RISE

RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

WORKING PAPER SERIES

October 2018

Testing

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

RISE-WP-18/025

Funded by:

Testing[*]

Annika B. Bergbauer, Eric A. Hanushek, and Ludger Woessmann[†]

Abstract

School systems regularly use student assessments for accountability purposes. But, as highlighted by our conceptual model, different configurations of assessment usage generate performance-conducive incentives of different strengths for different stakeholders in different school environments. We build a dataset of over 2 million students in 59 countries observed over 6 waves in the international PISA student achievement test 2000-2015. Our empirical model exploits the country panel dimension to investigate reforms in assessment systems over time, where identification comes from taking out country and year fixed effects along with a rich set of student, school, and country measures. We find that the expansion of standardized external comparisons, both school-based and student-based, is associated with improvements in student achievement. The effect of school-based comparison is stronger in countries with initially low performance. Similarly, standardized monitoring without external comparison has a positive effect in initially poorly performing countries. By contrast, the introduction of solely internal testing and internal teacher monitoring including inspectorates does not affect student achievement. Our findings point out the pitfalls of overly broad generalizations from specific country testing systems.

Keywords: student assessment, testing, accountability, student achievement, international, PISA

JEL classification: I28, H52, L15, D82, P51

September 14, 2018

---

[*] We gratefully acknowledge comments from participants at the Spring Meeting of Young Economists, the BGPE Research Workshop, and the center seminar of the ifo Center for the Economics of Education. This work was supported by the Smith Richardson Foundation. Woessmann acknowledges support by Research on Improving Systems of Education (RISE) which is funded by UK Aid and Australian Aid.

[†] Bergbauer: ifo Institute at the University of Munich, bergbauer@ifo.de; Hanushek: Hoover Institution, Stanford University, CESifo, IZA, and NBER, hanushek@stanford.edu; Woessmann: University of Munich, ifo Institute, CESifo, and IZA, woessmann@ifo.de.

1. Introduction

Use of student assessments for accountability purposes has grown rapidly around the world. While some have argued that this trend has been damaging to schooling (Hout and Elliott (2011); Andrews and coauthors (2014)), others have argued that even more student assessment is called for. In fact, the World Bank (2018), in evaluating the need for improved human capital development around the world, explicitly calls for expansion of student evaluations and concludes that “[t]here is too little measurement of learning, not too much” (p. 17). However, both critics and proponents of international and national testing often fail to differentiate among alternative forms and uses of testing, leading to a confused debate. For example, in the United States consideration of testing is mostly restricted to such accountability systems as exemplified by No Child Left Behind (NCLB). In reality, there are many other dimensions of student assessments. Testing students in order to provide external comparisons is very different from evaluating teachers on the basis of student performance or from making selections of which students should continue on to university. And standardized tests normed to a large population are very different than teacher-generated tests used to assess the pace of classroom learning. Understanding the overall impact of student testing requires careful consideration of how the assessments are used and what incentives they create.

This paper exploits international comparisons to estimate the effects of different types and dimensions of student assessments on overall levels of student achievement. It places the evaluation of student assessments into the general analysis of how information is translated into incentives for the actors and into behavioral results. The conceptual framework of a principal-agent model leads us to consider three dimensions of student assessments: varying strengths of incentives, different stakeholders on whom the incentives are focused, and dependence on particular school environments.

While there have been previous evaluations of the impact of accountability systems, largely within the United States (Figlio and Loeb (2011)), it is unclear how to generalize from these. These policies operate within a specific institutional environment of national school systems; as such, the evaluations necessarily neglect overall features that are common across a nation. Moreover, testing policies are often set at the national level, making it difficult to construct an adequate comparison group for evaluation of policy outcomes. By moving to international comparisons, it is possible to consider how overall institutional structures interact with the specifics of student assessments and school accountability systems. This cross-country approach allows us to investigate which aspects of student assessment systems generalize to larger settings and which do not. Of course, this advantage comes at a cost, because identifying the impact of various schooling policies across nations offers its own challenges.

Our empirical analysis uses data from the Programme for International Student Assessment (PISA) to construct a panel of country observations of student performance. Specifically, we pool the micro data of over two million students across 59 countries participating in six PISA waves between 2000 and 2015. PISA includes not only assessments of student outcomes, but also rich background information on both students and schooling institutions in the different countries. We derive a series of measures of different types of student assessments from these survey data and from other international data sources.

Because this is a period of rapid change in student assessment policies across countries, we can link policies to outcomes in fixed-effects panel models. Our identification relies on changes in student assessment regimes within countries over time. While using the individual student data for estimation at the micro level, we measure our treatment variables as country aggregates at each point in time to avoid bias from within-country selection of students into schools. Conditioning on country and year fixed effects allows us to account for unobserved time-invariant country characteristics as well as common time-specific shocks.[1]

Our analysis shows that some uses of student testing affect student learning, while others have no discernible impact. We create four categories of test usage that correspond to differing incentive patterns in our conceptual model. On the one hand, we find that expanded standardized testing that provides external comparisons is associated with increased performance on the international tests. This is true for both school-based and student-based forms of external comparisons and in math, science, and reading. On the other hand, internal testing that simply informs or monitors progress without external comparability and internal teacher monitoring including inspectorates have little discernible effect on overall performance. While not being related to student achievement on average, introducing standardized monitoring without external comparison has a positive effect in initially poorly performing countries, but not in initially

[1] Our analysis expands on the growing literature studying determinants of student achievement in a cross-country setting (Hanushek and Woessmann (2011); Woessmann (2016)). Methodologically, our approach builds on the analysis of school autonomy in Hanushek, Link, and Woessmann (2013).

highly performing countries. Similarly, the impact of school-based external comparisons differs across schooling systems with larger impacts being seen in poorer performing systems.

In a placebo test with leads of the assessment variables, we show that new usages of assessments are not systematically linked to prior outcome conditions. We also show that the results are not affected by any individual country; that they are robust to subsets of countries, to a long-difference specification, and to controlling for test exclusion rates; and that changes in PISA testing procedures are not affecting the results.

Sorting out the implications of alternative testing regimes is increasingly important from a policy perspective. As testing technologies change, it is becoming easier to expand assessments. Further, the linkage of accountability systems with ideas of reform and improvement has led to worldwide increases in testing for accountability purposes. At the same time, backlash to various applications of testing and monitoring of schools has placed assessment policies into open and often contentious public debate. Our analysis can inform this debate in a scientific way.

The next section develops a conceptual framework that highlights the achievement effects of different dimensions of student assessments. Section 3 introduces the data and Section 4 the empirical model. Section 5 presents our results including analyses of heterogeneous effects. Section 6 reports a placebo test and Section 7 a series of robustness analyses. Section 8 concludes.

2. An Incentive Framework of Different Dimensions of Assessments

To frame our thinking about potential effects of different uses and displays of student assessments, we develop a simple conceptual framework that focuses on how assessment regimes create incentives for teachers and students to focus on raising student achievement. We start with a basic principal-agent framework, discuss the technology of student assessment, and then analyze three dimensions of student assessments: different strengths of incentives, different addressees of incentives, and dependence on school environments.

2.1 Conceptual Framework: Principal-Agent Relationships

Our underlying framework is one in which parents are trying to ensure the welfare of their children. We take a very simplified view that highlights parental choices over the schooling investments of their children. Of course, parental choices and the activities of parents and children are much more complicated than the simplified views we express here, but we want to emphasize strategic choices about child investment and how these are affected by student assessment systems.

Abstracting from any other factors that enter parental considerations, let us assume that parents p aim to maximize the following value function V that balances long-run outcomes and short-run happiness of their child (student) s:

$$\text{Parents: max } V_p = f_p[A_s, R_s, E_s] \tag{1}$$

Specifically, parents care about their child's achievement A of knowledge and skills, which we believe directly affects their long-run economic outcomes (Card (1999); Hanushek et al. (2015)). The happiness of the child in the short run depends positively on any short-term reward R for learning and negatively on the effort E that the child has to put in.

Parents, however, cannot directly choose the elements of this value function but must work indirectly to achieve their ends. In particular, they may offer short-term rewards for learning R to their child and try as best as possible to observe and control child effort E. Similarly, achievement A is only partially controlled by parents but as a general rule relies heavily upon purchasing the services of schools. This is natural because of economies of scale in producing knowledge, of the limited ability of parents to provide the full array of school services, and of the benefits of specialization.

