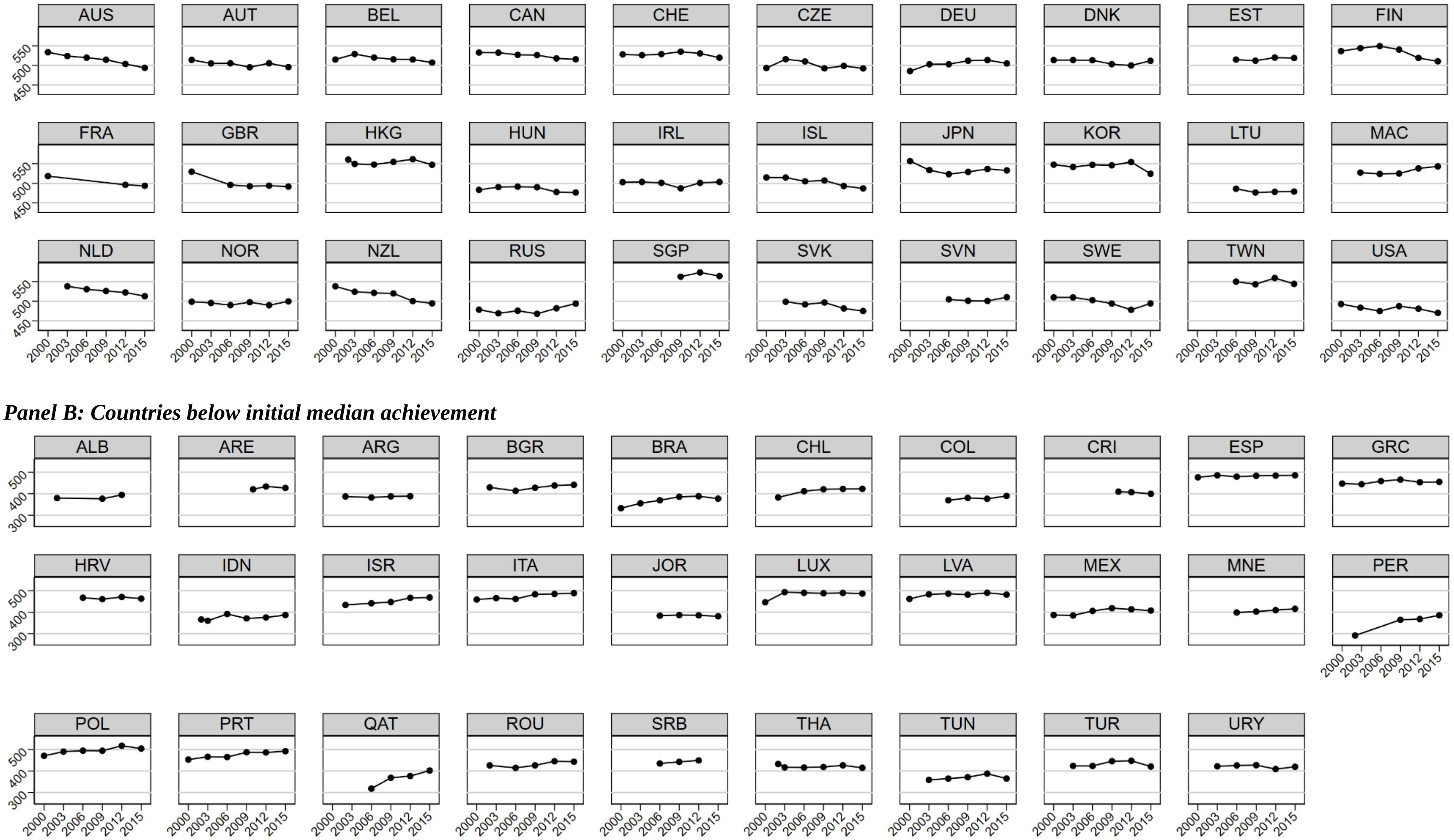
Leschnig, Lisa, Guido Schwerdt, and Katarina Zigova. 2017. "Central School Exams and Adult Skills: Evidence from PIAAC." Unpublished manuscript, University of Konstanz.

OECD. 2015. *Education at a Glance 2015: OECD Indicators*. Paris: Organisation for Economic Co-operation and Development.