The production of achievement A can thus be described through an educational production function that we write as

$$A_s = A_s(I, E_t, E_s) \tag{2}$$

For simplicity, child achievement A is a function of inputs I into the teaching process (including parental inputs, school inputs, and student ability), teacher effort E_t, and student effort E_s.

As effort levels of teachers and children cannot be perfectly observed or controlled by parents, this setup gives rise to a tree of standard principal-agent relationships (Laffont and Martimort (2002)).[2] In particular, parents act as principals that contract the teaching of their children to schools and teachers as agents. In the process of classroom instruction, teachers also act as principals themselves who cannot fully observe the learning effort of their students as

[2] The RISE conceptual framework (Pritchett (2015)) similarly uses a series of connected principal-agent relationships to analyze the performance of schooling systems in producing learning.

agents. Teaching in the classroom and studying at your desk may be viewed as classical examples of asymmetric information where the respective principal cannot fully monitor the behavior of the respective agent. Parents, teachers, and students each have specific objective functions that combine with the asymmetric information of the actors. Therefore, one cannot simply assume that the actions of children and teachers will lead to the optimal result for parents.

Let us assume that teachers maximize the following value function:

$$\text{Teachers: max } V_t = f_t\left[A_s\left(I, \underbrace{E_t}_{(+)}, E_s\right), R_t, \underbrace{E_t}_{(-)}\right] \tag{3a}$$

Teachers derive value from their students' achievement *A*, which is a positive function of their own effort E_t, as well as from other short-term rewards R_t. At the same time, their effort at teaching E_t is costly to them, directly entering their value function negatively.

The value function of students is very similar, except that the focus is their own rewards and effort:

$$\text{Students: max } V_s = f_s\left[A_s\left(I, E_t, \underbrace{E_s}_{(+)}\right), R_s, \underbrace{E_s}_{(-)}\right] \tag{4a}$$

Note that the students' value function has the same arguments as the parents' value function, only that, for several reasons, children and parents may put different weights to the short-run and long-run costs and rewards. For example, children may be less aware of the importance of achievement *A* for their long-run well-being than parents. Furthermore, children may be less willing or able to solve the dynamic optimization problem, leading to behavioral biases that prevent them from pursuing their own long-run well-being (Lavecchia, Liu, and Oreopoulos (2016)).

If parents had full information about the effort levels of teachers and students, they could effectively contract with each to maximize their own value function. However, because of the incomplete monitoring of effort and the differing value functions, the ensuing principal-agent problems may lead to suboptimal effort levels by teachers and by students.

2.2 The Technology of Student Assessment

Solving these problems can be accomplished if there is sufficient information about the effort levels of agents, but actually obtaining and monitoring effort levels is generally costly. The more common solution is to begin with outside assessments of the outcomes of interest *A*. Nonetheless, there are a number of complications with the usage of information about achievement, and these are the subject of many current policy deliberations and controversies. Because achievement is a function of both teacher and student effort, it is not easily possible to infer the effort of either with just information on achievement levels.

At a basic level, student assessments provide information on student outcomes. They use a testing technology τ to transform actual outcomes *A* into observed outcomes *O*:

$$O_s = \tau(A_s) \tag{5}$$

From this information on student outcomes, one can try to infer effort levels. This would then allow creating incentives that align agents' behavior more closely with the principals' objective function.

Historically, a variety of testing regimes have been developed that are designed to provide information about achievement levels. For our purposes, however, we have to consider how any of these assessments can be used to solve the underlying principal-agent problems. In reality, the emerging policy choices frequently assume specific features of the production function in arriving at solutions to these problems.

In a general way, we can think of providing rewards *R* to both teachers and students based on the outcome levels *O* observed by the student assessments:[3]

$$\text{Teachers: max } V_t = f_t\left[A_s\left(I, \underbrace{E_t}_{(+)}, E_s\right), R_t\left(\underbrace{O_s}_{(+)}\right), \underbrace{E_t}_{(-)}\right] \tag{3b}$$

$$\text{Students: max } V_s = f_s\left[A_s\left(I, E_t, \underbrace{E_s}_{(+)}\right), R_s\left(\underbrace{O_s}_{(+)}\right), \underbrace{E_s}_{(-)}\right] \tag{4b}$$

This effectively alters their value functions and introduces incentives for their behavior.

That is the focus of this paper: By creating outcome information, student assessments provide a mechanism for developing better incentives to elicit increased effort by teachers and students, thereby ultimately raising student achievement levels to better approximate the desires of the parents. We think of the potential rewards *R* for observed outcomes *O* in a very general

[3] Throughout, we have taken the simplifying assumption that there is a single teacher whose behavior is affected by incentive schemes. In reality, the incentive schemes almost certainly have an impact not only on the effort choices of existing teachers, but also on who becomes a teacher and the long-run supply of teachers.

way, including implicit and explicit rewards, material and non-material rewards, and ranging from simple observability of outcomes over parental gratification for students to consequences for teachers at school.

There are two issues that we have to consider. First, how do we separate the joint effort levels of teachers and students in order to provide the right incentives? Second, how do we deal with imperfect technologies that do not provide complete information on A? For expositional purposes, let us start with the assumption that actual achievement is perfectly observed, i.e., $O_s = A_s$. We will come back to the more realistic assumption that O_s is only an imperfect measure of actual achievement below.

The first issue is a classical identification problem. We want to know when we can infer effort levels of teachers and students from information on outcomes. If student efforts were constant over time, we could directly relate changes in achievement in a given classroom to the teacher and from that infer their effort levels. Alternatively, if we thought teacher effort was constant, we could attribute different performance of students to their own effort. The first is roughly the idea behind value-added modelling (Koedel, Mihaly, and Rockoff (2015); Chetty, Friedman, and Rockoff (2017)). The second is closer to providing consequential exit exams for student achievement (Bishop (1997)). Of course, in neither case is it realistic to assume constant effort by the other actor, but the policy choices implicitly assume that one form of effort is much more important than the other. These issues will be discussed more completely in Section 2.4 below.

The second issue recognizes the fact that no assessment technology τ today provides complete measurement of the relevant achievement for long-run well-being. Prior discussions of accountability systems have considered various dimensions of this problem (Figlio and Loeb (2011)). Perhaps the best-known conceptual discussion is the classic Holmstrom and Milgrom (1991) paper that considers how imperfect measurement of outcomes distorts incentives (see also Dixit (2002)). In particular, if there are multiple objectives and only a subset is measured, effort could be distorted to the observed outcomes to the detriment of unobserved outcomes. But there is also more general discussion of such topics as teaching to the test (Koretz (2017)),[4] gaming of

[4] There are two aspects of teaching to the test. On the one hand, teaching may unduly focus on the form and character of the test itself, which is not in the interest of parents. Creative and flexible designs of tests are required to prevent such activity. On the other hand, if the tests accurately sample from the domains of achievement that parents

tests (e.g., nutritious feeding on testing days, see Figlio and Winicki (2005)), and cheating (Jacob and Levitt (2003)). Each of these topics includes an element of testing technology and the accuracy of observed measures and is the subject of a much larger literature. Here, we simply want to note that the impact of different incentives will be conditioned by elements of the testing technology. The ultimate effects on achievement thus become an empirical question.

2.3 Assessment Dimension 1: Different Strengths of Incentives

Testing is a ubiquitous component of schooling, but not all tests have the same use or impact in helping to solve the underlying principal-agent problems. By far the most common type of testing is teacher-developed tests that are used both to guide instruction and to provide feedback to students and parents. The key feature of teacher-developed tests is that it is generally difficult if not impossible to compare results across teachers. Therefore, while these tests may be useful in providing incentives to students and related information to parents (O_s enters positively in R_s in equation (4b)), they do not solve the principal-agent problem between parents and teachers (O_s effectively does not enter R_t in equation (3b)). One would not expect the results of these tests to affect teacher effort levels. There is a blurry line between teacher-developed tests and periodic content testing that generally goes under the heading of formative assessments which may also be provided by external producers. In both cases, the information provided by the tests is just used internally by the teacher without parents being able to compare outcomes externally.

At the other end of the continuum of testing are standardized tests that have been normed to relevant population performance. These tests allow for direct comparisons of student outcomes in different circumstances and thus suggest the possibility of using them to provide incentives to teachers in addition to students.

Of course, the strength of any incentives relating to these various tests will depend upon how they enter into rewards for teachers and students in equations (3b) and (4b). On the one hand, results of student assessments may just provide information to some or all actors in the system.[5] On the other hand, performance on any test may also be linked directly to consequences

---

desire, focusing teaching towards the contents of the test is in fact part of the mechanism of aligning teaching with the parental value function.

[5] For example, school rankings may be published to the general public (see Koning and van der Wiel (2012), Burgess, Wilson, and Worth (2013), and Nunes, Reis, and Seabra (2015) for evidence from the Netherlands, Wales, and Portugal, respectively), and school report cards may provide information to local communities (see Andrab, Das, and Khwaja (2017) for evidence from a sample of villages in Pakistan).

– rewards and punishments to students (including retention and promotion) and teachers.[6] As a general principle, we would naturally expect attaching consequences to results to produce stronger incentives and larger behavioral changes.

2.4 Assessment Dimension 2: Different Addressees of Incentives

Previously, we described the overall problem as a tree of principal-agent relationships. We did that because the problem applies to the behavior and effort levels of a wide variety of actors in the schooling system. As a canonical description of the tree, we are concerned with the parent-child problem, the parent-teacher problem, and the teacher-child problem. Adding another layer to the system, parents often look beyond the individual teacher to school administrators at different levels, including the nation, the region, the school district, and the school. This suggests that there are parent-administrator problems, administrator-administrator problems, and administrator-teacher problems that are relevant to incentive design questions.

The optimal design of incentives generally calls for rewarding the results of behavior directly under the control of the actor and not rewarding results from other sources. The problem as sketched out above is that most testing includes the results of actions of multiple parties. While incentives found in various schooling circumstances are often implicitly discussed and instituted with one of these principal-agent problems in mind, it is easy to see how incentives may differ across the various actors and how solving one principal-agent problem may leave others untouched.

For example, centralized exit exams that have consequences for further schooling of students may have strong incentives for student effort (equation (4b)), but limited impact on teacher effort (equation (3b)).[7] On the other hand, testing that is directly linked to consequences for schools such as the NCLB legislation in the US may have limited relevance for students and

[6] Apart from systemic consequences, different parents will attach different consequences to their children for the same performance, likely contributing to achievement differences across socioeconomic groups.

[7] By affecting chances to enter specific institutions and fields of higher education as well as the hiring decisions of potential employers, central exit exams usually have real consequences for students; see Bishop (1997), Woessmann (2003), Woessmann et al. (2009), Jürges, Schneider, and Büchel (2005), Lüdemann (2011), and Schwerdt and Woessmann (2017) for further analysis of the effects of central exit exams.

their efforts.[8] Similarly, school inspectorates and inspections of teacher lessons may be more relevant for school and teacher effort than for student effort.

There is much public discussion of the implications of high-stakes testing, but this frequently is not accurately aligned with incentives for the different actors in the system. For example, differential rewards to teachers based upon test-score growth are high stakes for the teachers, but not for the students. At the same time, tests that have no consequences for any of the actors may be inconsequential for overall performance because nobody may take them seriously.

2.5 Assessment Dimension 3: Dependence on School Environments

The prior conceptual discussion is framed in terms of a series of individual two-way interactions. Understanding the implications of various testing schemes and their usage necessarily involves looking at performance across schools and, in our case, across countries. When we think in these larger terms, it is difficult to believe that behavior is uniform across systems even when confronted with the same incentive structure.[9]

For example, if we look at a set of high-performing schools, we may think that they know how to react to achievement signals and different rewards. Therefore, we may expect that any type of incentive structure created by student assessments has a stronger impact on them than on an otherwise comparable set of low-performing schools. But at the same time, we might think that the results are just the opposite: Low-performing schools have more room for improvement and may be in greater need to have their incentives focused on student outcomes. High-performing schools, by contrast, may have the capabilities and be subject to overall political and schooling institutions that already better reflect the desires of parents.

[8] For analyses of the effects of NCLB and predecessor reforms, see Hanushek and Raymond (2005), Jacob (2005), Dee and Jacob (2011), Reback, Rockoff, and Schwartz (2014), and Deming et al. (2016); see Figlio and Loeb (2011) for a survey.

[9] Another dimension of heterogeneity may be across parents within a system, in that different parents have different value functions (including different discount rates that affect the relative value of short-term and long-term outcomes) and/or different capacity to drive favorable results. Such differences may lie behind movements such as parents opting out of state-wide testing in the US, in that some parents may feel that the measured output does not provide much information about the type of achievement that they care about.

3. International Panel Data

For our analysis, we combine the student micro data of all available waves of the PISA international achievement test with measures of different types of student assessment policies over a period of 15 years. We describe each of the two components in turn.

3.1 Six Waves of PISA Student Achievement Tests

In 2000, the Organisation for Economic Co-operation and Development (OECD) conducted the first wave of the international student achievement test called Programme for International Student Assessment (PISA). Since then, PISA has tested the math, science, and reading achievement of representative samples of 15-year-old students in all OECD countries and in an increasing number of non-OECD countries on a three-year cycle (OECD (2016)).[10] PISA makes a concerted effort to ensure random sampling of schools and students and to monitor testing conditions in participating countries. Data are not reported for countries that do not meet the standards.[11] PISA does not follow individual students over time. But the repeated testing of representative samples of students creates a panel structure of countries observed every three years.

In our analyses, we consider all countries that have participated in at least three of the six PISA waves between 2000 and 2015.[12] This yields a sample of 59 countries observed in 303 country-by-wave observations. We perform our analysis at the individual student level, encompassing a total sample of 2,187,415 students in reading and slightly less in math and science. The sample, listed in Table 1, includes 35 OECD and 24 non-OECD countries that encompass a wide range of levels of economic development and student achievement.

PISA uses a broad set of tasks of varying difficulty to create a comprehensive indicator of the continuum of students' competencies in each of the three subjects. Overall testing lasts for up

[10] The target population contains all 15-year-old students irrespective of the educational institution or grade that they attend. Most countries employ a two-stage sampling design, first drawing a random sample of schools in which 15-year-old students are enrolled (with sampling probabilities proportional to schools' number of 15-year-old students) and second randomly sampling 35 students of the 15-year-old students in each school.

[11] In particular, due to deviations from the protocol, the data exclude the Netherlands in 2000, the United Kingdom in 2003, the United States in the reading test 2006, and Argentina, Kazakhstan, and Malaysia in 2015.

[12] We include the tests conducted in 2002 and 2010 in which a number of previously non-participating countries administered the 2000 and 2009 test, respectively. We exclude any country-by-wave observation for which the whole information of a background questionnaire is missing. This applies to France from 2003-2009 (missing school questionnaire) and Albania in 2015 (missing student questionnaire). Due to its small size, Liechtenstein was also dropped.

to two hours. Using item response theory, achievement in each domain is mapped on a scale with a mean of 500 test-score points and a standard deviation of 100 test-score points for OECD-country students in the 2000 wave. The test scales are then psychometrically linked over time.[13] Until 2012, PISA employed paper and pencil tests. In 2015, the testing mode was changed to computer-based testing, a topic we will come back to in our robustness analysis below.

Figure 1 depicts the evolution of math achievement of each country over the 15-year period. While average achievement across all countries was quite stable between 2000 and 2015, achievement has moved significantly up in some countries and significantly down in others. In 14 countries, achievement improved by at least 20 percent of a standard deviation compared to their initial achievement (in decreasing order, Peru, Qatar, Brazil, Luxembourg, Chile, Portugal, Israel, Poland, Italy, Mexico, Indonesia, Colombia, Latvia, and Germany). On the other hand, achievement decreased by at least 20 percent of a standard deviation in eleven countries (United States, Korea, Slovak Republic, Japan, France, Netherlands, Finland, Iceland, United Kingdom, Australia, and New Zealand).