Appendix Figures and Tables

Figure A1: PISA math achievement in 2000-2015

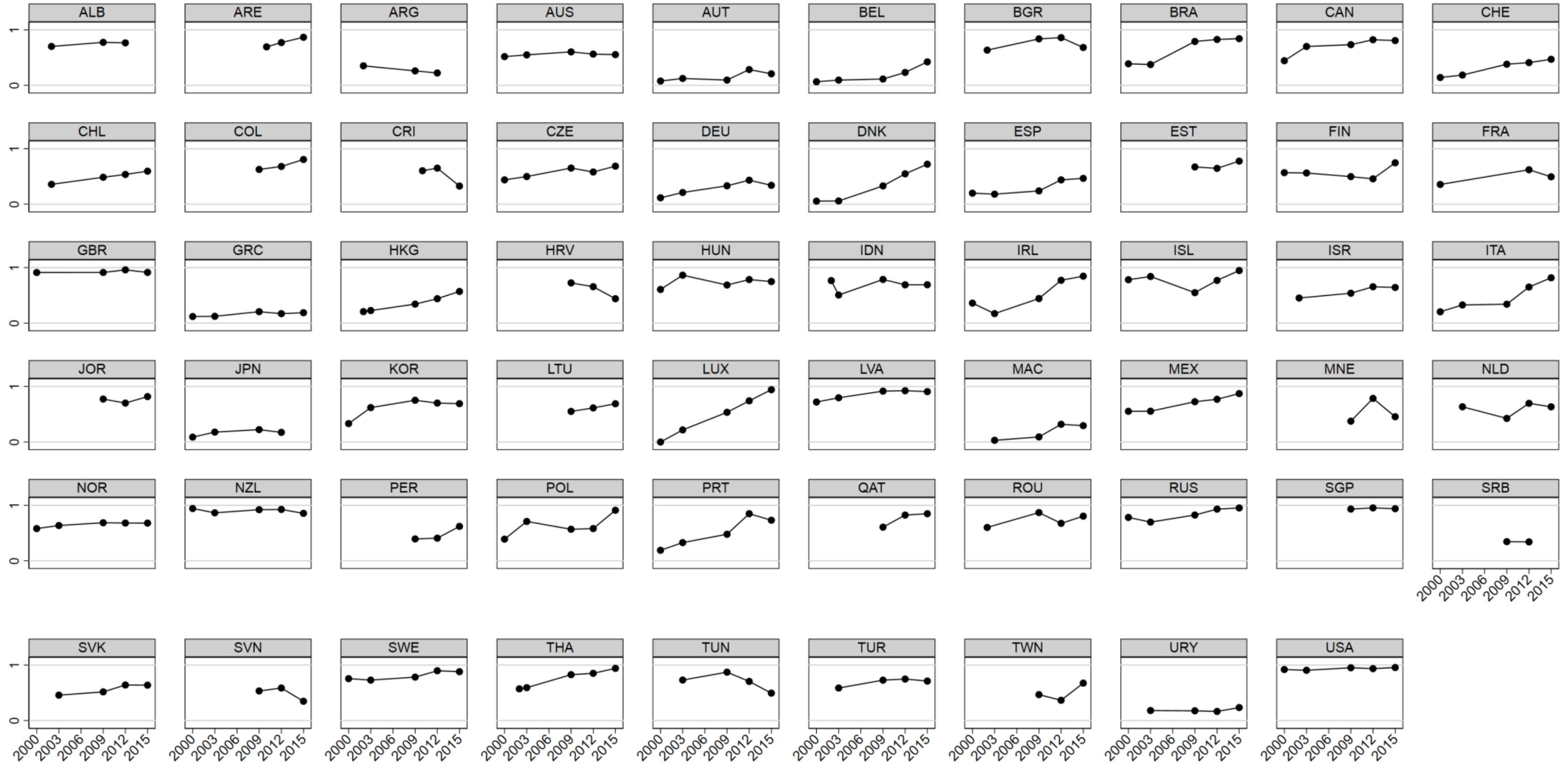
Panel A: Countries above initial median achievement



Panel B: Countries below initial median achievement

Notes: Country mean achievement in PISA math test. Country sample split at median of initial achievement level for expositional reasons. Country identifiers are listed in Appendix Table A1. Own depiction based on PISA micro data.

Figure A2: School-focused external comparison in 2000-2015



Notes: Country share of schools with assessments for external comparison. Country identifiers are listed in Appendix Table A1. Own depiction based on PISA micro data.