In student and school background questionnaires, PISA provides a rich array of background information on the participating students and schools. Students are asked to provide information on their personal characteristics and family background, and school principals provide information on the schools' resources and institutional setting. While some questionnaire items, such as student gender and age, remain the same across the six PISA assessment cycles, other information is not available in or directly comparable across all waves. We therefore select a set of core variables of student characteristics, family backgrounds, and school environments that are available in each of the six waves and merge them with the test score data into one dataset comprising all PISA waves.

Our vector of control variables allows us to condition on a rich set of observed characteristics of students, schools, and countries. The student-level controls include student gender, age, first- and second-generation immigration status, language spoken at home, parental education (measured in six categories), parental occupation (four categories), and books at home (four categories). The school-level controls include school size (number of students), community location (five categories), share of fully certified teachers, principals' assessments of the extent

[13] The math (science) test was re-scaled in 2003 (2006), any effect of which should be captured by the year fixed effects included in our analysis.

to which learning in their school is hindered by teacher absenteeism (four categories), shortage of math teachers, private operation, and share of government funding. At the country level, we include GDP per capita and, considering the results in Hanushek, Link, and Woessmann (2013), the share of schools with academic-content autonomy and its interaction with initial GDP per capita. To avoid sample selection bias from non-response in the survey data, we impute missing values in the student and school background variables by using the respective country-by-wave mean.[14] To ensure that imputed data are not driving our results, all our regressions include a set of dummy variables – one for each variable with missing data – that are set to one for imputed values and zero otherwise.

3.2 Categories of Assessment Usage

From the PISA school background questionnaires and other sources, we derive a series of measures of different categories of the use of student assessments over the period 2000-2015. The central insight of our conceptual modeling is that different kinds of tests and different uses of these tests create varied incentives, and these are likely to show up in different achievement outcomes. To be useful for the analysis, we need information on different testing practices that is consistent both across countries and across time. There are several sources that provide relevant data while meeting these stringent requirements. Obviously, survey designers and organizations supplying information about assessments have not had our conceptual model in mind when initiating their work. Thus, we have questions that cover a wide range of narrow aspects of testing, and for our empirical analysis it is useful to collapse several individual items into more general categories.

Here we summarize the categories of testing that we construct, while the details of questions and sources can be found in the Data Appendix. From a combination of the surveys for principals that accompany the PISA assessments, of the regular publications and data collection of other parts of the OECD, and from data compiled under the auspices of the European Commission, we have 13 separate indicators of the use and purpose of testing, each measured at the country-by-wave level.[15] We combine these into four separate categories that represent quite different aspects of testing in the schools. They differ by the degree of standardization of the

[14] The share of missing values is generally very low, see Appendix Table A1.

[15] Appendix Table A2 provides an overview of the different underlying assessment indicators. Appendix Table A3 indicates the number of country observations by wave for each indicator.

assessment data and the specific actors – administrators, teachers, and students – most affected. We construct these aggregate measures because of overlap and correlations among the individual questions and because of potential measurement error in different individual questions of similar content areas.

Standardized External Comparisons. The first category relates to four separate data sources that identify use of standardized assessments constructed outside of the schools and used explicitly to allow comparisons of student outcomes across schools and students. This category includes the proportion of schools where (according to the principals of schools participating in PISA) performance of 15-year-olds is regularly compared through external examinations to students across the district or the nation (which we term "school-based external comparisons"). It also includes indicators of whether central examinations affect student placement at lower secondary level (two sources) and whether central exit exams determine student outcomes at the end of secondary school (which, together, we term "student-based external comparisons").[16] This overall category of exams has strong incentives through the rewards to students but also affects rewards to administrators and teachers by making external information available to parents and policy makers. While not fully explicit from the surveys, the items in this category are roughly ones where consequential outcomes are related to student scores, making for stronger incentives.[17]

Standardized Monitoring. In other instances, standardized assessments are used to monitor the performance of students, teachers, or schools without necessarily involving any external comparison or public recording. Three questions in the PISA survey provide information on the prevalence of different aspects of this usage: standardized testing in the tested grade, monitoring of teacher practices by assessments, and tracking of achievement data by an administrative authority. While not always clear, these test usages appear closer to report card systems without external comparison and imply less powerful incentives than in the previous category of external comparisons.

[16] As discussed in the Data Appendix, data on assessments used for student placement are available for only a subset of countries, largely the OECD countries.

[17] In prior work on U.S. accountability, accountability that had consequential impacts on schools were much more closely related to student performance than accountability that was confined to report card information (Hanushek and Raymond (2005)).

Internal Testing. This category would generally cover testing – either standardized or unstandardized – that is used for general pedagogical management including informing parents of student progress, public posting of outcomes, and tracking school outcomes across cohorts. The data come from three separate PISA questions and, in our conceptual framework, represent low-level incentives because of the lack of comparability across student groups.

Internal Teacher Monitoring. In addition to the general use of internal assessments covered in the previous category, this final category covers internal assessments that are directly focused on teachers. Specifically, this category, again derived directly from the principal surveys in PISA, combines schools' use of assessments to judge teacher effectiveness and the monitoring of teacher practice by principals and by external inspectorates. These assessments would have minimal incentives for students and uncertain but generally small impacts on teacher rewards because of the lack of comparability across settings.

Aggregation of Separate Indicators. We combine the original 13 separate indicators of assessment practices into four main categories as the simple average of the observed indicators in each category.[18] Constructing the aggregate categories serves several purposes. In various instances the survey items are measuring very similar concepts, so that the aggregation acts to reduce individual measurement error and to limit multicollinearity at the country level (which is key in our identification strategy). For example, using our aggregate country-by-wave data, some individual indicators are correlated above 0.5 even after extracting country and year fixed effects. Additionally, the aggregation permits including the added information from some more specialized OECD and EU sources while not forcing elimination of other countries outside these boundaries.

Some Descriptive Statistics. Table 2 provides descriptive statistics both for the individual indicators of student assessment and for the four combined assessment categories. The measures derived from the PISA background questionnaires are shares bounded between 0 and 1, whereas the other assessments measures are dummy variables.[19] As is evident, some assessment practices

[18] The variables in each category are calculated as proportionate usage in terms of the specific indicators for each country and wave. Note also that indicator data entirely missing for specific PISA waves are imputed by country-specific linear interpolation of assessment usages, a procedure that retains the entire country-by-wave information but that does not influence the estimated impact of the test category because of the inclusion of imputation dummies in the panel estimates (see Data Appendix for details).

[19] In federal countries, the dummy variables capture whether the majority of the student population in a country is subject to the respective assessment policy.

are more common than others. For example, 89 percent of schools in our country-by-wave observations use some form of assessment to inform parents, but only 29 percent have national standardized exams in lower secondary school. Table 1 provides country-by-country statistics of the initial and final value of three selected measures of standardized external comparison. Of particular relevance, there is a tendency for increased prevalence of the measures of standardized external comparison over time.

The important aspect of our test usage data is the amount of variation over time within individual countries. To give some understanding of the patterns of change, Figure 2 provides a depiction of the evolution of using standardized assessments for school-based external comparison from 2000 to 2015 for each country. The increasing use of such external assessments in many countries is quite evident. For example, in five countries, the share of schools that are externally compared with student assessments increased by more than 50 percentage points (Luxembourg, Denmark, Italy, Portugal, and Poland), and in another 18 countries, the share increased by more than 20 percentage points. In three countries, by contrast, the share decreased by more than 20 percentage points (Tunisia, Costa Rica, and Croatia).[20]

4. Empirical Model

Identifying the impacts of testing in a cross-country analysis is of course challenging. Assessments are not exogenously distributed across schools and countries. At the student level, an obvious potential source of bias stems from the selection of otherwise high-performing students into schools that have specific assessment practices. At the country level, there may also be reverse causality if poorly performing countries introduce assessment systems in order to improve their students' achievement. Ultimately, any omitted variable that is associated both with the existence of student assessments and with student achievement levels will lead to bias in conventional estimation. In the cross-country setting, for example, unobserved country-level

[20] It is beyond the scope of this paper to provide detailed anecdotal narratives of specific policy reforms that underlie the changes in student assessment measures documented by the PISA school background questionnaires. However, on a number of occasions, it is straightforward to link major policy reforms directly to the overall pattern of expanded accountability measures. For example, the strong increase in school-based assessments used for external comparison in Italy in 2009, clearly visible in Figure 2, coincides with the introduction of the *Invalsi* national test (https://it.wikipedia.org/wiki/Test_INVALSI). Similarly, the increased external assessment in Denmark in 2006 reflects the 2006 *Folkeskole* Act which introduced a stronger focus on evaluation, assessment, and accountability including national tests (Shewbridge et al. (2011)). And the strong increase in external assessments in Luxembourg shows the introduction of standardized national assessments that monitor student outcomes in French, German, and mathematics (Shewbridge et al. (2012)).

factors such as culture, the general valuation of educational achievement, or other government institutions may introduce omitted-variable bias.