Table A1: Selected indicators by country

	OECD	PISA math score		School-focused external comparison		National standardized exams in lower sec. school		National tests for career decisions		Central exit exams	
	2015	2000	2015	2000	2015	2000	2015	2000	2015	2000	2015
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Albania (ALB) [a]	0	380	395	0.70	0.77	.	.	.	.	.	.
Argentina (ARG) [a]	0	387	389	0.35	0.22	.	.	.	.	.	.
Australia (AUS)	1	534	494	0.52	0.55	0	0	.	.	0.80	1
Austria (AUT)	1	514	496	0.08	0.21	0	0	.	.	0	0
Belgium (BEL)	1	515	507	0.07	0.42	0	0.32	0	0.32	.	.
Brazil (BRA)	0	333	377	0.39	0.84	0	0	.	.	.	.
Bulgaria (BGR) [a]	0	430	442	0.64	0.68	.	.	0	1	.	.
Canada (CAN)	1	533	516	0.44	0.81	0	0	.	.	0.54	0.54
Chile (CHL) [a]	1	383	423	0.36	0.60	0	0	.	.	0	0
Colombia (COL) [c]	0	370	390	0.63	0.81	0	0	.	.	.	.
Costa Rica (CRI) [e]	0	410	400	0.61	0.33	.	.	.	.	.	.
Croatia (HRV) [c]	0	467	463	0.73	0.44	.	.	.	.	.	.
Czech Republic (CZE)	1	493	492	0.44	0.69	0	0	0	0	0	1
Denmark (DNK)	1	514	512	0.06	0.72	1	1	1	1	1	1
Estonia (EST) [c]	1	515	519	0.67	0.78	1	1	.	.	1	0
Finland (FIN)	1	536	511	0.57	0.75	0	0	.	.	1	1
France (FRA)	1	518	494	0.36	0.50	1	1	.	.	1	1
Germany (DEU)	1	485	505	0.12	0.34	.	.	0	1	0.43	0.95
Greece (GRC)	1	447	455	0.12	0.19	0	0	0	0	1	0
Hong Kong (HKG) [a]	0	560	547	0.21	0.57	.	.	.	.	.	.
Hungary (HUN)	1	483	477	0.61	0.75	0	0	.	.	.	.
Iceland (ISL)	1	515	487	0.78	0.95	0	0	1	0	.	.
Indonesia (IDN) [a]	0	366	387	0.77	0.69	.	.	.	.	1	1
Ireland (IRL)	1	503	504	0.36	0.85	1	1	1	1	1	1
Israel (ISR) [a]	1	434	468	0.45	0.64	0	0	.	.	1	1
Italy (ITA)	1	459	489	0.21	0.82	1	1	0	1	1	1
Japan (JPN)	1	557	533	0.09	0.17	0	0	.	.	1	1
Jordan (JOR) [c]	0	384	381	0.77	0.82	.	.	.	.	.	.
Korea (KOR)	1	548	524	0.33	0.69	0	0	.	.	1	1
Latvia (LVA)	1	462	482	0.72	0.91	1	1	1	1	.	.

(continued on next page)

Table A1 (continued)

	OECD	PISA math score		School-focused external comparison		National standardized exams in lower sec. school		National tests for career decisions		Central exit exams	
	2015	2000	2015	2000	2015	2000	2015	2000	2015	2000	2015
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Lithuania (LTU) [c]	0	486	479	0.55	0.69	.	.	0	0	1	1
Luxembourg (LUX) [b]	1	446	487	0.00	0.94	0	0	1	1	.	.
Macao (MAC)	0	527	543	0.03	0.30	.	.	.	.	.	.
Mexico (MEX)	1	387	408	0.55	0.87	0	0	.	.	.	.
Montenegro (MNE) [c]	0	399	416	0.38	0.46	.	.	.	.	.	.
Netherlands (NLD) [b]	1	538	513	0.64	0.63	1	1	1	1	1	1
New Zealand (NZL)	1	538	494	0.94	0.86	0	0	.	.	1	1
Norway(NOR)	1	499	500	0.58	0.68	0	1	0	1	1	1
Peru (PER) [a]	0	292	386	0.40	0.62	.	.	.	.	.	.
Poland (POL)	1	471	505	0.39	0.91	0	1	0	1	0	1
Portugal (PRT)	1	453	493	0.19	0.73	0	1	0	1	.	.
Qatar (QAT) [c]	0	318	402	0.61	0.85	.	.	.	.	.	.
Romania (ROU) [a]	0	426	443	0.60	0.81	.	.	0	1	.	.
Russia (RUS)	0	478	494	0.78	0.95	.	.	.	.	.	.
Serbia (SRB) [c]	0	435	449	0.35	0.34	.	.	.	.	.	.
Singapore (SGP) [d]	0	563	564	0.93	0.94	.	.	.	.	1	1
Slovak Republic (SVK) [b]	1	499	475	0.46	0.64	0	0	.	.	0	1
Slovenia (SVN) [c]	1	505	510	0.54	0.35	0	0	0	0	1	1
Spain (ESP)	1	476	486	0.20	0.47	0	0	.	.	0	0
Sweden (SWE)	1	510	494	0.76	0.88	0	0	1	1	0	0
Switzerland (CHE)	1	528	520	0.14	0.47	.	.	.	.	.	.
Taiwan (TWN) [c]	0	550	544	0.47	0.68	.	.	.	.	.	.
Thailand (THA) [a]	0	433	415	0.57	0.94	.	.	.	.	.	.
Tunisia (TUN) [b]	0	359	365	0.73	0.50	.	.	.	.	.	.
Turkey (TUR) [b]	1	424	421	0.59	0.71	1	1	.	.	0	0
United Arab Emirates (ARE) [e]	0	421	427	0.69	0.87	.	.	.	.	.	.
United Kingdom (GBR)	1	530	492	0.91	0.91	0	0	0.87	0	1	1
United States (USA)	1	493	470	0.92	0.96	0	1	.	.	0.07	0.07
Uruguay (URY) [b]	0	422	420	0.18	0.24	.	.	.	.		
Country average	0.59	465	469	0.48	0.66	0.23	0.35	0.39	0.67	0.66	0.72