In our empirical model, we address leading concerns of bias in cross-country estimation by formulating a fixed-effects panel model of the following form:

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + \beta X_{ct} + \mu_c + \mu_t + \varepsilon_{ict} \quad (6)$$

In this empirical version of an education production function, achievement *A* of student *i* in country *c* at time *t* is expressed as a linearly additive function of vectors of input factors at the level of students *I*, schools *S*, and countries *C*, as well as the measures of student assessment *X*. The parameters μ_c and μ_t are country and year fixed effects, respectively, and ε_{ict} is an individual-level error term. Because of potential multicollinearity between the four categories of student assessment, we start by estimating separate models for each assessment category and subsequently report models that consider all four categories simultaneously.

Our fixed-effects panel model identifies the effect of assessment practices on student achievement only from country-level within-country variation over time. First, note that the treatment variable, X_{ct}, is aggregated to the country-by-wave level. By measuring the average extent of student assessments in a country at any given point in time, this specification avoids bias from within-country selection of students into schools that use student assessments. This does not, however, address concerns of bias from unobserved features at the country level.

Therefore, we secondly include country fixed effects μ_c, which effectively address any potential omitted variable bias that arises from unobserved time-invariant country characteristics that may be correlated with both assessments and achievement. The specification exploits the fact that different countries have reformed their assessment systems at different points in time. Being identified from country-level variation over time, our parameter of interest β will not be affected by systematic, time-invariant differences across countries. This implies that countries that do not change their assessment practices over the observation period will not enter into the estimation of β.

To avoid bias from the fact that the global trend towards more assessment may coincide with other trends that are relevant for student achievement, the model also includes time fixed effects μ_t. These also capture any common shocks that affect testing in a specific PISA wave, as well as any changes in the testing instruments in a given wave.

The key identifying assumption of our model is the standard assumption of fixed-effects panel models. Conditional on the rich set of control variables at the student, school, and country level included in our model, in the absence of reform the change in student achievement in countries that have introduced or extended assessment practices would have been similar to the change in student achievement in countries that did not reform at the given point in time. We will come back to a discussion of potential violations of this identifying assumption and thus potential remaining bias in the panel estimates in our further analyses below.

5. Results

The conceptual model identified three primary dimensions of the outcome implications of alternative assessment usage: strength of incentives, addressee of the primary incentives, and interactions with the overall environment. Here we sequentially consider the estimated impact of each of these dimensions.

5.1 Strength of Incentives across Usage Categories

We start our discussion of results with the average effects of the different categories of student assessment in our country sample. Table 3 presents the results for the combined measures of the four assessment categories, first entered separately (columns 1-4) and then jointly (columns 5-7). All models are estimated as panel models with country and year fixed effects, conditioning on the rich set of control variables at the student, school, and country level indicated above.[21] Regressions are weighted by students' sampling probabilities within countries, giving equal weight to each country-by-wave cell across countries and waves. Standard errors are clustered at the country level throughout.

Overall, the basic impact results displayed in Table 3 suggest that different forms and dimensions of student assessments have very different effects on student achievement. Among the four assessment categories, only standardized testing that is used for external comparisons has a strong and statistically significant positive effect on student outcomes. The coefficients on

[21] Appendix Table A1 shows the coefficients on all control variables for the specification of the first column in Table 5. Note that our results confirm the finding of Hanushek, Link, and Woessmann (2013) that the effect of school autonomy on student achievement is negative in developing countries but positive in developed countries in this extended setting. With six rather than four PISA waves and with 303 rather than 155 country-by-wave observations, we show that the previous results about autonomy are robust to the consideration of the effects of student assessment reforms.

standardized monitoring and internal testing are insignificant and close to zero, whereas there is quite a sizeable negative coefficient on internal teacher monitoring. These different impacts are consistent with the predictions on differing strengths of incentives from the conceptual discussion.

The point estimate for standardized external comparisons suggests that a change from not used to complete standardized external comparison is related to an increase in math achievement by more than one quarter of a standard deviation. The point estimates and the statistical significance of the category impacts are very similar between the regressions that include each category of test usage individually and the regression that includes all four categories simultaneously (column 5), indicating that there is enough independent variation in the different assessment categories for estimation and that the effect of standardized external comparison does not reflect reforms in other assessment categories. In the inclusive regression, the negative coefficient on internal teacher monitoring even turns significant in math. With that nuanced exception, results for science and reading achievement are very similar to those for math (columns 6 and 7).

Individual results for each of the 13 underlying country-level indicators of student assessment going into our test usage categories are shown in Appendix Table A4, where each cell represents a separate regression.[22] Of particular interest, each of the four elements of the external comparison composite, with one exception, has a significantly positive impact on student performance in the three subjects. The exception is the use of central exit examinations, which could simply reflect that student performance measured by PISA at age 15 is not very responsive to rewards that only occur at the end of secondary school (when students are usually aged around 18 or 19). While the point estimates are positive in all three subjects, they do not reach statistical significance.[23] The estimated coefficients for the other three indicators taken separately are substantially smaller than the combined measure. As noted, this probably reflects both a reduction in measurement error for the correlated indicators and the fact that the different

[22] In the separate regressions of Appendix Table A4, the number of countries and waves included in each estimation varies and is determined by the availability of the specific assessment indicator.

[23] Consistent with the weaker evidence on central exit exams, constructing the combined measure of standardized external comparison without the central exit exam measure (i.e., based on the other three underlying indicators) yields a slightly larger coefficient estimate of 30.926 in the specification of column 5 of Table 3.

incentives are additive.[24] We return below to a consideration of separate components of external comparisons.

At the individual indicator level in Appendix Table A4, there is also some evidence of positive effects of standardized testing in the relevant grade for PISA, and some indication of impact from the use of assessment to inform parents. None of the other indicators of standardized monitoring without external comparison, of internal testing, and of internal teacher monitoring is significantly related to student achievement on average. The individual estimates suggest that the potential negative impact of the internal monitoring of teachers is driven by the two subjective components – monitoring by the principal and by external inspectorates. The aggregate categorical variable is larger than these two subcomponents, potentially again reflecting a reduction in measurement error and possible additivity.

Overall, the results indicate that, when assessing the effects of student assessments, it is important to differentiate among alternative forms and dimensions of student assessments. Across the different measures and subjects, the results for the effects of standardized external comparisons consistently suggest that introducing such assessments leads to higher achievement. By contrast, student assessments that are only used for internal testing and inspection do not seem to matter much for average student achievement. The findings suggest that clearer, more targeted information creates stronger incentives.

5.2 School-based versus Student-Based External Comparisons

The previous section highlighted the impacts of having standardized examinations that were used for external comparisons. The category of external comparisons, however, actually aggregates two quite distinct sets of incentives. One component (from the PISA questionnaires) considers the general use of standardized assessments for external comparison of schools to district or national performance. This category mainly indicates incentives to schools, potentially having its greatest effect on administrators and teachers. The second category combines three different measures of using tests to determine school and career placement decisions for students with the clear locus of incentives on the students themselves.

[24] A third possibility is that the estimation samples for the separate indicators are varied and smaller than for the combined indicator. However, we reject this explanation because estimating the combined model in column 5 of Table 3 just for the smallest sample of countries in the separate indicator models yields a virtually identical coefficient for external comparisons.

Table 4 disaggregates the standardized external comparisons into school-based and student-based external comparisons (each of which is based on standardized exams that have meaning across schools).[25] This table presents simultaneous estimates that include the other three categories. Both school and student incentives are strongly positive and statistically significant, with estimates for the school-based incentives being somewhat larger than for the individual student incentives. At the same time, none of the estimates for the remaining categories are qualitatively affected. The results suggest that focusing incentives on different actors yields different responses and leads to separate effects on outcomes.