Notes: PISA data: Country means, based on non-imputed data for each variable, weighted by sampling probabilities. "." = not available. [a-e] "2000" PISA data refer to country's initial PISA participation in [a] 2002, [b] 2003, [c] 2006, [d] 2009, [e] 2010.

Table A2: Descriptive statistics and complete model of basic specification

	Descriptive statistics			Basic model	
	Mean	Std. dev.	Share imputed	Coeff.	Std. err.
Standardized testing with external comparison (SCOMP)				37.304***	*(6.530)*
× initial score				-0.246***	*(0.085)*
Standardized testing for internal comparison (SINT)				67.772***	*(17.139)*
× initial score				-0.776***	*(0.175)*
Internal reporting (IRPT)				-13.858	*(12.216)*
× initial score				0.161	*(0.100)*
Teacher monitoring (TMON)				10.432	*(25.005)*
× initial score				-0.478*	*(0.249)*
Student and family characteristics					
Female	0.500	0.500	0.001	-11.557***	*(0.946)*
Age (years)	15.77	0.298	0.001	12.284***	*(0.921)*
Immigration background					
Native student	0.891	0.306	0.034		
First generation migrant	0.051	0.216	0.034	-8.322	*(4.635)*
Second generation migrant	0.058	0.230	0.034	-2.772	*(2.736)*
Other language than test language or national dialect spoken at home	0.107	0.301	0.061	-15.133***	*(2.309)*
Parents' education					
None	0.018	0.132	0.031		
Primary	0.064	0.243	0.031	9.138***	*(2.228)*
Lower secondary	0.100	0.295	0.031	10.814***	*(2.421)*
Upper secondary I	0.089	0.280	0.031	20.951***	*(2.984)*
Upper secondary II	0.271	0.438	0.031	26.363***	*(2.559)*
University	0.457	0.490	0.031	36.135***	*(2.538)*
Parents' occupation					
Blue collar low skilled	0.082	0.269	0.041		
Blue collar high skilled	0.094	0.286	0.041	8.401***	*(1.153)*
White collar low skilled	0.169	0.366	0.041	15.520***	*(1.108)*
White collar high skilled	0.335	0.463	0.041	35.601***	*(1.552)*
Books at home					
0-10 books	0.168	0.369	0.026		
11-100 books	0.478	0.493	0.026	30.297***	*(1.908)*
101-500 books	0.280	0.444	0.026	64.817***	*(2.426)*
More than 500 books	0.074	0.258	0.026	73.718***	*(3.433)*

(continued on next page)

Table A2 (continued)

	Descriptive statistics			Basic model	
	Mean	Std. dev.	Share imputed	Coeff.	Std. err.
School characteristics					
Number of students	841.7	717.2	0.093	0.012***	*(0.002)*
Privately operated	0.196	0.388	0.071	7.500*	*(4.396)*
Share of government funding	0.829	0.269	0.106	-16.293***	*(4.596)*
Share of fully certified teachers at school	0.849	0.269	0.274	6.662**	*(2.793)*
Shortage of math teachers	0.196	0.390	0.041	-5.488***	*(1.031)*
Teacher absenteeism					
No	0.336	0.429	0.213		
A little	0.475	0.448	0.213	-0.325	*(1.175)*
Some	0.145	0.315	0.213	-6.089***	*(1.556)*
A lot	0.043	0.183	0.213	-7.715***	*(2.413)*
School's community location					
Village or rural area (<3,000)	0.103	0.298	0.056		
Town (3,000-15,000)	0.202	0.393	0.056	5.238***	*(1.768)*
Large town (15,000-100,000)	0.312	0.454	0.056	9.935***	*(2.148)*
City (100,000-1,000,000)	0.242	0.420	0.056	14.209***	*(2.594)*
Large city (>1,000,000)	0.141	0.343	0.056	17.482***	*(3.447)*
Country characteristics					
Academic-content autonomy	0.611	0.264	-	-11.666	*(8.826)*
Academic-content autonomy × Initial GDP p.c.	4.998	8.153	-	1.871***	*(0.475)*
GDP per capita (1,000 $)	26.51	21.51	-	0.009	*(0.123)*
Country fixed effects; year fixed effects				Yes	
Student observations	2,193,026			2,094,856	
Country observations	59			59	
Country-by-wave observations	303			303	
R^2				0.393	