5.3 Environmental Differences in Usage Impact

Results so far were distinguished by the first two assessment dimensions stressed by our conceptual framework, different strengths of incentives and different addressees of incentives. This section turns to the third assessment dimension, the extent to which effects vary by different school environments.

Countries enter our observation period at very different stages of educational development, and almost certainly with environments that have both different amounts of information about schools and different degrees of policy interactions among parents, administrators, and teachers. One straightforward way to parameterize these differences is to explore how incentive effects vary with a country's initial level of achievement.

We introduce an interaction term between the specific assessment measure X_{ct} and a country's average achievement level when it first participated in PISA, A_{c0}:

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + \beta_1 X_{ct} + \beta_2 (X_{ct} \times A_{c0}) + \mu_c + \mu_t + \varepsilon_{ict} \quad (7)$$

The parameter β_2 indicates whether the assessment effect varies between countries with initially low or high performance. Note that the initial performance level is a country feature that does not vary over time, so that any main effect is captured by the country fixed effects μ_c included in the model.

Table 5 presents estimates of the interacted model for the three subjects. The left three columns provide results for the aggregate category of standardized external comparisons, while

[25] The measure of student-based external comparison is the simple average of the three underlying indicators of standardized external comparison except for the one on school-based external comparison. Note that the estimates of Table 4 are based on smaller student samples from fewer countries, because data on student-based external comparison are available for few countries beyond OECD and European Union countries.

the right three columns divide the external comparisons into school-based and student-based comparisons. The initial score is centered on 400 PISA points (one standard deviation below the OECD mean). The precise patterns of estimated effects by initial achievement with confidence intervals are displayed in Figure 3 for math performance.

In broad generalities, the picture of how the overall achievement environment interacts with the incentives from different test usage can be summarized as follows. First, the impact of standardized external comparisons is stronger in lower achieving countries and goes to zero for the highest achieving countries. In particular, at an initial country level of 400 PISA points the introduction of standardized external comparison leads to an increase in student achievement of 37.3 percent of a standard deviation in math. With each 100 initial PISA points, this effect is reduced by 24.6 percent of a standard deviation. Second, standardized monitoring similarly creates significant incentives in initially low-achieving countries, with effects disappearing for higher-achieving countries (i.e., those with initial scores of roughly above 490 in all subjects). Third, the estimate of internal testing is insignificant throughout the initial-achievement support. Fourth, the estimates for internal teacher monitoring are insignificant for most of the initial-achievement distribution and turn negative only at high levels of initial achievement in math (perhaps reflecting the purely linear interaction). Fifth, when external comparisons are disaggregated into school-based and student-based components, school-based comparisons follow essentially the same heterogeneous pattern as overall standardized external comparisons but go to zero for a somewhat larger set of initially high-achieving countries. By contrast, the impact of student-based external comparisons does not vary significantly with initial achievement levels.

The disaggregated underlying individual indicators of standardized external comparison consistently show the pattern of significantly stronger effects in initially poorly performing countries (Appendix Table A5).[26] Interestingly, the introduction of central exit exams – which did not show a significant effect on average – also shows the pattern of decreasing effects with higher initial achievement, in particular in science. Similarly, all three underlying indicators of standardized monitoring also show the same pattern of significant positive effects at low levels of achievement and significantly decreasing effects with initial achievement. Thus, the positive

[26] There is no significant heterogeneity in the effect of the Eurydice measure of national testing, which is likely due to the fact that this measure is available only for 18 European countries which do not feature a similarly wide range of initial achievement levels.

effect of standardized testing in low-achieving countries appears to be quite independent of whether the standardized tests are used for external comparison or just for monitoring. This finding supports the World Bank report that focused on low achieving countries: “There is too little measurement of learning, not too much” (World Bank (2018), p. 17).[27]

In contrast to the significant interactions with initial achievement levels, we do not find evidence of consistent heterogeneities in several other environmental dimensions (not shown). In particular, the effects of the four assessment categories do not significantly interact with countries’ initial level of GDP per capita, which contrasts with the heterogeneous effects found for school autonomy in that dimension in Hanushek, Link, and Woessmann (2013). Similarly, there are no significant interactions of the assessment categories with the level of school autonomy in a country. In addition, the use of standardized external comparisons does not significantly interact with the other three categories of student assessments.

Overall, the heterogeneity analysis suggests that the use of standardized assessments is particularly fruitful in countries with relatively poor achievement, irrespective of whether they are used for external comparison or only for internal monitoring.

6. A Placebo Test with Leads of the Assessment Variables

Our fixed-effects panel model identifies the effect of assessment policies on student achievement from policy changes within countries over time. Bias from non-random within-country selection of students into schools is avoided through aggregating the assessment variables to the country level. Bias from common shocks or specific issues of particular PISA waves is taken care of through the inclusion of year fixed effects. Bias from any unobserved country features is taken care of through the inclusion of country fixed effects to the extent that the country features do not vary systematically over time. The rich set of student, school, and country background factors considered in our model takes out country-specific variation over time to the extent that it is observed in these variables.

[27] An interesting outlier in the individual-indicator analysis is the use of assessments to inform parents, which shows the opposite type of heterogeneity (significantly so in math and science): The expansion of using assessments to inform parents about their child’s progress does not have a significant effect at low levels of initial achievement, but the effect gets significantly more positive at higher levels. Among initially high-performing countries, informing parents leads to significant increases in student achievement; e.g., at an initial achievement level of 550 PISA points, there is a significantly positive effect on science achievement of 37.0 percent of a standard deviation. It seems that addressing assessments at parents is only effective in raising student achievement in environments that already show a high level of achievement, capacity, and responsiveness of schools.

A leading remaining concern of the fixed-effects model is that reforms may be endogenous, in the sense that reforming countries may already be on a different trajectory than non-reforming countries for other reasons, thus violating the usual common-trend assumption of the fixed-effects model.

Our panel setup lends itself to an informative placebo test. In particular, any given reform should *not* have a causal effect on the achievement of students in the wave *before* it is implemented. But, if the reform were endogenous, we should in fact see an association between prior achievement and subsequent reform. Therefore, including leads of the assessment measures – i.e., additional variables that indicate the assessment status in the *next* PISA wave – provides a placebo test of this.

Table 6 reports the results of this placebo test. As is evident, none of the lead variables of the four assessment categories is significantly related to student achievement (i.e., in the wave before reform implementation). At the same time, the results of the contemporaneous assessment measures are fully robust to conditioning on the lead variables: The use of standardized external comparison has a significant positive effect on the math, science, and reading achievement of students *in the year in which it is implemented*, but not in the wave in which it is not implemented yet. Moreover, the estimated coefficients for the usage categories are qualitatively similar to those in Table 3.

The fact that the leads of the assessment variables are insignificant also indicates that lagged achievement does not predict assessment reforms. In that sense, the results speak against the possibility that endogeneity of assessment reforms to how a school system is performing is a relevant concern for the interpretation of our results.

Estimating the full interacted model with all four assessment categories and their leads interacted with initial achievement is overly demanding to the data. Nevertheless, focusing just on the main results of Section 5.3, an interacted model that includes just standardized external comparison, its lead, and their interactions with initial achievement gives confirmatory results: standardized external comparison is significantly positive, its interaction with initial achievement is significantly negative, and both the lead variable and its interaction with initial achievement are statistically insignificant (not shown).

No similar test is possible for the lag of the assessment variables, as lagged assessment policies may in fact partly capture the effect of previously implemented reforms to the extent that reforms take time to generate their full effects. In a specification that includes the contemporaneous, lead, and lagged variable, both the contemporaneous and the lag of the standardized external comparison variable are statistically significant while the lead remains insignificant (not shown).

In sum, there is no evidence of the introduction of different test usage regimes in response to prior educational circumstances.

7. Robustness Analyses

Our results prove robust to a number of interesting alternative specifications. To begin with, we want to make sure that none of our results are driven by the peculiarity of any specific country. Therefore, we re-ran all our main models (the simultaneous regressions of columns 5-7 in Table 3 and columns 1-3 in Table 5) excluding one country at a time. The qualitative results are insensitive to this, with all significant coefficients remaining significant in all regressions (not shown).