Notes: Descriptive statistics: Mean: international mean (weighted by sampling probabilities). Std. dev.: international standard deviation. Share imputed: share of missing values in the original data, imputed in the analysis. Basic model: Full results of the specification reported in first column of Table 3. Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability. Regression includes imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A3: Measures of student testing: Sources and definitions

	Source	Countries	Waves	Definition	Deviation in wording in specific waves
	(1)	(2)	(3)	(4)	(5)
Standardized testing with external comparison (SCOMP)					
School-focused external comparison	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To compare the school to district or national performance.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
National standardized exams in lower secondary school	OECD (2015)	OECD EAG sample	2000-2015	National/central examinations (at the lower secondary level), which apply to nearly all students, are standardized tests of what students are expected to know or be able to do that have a formal consequence for students, such as an impact on a student's eligibility to progress to a higher level of education or to complete an officially recognized degree.	
National tests for career decisions	Eurydice (2009)	EU countries	2000-2015	Year of first full implementation of national testing, ISCED levels 1 and 2: Tests for taking decisions about the school career of individual pupils, including tests for the award of certificates, or for promotion at the end of a school year or streaming at the end of ISCED levels 1 or 2.	
Central exit exams	Leschnig, Schwerdt, and Zigova (2017)	PIAAC sample	2000-2015	Exit examination at the end of secondary school: A central exam is a written test at the end of secondary school, administered by a central authority, providing centrally developed and curriculum based test questions and covering core subjects. (See text for additional detail.)	
Standardized testing for internal comparison (SINT)					
Standardized testing in tested grade	PISA school questionnaire	PISA sample	2000, 2003, 2009, 2015	Generally, in your school, how often are 15-year-old students assessed using standardized tests? More than "never."	2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2009: "using the following methods:" "standardized tests"; 2015: "using the following methods:" "mandatory standardized tests" or "non-mandatory standardized tests".
Student tests to monitor teacher practice	PISA school questionnaire	PISA sample	2003, 2009-2015	During the last year, have any of the following methods been used to monitor the practice of teachers at your school? Tests or assessments of student achievement.	2003 and 2012: "mathematics teachers" instead of "teachers"; 2009: "<test language> teachers" instead of "teachers"
Achievement data tracked by administrative authority	PISA school questionnaire	PISA sample	2006-2015	In your school, are achievement data used in any of the following accountability procedures? Achievement data are tracked over time by an administrative authority.	

(continued on next page)

Table A3 (continued)

	Source (1)	Countries (2)	Waves (3)	Definition (4)	Deviation in wording in specific waves (5)
Internal reporting (IRPT)					
Assessments to inform parents	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To inform parents about their child's progress.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
Assessments to monitor school progress	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To monitor the school's progress from year to year.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
Achievement data posted publicly	PISA school questionnaire	PISA sample	2006-2015	In your school, are achievement data used in any of the following accountability procedures? Achievement data are posted publicly (e.g. in the media).	
Teacher monitoring (TMON)					
Teacher effectiveness judged by assessments	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To make judgements about teachers' effectiveness.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
Teacher practice monitored by principal	PISA school questionnaire	PISA sample	2003, 2009-2015	During the last year, have any of the following methods been used to monitor the practice of teachers at your school? Principal or senior staff observations of lessons.	2003 and 2012: "mathematics teachers" instead of "teachers"; 2009: "<test language> teachers" instead of "teachers"
Teacher practice monitored by external inspectors	PISA school questionnaire	PISA sample	2003, 2009-2015	During the last year, have any of the following methods been used to monitor the practice of teachers at your school? Observation of classes by inspectors or other persons external to the school.	2003 and 2012: "mathematics teachers" instead of "teachers"; 2009: "<test language> teachers" instead of "teachers"

Notes: Own depiction based on indicated sources.

Table A4: Country observations by wave

	2000/02	2003	2006	2009/10	2012	2015	Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Standardized testing with external comparison (SCOMP)							
School-focused external comparison	39	37	–	58	59	55	248
National standardized exams in lower secondary school	30	29	35	35	36	36	201
National tests for career decisions	17	15	21	21	21	21	116
Central exit exams	23	22	28	29	30	30	162
Standardized testing for internal comparison (SINT)							
Standardized testing in tested grade	38	35	–	58	–	51	182
Student tests to monitor teacher practice	–	36	–	57	59	56	208
Achievement data tracked by administrative authority	–	–	53	58	59	56	226
Internal reporting (IRPT)							
Assessments to inform parents	40	37	–	58	59	55	249
Assessments to monitor school progress	40	37	–	58	59	55	249
Achievement data posted publicly	–	–	53	58	59	56	226
Teacher monitoring (TMON)							
Teacher effectiveness judged by assessments	40	37	–	58	59	55	249
Teacher practice monitored by principal	–	37	–	58	59	56	210
Teacher practice monitored by external inspectors	–	37	–	58	59	56	210

Notes: Own depiction based on PISA data and other sources. See Data Appendix for details.

Table A5: Correlation of four testing categories

	SCOMP	SINT	IRPT	TMON
	(1)	(2)	(3)	(4)
Raw correlations				
Standardized testing with external comparison (SCOMP)	1			
Standardized testing for internal comparison (SINT)	0.478	1		
Internal reporting (IRPT)	0.342	0.583	1	
Teacher monitoring (TMON)	0.278	0.562	0.364	1
Correlations after taking out country and year fixed effects				
SCOMP	1			
SINT	0.178	1		
IRPT	0.188	0.231	1	
TMON	0.169	0.298	0.485	1

Notes: Correlation coefficients in pooled dataset of 303 country-by-wave observations. All reported correlations are statistically significant at the 1 percent level.

Table A6: Disaggregation of standardized external comparison into school-focused and student-focused comparison

	Math	Science	Reading
	(1)	(2)	(3)
School-focused external comparison	25.015***	21.317**	23.480***
	(7.667)	*(8.246)*	*(7.291)*
Student-focused external comparison	17.309***	15.198***	14.481***
	(3.620)	*(3.883)*	*(3.753)*
Standardized testing for internal comparison (SINT)	-4.658	-8.333	-8.400
	(16.599)	*(15.007)*	*(14.602)*
Internal reporting (IRPT)	4.896	13.419	-16.890
	(13.686)	*(15.306)*	*(18.616)*
Teacher monitoring (TMON)	-35.424**	-27.374	-18.372
	(15.165)	*(16.656)*	*(16.373)*
Control variables	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Student observations	1,672,041	1,671,914	1,751,351
Country observations	42	42	42
Country-by-wave observations	230	230	230
R^2	0.348	0.315	0.321

Notes: Dependent variable: PISA test score in subject indicated in the header. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A7: Estimations for separate underlying testing indicators: Specification with average effects

	Math (1)	Science (2)	Reading (3)	Observations (4)	Countries (5)	Waves (6)	R^2 (7)
Standardized testing with external comparison (SCOMP)							
School-focused external comparison	13.797*	13.147*	16.058**	1,703,142	59	5	0.382
	(7.417)	*(6.598)*	*(6.227)*				
National standardized exams in lower secondary school	13.400**	14.272**	14.568**	1,517,693	36	6	0.326
	(5.508)	*(5.336)*	*(5.418)*				
National tests for career decisions	15.650***	11.144***	11.002***	676,732	21	6	0.264
	(1.701)	*(2.377)*	*(2.932)*				
Central exit exams	3.694	8.242	9.806	1,141,162	30	6	0.308
	(7.041)	*(6.575)*	*(6.551)*				
Standardized testing for internal comparison (SINT)							
Standardized testing in tested grade	15.497**	11.051	19.380***	1,198,463	59	4	0.386
	(7.244)	*(6.901)*	*(7.169)*				
Student tests to monitor teacher practice	-19.266*	0.305	-10.046	1,537,802	59	4	0.385
	(9.625)	*(9.785)*	*(6.329)*				
Achievement data tracked by administrative authority	-3.555	5.173	-1.677	1,713,976	59	4	0.394
	(9.266)	*(9.578)*	*(12.787)*				
Internal reporting (IRPT)							
Assessments to inform parents	7.923	14.664**	4.234	1,705,602	59	5	0.385
	(6.594)	*(6.974)*	*(7.912)*				
Assessments to monitor school progress	1.480	7.283	-1.598	1,705,602	59	5	0.385
	(5.343)	*(7.630)*	*(7.308)*				
Achievement data posted publicly	0.344	0.571	-16.954	1,713,976	59	4	0.394
	(8.371)	*(7.630)*	*(10.165)*				
Teacher monitoring (TMON)							
Teacher effectiveness judged by assessments	-4.065	3.110	-1.981	1,705,602	59	5	0.385
	(8.249)	*(9.619)*	*(7.810)*				
Teacher practice monitored by principal	-19.751	-10.893	-14.239	1,588,962	59	4	0.385
	(14.072)	*(10.793)*	*(10.062)*				
Teacher practice monitored by external inspectors	-13.152	-13.524	-17.553*	1,588,962	59	4	0.385
	(10.038)	*(8.898)*	*(10.306)*				