To test whether results differ between developed and less developed countries, we split the sample into OECD and non-OECD countries. As the first two columns of Table 7 show, qualitative results are similar in the two subgroups of countries, although the positive effect of standardized external comparison is larger in OECD countries. Patterns of heterogeneity are less precisely identified within the two more homogeneous subgroups (Table 8). In the group of OECD countries, the significant effect of standardized external comparison does not vary significantly with initial achievement, but the demands of the fully interacted model make estimation difficult with just the 35-country sample. When we drop the insignificant interactions (column 2), the point estimate of the use of standardized scores for comparisons is significant. The heterogeneous effect of standardized monitoring is somewhat more pronounced in OECD countries. But overall, the patterns do not differ substantively between the two country groups.

Our main model is identified from changes that occur from one PISA wave to the next, i.e., from three-year changes. To identify from less frequent changes and to gauge the long-run relevance of the policy reforms, column 3 of Table 7 estimates a model in long differences that restricts the analysis to the 15-year change from the first to the last PISA wave. Our main finding is robust in this long-difference specification, with the estimate of the positive effect of standardized external comparison being even larger when considering only long-run changes and with the estimates of effects of the other three assessment categories remaining insignificant. Similarly, while obviously less precise, the pattern of heterogeneity by initial achievement is evident when the analysis is restricted to the category of standardized external comparisons (see column 5 of Table 8).[28]

While PISA has stringent sampling standards, there is some variation over countries and time in the extent to which specific schools and students are excluded from the target population. Main reasons for possible exclusions are inaccessibility in remote regions or very small size at the school level and intellectual disability or limited test-language proficiency at the student level (OECD (2016)). The average total exclusion rate is below 3 percent, but it varies from 0 percent to 9.7 percent across countries and waves. To test whether this variation affects our analysis, the next columns in Tables 7 and 8 control for the country-by-wave exclusion rates reported in each PISA wave. As is evident, results are hardly affected.

Finally, in 2015 PISA instituted a number of major changes in testing methodology (OECD (2016)). Most importantly, PISA changed its assessment mode from paper-based to computer-based testing. In addition, a number of changes in the scaling procedure were undertaken, including changing from a one-parameter Rasch model to a hybrid of a one- and two-parameter model and changing the treatment of non-reached testing items. We performed three robustness tests to check whether these changes in testing methodology affect our results.

First, the simplest test of whether our analysis is affected by the 2015 changes in testing methodology is to drop the 2015 wave from our regressions. As is evident from column 5 in Table 7 and column 7 in Table 8, qualitative results do not change when estimating the model just on the PISA waves from 2000 to 2012, indicating that our results cannot be driven by the indicated changes in testing mode.

Second, to address the changes in the psychometric scaling procedure, PISA recalculated countries' mean scores in the three subjects for all PISA waves since 2006 using the new 2015 scaling approach. In the final columns of Tables 7 and 8, we run our models with these rescaled country mean scores instead of the original individual scores as the dependent variable for the PISA waves 2006 to 2015. Again, qualitative results do not change, indicating that the changes in scaling approach do not substantively affect our analysis.

---

[28] Similarly, a model restricted to the category of standardized monitoring yields a significantly positive main effect and a significantly positive interaction (not shown).

Third, while no similar analysis is possible for the change in testing mode, we analyzed whether countries' change in PISA achievement from paper-based testing in 2012 to computer-based testing in 2015 is correlated with a series of indicators of the computer familiarity of students and schools that we derive from the PISA school and student background questionnaires in 2012. As indicated by Appendix Table A6, indicators of computer savviness in 2012 do not predict the change in test scores between 2012 and 2015 across countries. In particular, the change in countries' test achievement is uncorrelated with several measures of schools' endowment with computer hardware, internet connectivity, and software, as well as with several measures of students' access to and use of computers, internet, and software at home. The only exception is that the share of schools' computers that are connected to the internet is in fact *negatively* correlated with a country's change in science achievement, speaking against an advantage of computer-savvy countries profiting from the change in testing mode.

8. Conclusions

The extent of student testing and its usage in school operations have become items of heated debate in many countries, both developed and developing. Some simply declare that high-stakes tests – meaning assessments that enter into reward and incentive systems for some individuals – are inappropriate (Koretz (2017)). Others argue that increased use of testing and accountability systems are essential for the improvement of educational outcomes (World Bank (2018)) and, by extension, of economic outcomes (Hanushek and Woessmann (2015); Hanushek et al. (2015)).

Many of these discussions, however, fail to distinguish among alternative uses of tests. And, most applications of expanded student assessments used for accountability purposes have not been adequately evaluated, largely because they have been introduced in ways that make clear identification of impacts very difficult. Critically, the expansion of national testing programs has faced a fundamental analytical issue of the lack of suitable comparisons.

Our analysis turns to international comparisons to address the key questions of when student assessments can be used in ways that promote higher achievement. The conceptual framework behind the empirical analysis is a principal-agent model that motivates focusing on the strength of incentives to teachers and students, on the specific addressees of incentives created by differing test usage, and on environmental factors that affect the country-specific results of testing regimes.

The empirical analysis employs the increasingly plentiful international student assessment data that now support identification of causal implications of national testing.[29] Specifically, the six waves of the PISA assessments of the OECD between 2000 and 2015 permit country-level panel estimation that relies on within-country over-time analysis of country changes in assessment practices. We combine data across 59 countries to estimate how varying testing situations and applications affect student outcomes.

Our results indicate that accountability systems that use standardized tests to compare outcomes across schools and students produce greater student outcomes. These systems tend to have consequential implications and produce higher student achievement than those that simply report the results of standardized tests. They also produce greater achievement results than systems relying on localized or subjective information that cannot be readily compared across schools and classrooms, which have little or negative impacts on student achievement.

Moreover, both rewards to schools and rewards to students for better outcomes result in greater student learning. General comparisons of standardized testing at the school level appear to lead to somewhat stronger results than direct rewards to students that come through sorting across educational opportunities and subsequent careers.

Most interestingly from an international perspective is the finding that testing and accountability systems are more important for school systems that are performing poorly. It appears that systems that are showing strong results know more about how to boost student performance and are less in need of strong accountability systems.

Overall, the results from international comparisons of performance suggest that school systems gain from measuring how their students and schools are doing and where they stand in a comparative way. Comparative testing appears to create incentives for better performance and allows rewarding those who are contributing most to educational improvement efforts.

[29] Interestingly, even the international testing – conducted on a voluntary basis in a low-stakes situation – has come under attack for potentially harming the educational programs of countries. Recent analysis, however, rejects this potential problem (Ramirez, Schofer, and Meyer (2018)).

References

Andrab, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report cards: The impact of providing school and child test scores on educational markets." *American Economic Review* 107, no. 6: 1535-1563.

Andrews, Paul, and coauthors. 2014. "OECD and Pisa tests are damaging education worldwide." *The Guardian*: https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics (accessed June 20, 2018).

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

Braga, Michela, Daniele Checchi, and Elena Meschi. 2013. "Educational policies in a long-run perspective." *Economic Policy* 28, no. 73: 45-100.

Burgess, Simon, Deborah Wilson, and Jack Worth. 2013. "A natural experiment in school accountability: The impact of school performance information on pupil progress." *Journal of Public Economics* 106: 57-67.

Card, David. 1999. "The causal effect of education on earnings." In *Handbook of Labor Economics, Vol. 3A*, edited by Orley Ashenfelter and David Card. Amsterdam: North-Holland: 1801-1863.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2017. "Measuring the impacts of teachers: Reply." *American Economic Review* 107, no. 6: 1685-1717.

Dee, Thomas S., and Brian A. Jacob. 2011. "The impact of No Child Left Behind on student achievement." *Journal of Policy Analysis and Management* 30, no. 3: 418-446.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. "School accountability, postsecondary attainment, and earnings." *Review of Economics and Statistics* 98, no. 5: 848-862.

Dixit, Avinash. 2002. "Incentives and organizations in the public sector: An interpretative review." *Journal of Human Resources* 37, no. 4: 696-727.

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.