Notes: Each cell presents results of a separate regression. Dependent variable: PISA test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Number of observations and R^2 refer to the math specification. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A8: Correlation of testing reforms with other school policy measures

	Standardized testing with external comparison (SCOMP)	Standardized testing for internal comparison (SINT)	Internal reporting (IRPT)	Teacher monitoring (TMON)
	(1)	(2)	(3)	(4)
School autonomy	0.157	-0.007	-0.010	0.075
	(0.006)	*(0.899)*	*(0.857)*	*(0.195)*
School size	0.063	0.115	0.038	0.015
	(0.278)	*(0.046)*	*(0.507)*	*(0.801)*
Share of fully certified teachers at school	0.000	0.039	-0.022	-0.125
	(0.997)	*(0.494)*	*(0.708)*	*(0.030)*
Shortage of math teachers	0.019	0.118	-0.012	0.212
	(0.742)	*(0.040)*	*(0.834)*	*(0.000)*
Private vs. public school management	0.038	0.012	-0.115	0.021
	(0.509)	*(0.841)*	*(0.045)*	*(0.720)*
Share of government funding at school	-0.070	-0.103	0.089	0.054
	(0.223)	*(0.075)*	*(0.121)*	*(0.347)*

Notes: Correlation coefficients in pooled dataset of 303 country-by-wave observations, after taking out country and year fixed effects.

Table A9: Specification tests: Specification with average effects

	No teacher controls	No controls	Long difference (2000+2015 only)
	(1)	(2)	(3)
Standardized testing with external comparison (SCOMP)	28.429^{***} *(6.067)*	29.902^{***} *(6.619)*	61.184^{***} *(9.981)*
Standardized testing for internal comparison (SINT)	-4.271 *(14.502)*	0.218 *(13.187)*	-16.515 *(19.191)*
Internal reporting (IRPT)	10.776 *(12.001)*	13.052 *(10.514)*	19.131 *(26.395)*
Teacher monitoring (TMON)	-42.255^{***} *(15.604)*	-30.877^{*} *(16.250)*	-13.438 *(23.881)*
Teacher control variables	No	No	Yes
Other control variables	Yes	No	Yes
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Student observations	2,094,856	2,094,856	404,344
Country observations	59	59	38
Country-by-wave observations	303	303	76
R^2	0.390	0.256	0.365

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A10: Robustness tests: Specification with average effects

	OECD countries	Non-OECD countries	Control for exclusion rates	Without 2015	Rescaled test scale
	(1)	(2)	(3)	(4)	(5)
Standardized testing with external comparison (SCOMP)	29.303*** *(7.471)*	16.429* *(8.387)*	27.431*** *(6.160)*	31.205*** *(5.996)*	33.247*** *(8.937)*
Standardized testing for internal comparison (SINT)	4.671 *(15.292)*	-10.835 *(19.542)*	-5.817 *(13.900)*	-10.664 *(15.272)*	-10.906 *(15.499)*
Internal reporting (IRPT)	1.727 *(13.704)*	15.001 *(14.846)*	5.665 *(10.619)*	6.381 *(16.582)*	5.434 *(9.393)*
Teacher monitoring (TMON)	-25.693 *(16.190)*	-22.625 *(21.114)*	-35.308** *(15.460)*	-46.460** *(20.489)*	-29.108 *(21.312)*
Control variables	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Student observations	1,434,355	660,501	2,045,454	1,679,250	1,698,971
Country observations	35	24	59	59	58
Country-by-wave observations	197	106	289	247	223
R^2	0.283	0.441	0.388	0.399	n.a.

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. Rescaled test scale available for waves 2006-2015 only. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A11: Correlation of computer indicators in 2012 with change in PISA score from 2012 to 2015 at the country level

	Math	Science	Reading
	(1)	(2)	(3)
School			
Ratio of computers for education to students in respective grade	-0.015	-0.045	0.091
	(0.912)	*(0.744)*	*(0.503)*
Share of computers connected to Internet	-0.223*	-0.395***	-0.125
	(0.099)	*(0.003)*	*(0.360)*
School’s capacity to provide instruction hindered by:			
Shortage or inadequacy of computers for instruction	0.000	0.028	-0.029
	(0.998)	*(0.837)*	*(0.834)*
Lack or inadequacy of Internet connectivity	0.106	0.247*	0.040
	(0.438)	*(0.066)*	*(0.771)*
Shortage or inadequacy of computer software for instruction	0.091	0.059	0.083
	(0.503)	*(0.666)*	*(0.541)*
Student			
Computer at home for use for school work	0.034	0.240*	-0.162
	(0.805)	*(0.075)*	*(0.233)*
Number of computers at home	0.083	-0.043	0.181
	(0.544)	*(0.751)*	*(0.182)*
Educational software at home	-0.111	0.044	-0.238*
	(0.414)	*(0.746)*	*(0.077)*
Link to the Internet at home	0.043	0.221	-0.116
	(0.752)	*(0.102)*	*(0.394)*
Frequency of programming computers at school and outside of school	-0.150	-0.110	-0.003
	(0.270)	*(0.419)*	*(0.980)*
Weekly time spent repeating and training content from school lessons by working on a computer	0.095 *(0.485)*	0.071 *(0.604)*	0.030 *(0.826)*

Notes: Correlation between the respective computer indicator (2012) indicated in the first column with the change in PISA test scores (2012-215) in the subject indicated in the header. Sample: 56 country-level observations of countries participating in the PISA waves 2012 and 2015. *p*-values in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A12: Two-stage estimation: Panel model estimated at country-by-wave level

	Math	Science	Reading
	(1)	(2)	(3)
Standardized testing with external comparison (SCOMP)	38.088***	26.759**	46.405***
	(8.303)	*(10.291)*	*(10.236)*
× initial score	-0.223**	-0.114	-0.333***
	(0.104)	*(0.115)*	*(0.117)*
Standardized testing for internal comparison (SINT)	64.224***	90.318***	83.557***
	(22.277)	*(27.268)*	*(25.078)*
× initial score	-0.740***	-1.048***	-0.958***
	(0.226)	*(0.330)*	*(0.285)*
Internal reporting (IRPT)	-17.208	-14.606	-24.331
	(13.672)	*(17.558)*	*(19.135)*
× initial score	0.152	0.247	0.064
	(0.108)	*(0.167)*	*(0.210)*
Teacher monitoring (TMON)	12.788	16.643	-28.620
	(31.933)	*(31.970)*	*(36.782)*
× initial score	-0.476	-0.405	0.164
	(0.316)	*(0.350)*	*(0.344)*
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Country observations	59	59	59
Country-by-wave observations	303	303	302

Notes: Dependent variable: country-level aggregation of the residuals of a first-stage student-level regression that regresses the PISA test score in the subject indicated in the header on student gender, age, parental occupation, parental education, books at home, immigration status, language spoken at home, school location, school size, share of fully certified teachers at school, teacher absenteeism, shortage of math teachers, private vs. public school management, share of government funding at school, country's GDP per capita, school autonomy, GDP-autonomy interaction, imputation dummies, country fixed effects and year fixed effects. Least squares regression at country-by-wave level, including country and year fixed effects. Sample: country-level observations in six PISA waves 2000-2015. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A13: Two-stage estimation: Panel model estimated at country-by-wave level, specification with average effects

	Math	Science	Reading
	(1)	(2)	(3)
Standardized testing with external comparison (SCOMP)	30.756***	24.357***	27.046***
	(7.236)	*(7.472)*	*(6.621)*
Standardized testing for internal comparison (SINT)	-4.765	0.402	-1.317
	(16.974)	*(17.391)*	*(14.641)*
Internal reporting (IRPT)	5.404	15.201	-11.428
	(15.291)	*(17.128)*	*(17.067)*
Teacher monitoring (TMON)	-36.953**	-31.555*	-26.154
	(18.188)	*(16.476)*	*(17.414)*
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Country observations	59	59	59
Country-by-wave observations	303	303	303

Notes: Dependent variable: country-level aggregation of the residuals of a first-stage student-level regression that regresses the PISA test score in the subject indicated in the header on student gender, age, parental occupation, parental education, books at home, immigration status, language spoken at home, school location, school size, share of fully certified teachers at school, teacher absenteeism, shortage of math teachers, private vs. public school management, share of government funding at school, country's GDP per capita, school autonomy, GDP-autonomy interaction, imputation dummies, country fixed effects and year fixed effects. Least squares regression at country-by-wave level, including country and year fixed effects. Sample: country-level observations in six PISA waves 2000-2015. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.