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

Figlio, David, and Susanna Loeb. 2011. "School accountability." In *Handbook of the Economics of Education, Vol. 3,* edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 383-421.

Figlio, David N., and Joshua Winicki. 2005. "Food for thought: The effects of school accountability plans on school nutrition." *Journal of Public Economics* 89, no. 2-3: 381-394.

Hanushek, Eric A., Susanne Link, and Ludger Woessmann. 2013. "Does school autonomy make sense everywhere? Panel estimates from PISA." *Journal of Development Economics* 104: 212-232.

Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does school accountability lead to improved student performance?" *Journal of Policy Analysis and Management* 24, no. 2: 297-327.

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

Hanushek, Eric A., and Ludger Woessmann. 2011. "The economics of international differences in educational achievement." In *Handbook of the Economics of Education, Vol. 3,* edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 89-200.

Hanushek, Eric A., and Ludger Woessmann. 2015. *The knowledge capital of nations: Education and the economics of growth*. Cambridge, MA: MIT Press.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *Journal of Law, Economics and Organization* 7: 24-52.

Hout, Michael, and Stuart W. Elliott, eds. 2011. *Incentives and test-based accountability in education*. Washington, DC: National Academies Press.

Jacob, Brian A. 2005. "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89, no. 5-6: 761-796.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten apples: An investigation of the prevalence and predictors of teacher cheating." *Quarterly Journal of Economics* 118, no. 3: 843-877.

Jürges, Hendrik, Kerstin Schneider, and Felix Büchel. 2005. "The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany." *Journal of the European Economic Association* 3, no. 5: 1134-1155.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. "Value-added modeling: A review." *Economics of Education Review* 47: 180-195.

Koning, Pierre, and Karen van der Wiel. 2012. "School responsiveness to quality rankings: An empirical analysis of secondary education in the Netherlands." *De Economist* 160, no. 4: 339-355.

Koretz, Daniel. 2017. *The testing charade: Pretending to make schools better*. Chicago: University of Chicago Press.

Laffont, Jean-Jacques, and David Martimort. 2002. *The theory of incentives: The principal-agent model*. Princeton, NJ: Princeton University Press.

Lavecchia, Adam M., Heidi Liu, and Philip Oreopoulos. 2016. "Behavioral economics of education: Progress and possibilities." In *Handbook of the Economics of Education, Vol. 5*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 1-74.

Leschnig, Lisa, Guido Schwerdt, and Katarina Zigova. 2017. "Central school exams and adult skills: Evidence from PIAAC." Unpublished manuscript, University of Konstanz.

Lüdemann, Elke. 2011. "Intended and unintended short-run effects of the introduction of central exit exams: Evidence from Germany." In Elke Lüdemann, *Schooling and the formation of cognitive and non-cognitive outcomes*. ifo Beiträge zur Wirtschaftsforschung 39. Munich: ifo Institute.

Nunes, Luis C., Ana Balcão Reis, and Carmo Seabra. 2015. "The publication of school rankings: A step toward increased accountability?" *Economics of Education Review* 49: 15-23.

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

OECD. 2016. *PISA 2015 results (volume I): Excellence and equity in education*. Paris: Organisation for Economic Co-operation and Development.

Pritchett, Lant. 2015. Creating education systems coherent for learning outcomes: Making the transition from schooling to learning. RISE Working Paper 15/005. Oxford: Research on Improving Systems of Education (RISE).

Ramirez, Francisco O., Evan Schofer, and John W. Meyer. 2018. "International tests, national assessments, and educational development (1970-2012)." *Comparative Education Review* 62, no. 3.

Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. 2014. "Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind." *American Economic Journal: Economic Policy* 6, no. 3: 207-241.

Schwerdt, Guido, and Ludger Woessmann. 2017. "The information value of central school exams." *Economics of Education Review* 56: 65-79.

Shewbridge, Claire, Melanie Ehren, Paulo Santiago, and Claudia Tamassia. 2012. *OECD Reviews of Evaluation and Assessment in Education: Luxembourg*. Paris: OECD.

Shewbridge, Claire, Eunice Jang, Peter Matthews, and Paulo Santiago. 2011. *OECD Reviews of Evaluation and Assessment in Education: Denmark*. Paris: OECD.

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. 2016. "The importance of school systems: Evidence from international differences in student achievement." *Journal of Economic Perspectives* 30, no. 3: 3-32.

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. 2018. *World Development Report 2018: Learning to realize education's promise*. Washington, DC: World Bank.

Data Appendix: Sources and Construction of Assessment Measures

We derive a series of measures of different categories of the use of student assessments over the period 2000-2015 from the PISA school background questionnaires and other sources. Information on testing usage is classified into four groups with varying strength of generated incentives: standardized external comparison, standardized monitoring, internal testing, and internal teacher monitoring. We aggregate each assessment measure to the country-by-wave level. Below, we also discuss how we combine the different indicators into an aggregate measure for each of the four assessment categories. Details on the precise underlying survey questions and any changes in question wording over time are found in Appendix Table A2.

A.1 Standardized External Comparison

Drawing on four different sources, we combine four separate indicators of standardized testing usage 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 used 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 2 in the text 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.[30]

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

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

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 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 Monitoring

Beyond externally comparative testing, the PISA school background questionnaire also provides three additional measures of standardized testing used for different types of monitoring purposes.

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 Testing

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 also report on the use of assessments of 15-year-old students in their school for purposes other than external comparisons. Our first measure of internal testing 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 Internal Teacher Monitoring

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

First, again reporting on the use 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 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 Assessment Categories

Many of the separate assessment indicators are obviously correlated with each other, in particular within each of the four groups of assessment categories. For example, the correlation between the EAG measure of national standardized exams in lower secondary school and the Eurydice measure of national tests used 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 using assessments to inform parents and using 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 assessment categories, but in the appendix tables below 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 A3. Before combining the indicators, we therefore impute missing observations in the aggregate country-by-wave dataset from a linear time prediction within each country. We then construct the combined measures of the four assessment categories as the simple average of the individual imputed indicators in each category. 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 assessment categories are also correlated with each other. In the pooled dataset of 303 country-by-wave observations, the correlations range from 0.278 between standardized external comparison and internal teacher monitoring to 0.583 between standardized monitoring and internal testing. After taking out country and year fixed effects, the correlations are lowest between standardized external comparison and all other categories (all below 0.2), moderate between standardized monitoring and the other categories (all below 0.3), and largest between internal testing and internal teacher monitoring (0.485). Because of potential multicollinearity, we first run our analyses for each aggregate assessment category separately and then report a model that considers all four categories simultaneously.

Table A1: Descriptive statistics and complete model of basic interacted specification

	Descriptive statistics			Basic model	
	Mean	Std. dev.	Share imputed	Coeff.	Std. err.
Standardized external comparison				37.304***	*(6.530)*
X initial score				-0.246***	*(0.085)*
Standardized monitoring				67.772***	*(17.139)*
X initial score				-0.776***	*(0.175)*
Internal testing				-13.858	*(12.216)*
X initial score				0.161	*(0.100)*
Internal teacher monitoring				10.432	*(25.005)*
X initial score				-0.478*	*(0.249)*
Student and family characteristics					
Female	0.504	0.500	0.001	-11.557***	*(0.946)*
Age (years)	15.78	0.295	0.001	12.284***	*(0.921)*
Immigration background					
Native student	0.892				
First generation migrant	0.054	0.221	0.034	-8.322	*(4.635)*
Second generation migrant	0.054	0.223	0.034	-2.772	*(2.736)*
Other language than test language or national dialect spoken at home	0.111	0.305	0.061	-15.133***	*(2.309)*
Parents' education					
None	0.088	0.278	0.031		
Primary	0.019	0.134	0.031	9.138***	*(2.228)*
Lower secondary	0.062	0.238	0.031	10.814***	*(2.421)*
Upper secondary I	0.108	0.307	0.031	20.951***	*(2.984)*
Upper secondary II	0.077	0.262	0.031	26.363***	*(2.559)*
University	0.265	0.435	0.031	36.135***	*(2.538)*
Parents' occupation					
Blue collar low skilled	0.08	0.265	0.041		
Blue collar high skilled	0.088	0.278	0.041	8.401***	*(1.153)*
White collar low skilled	0.168	0.366	0.041	15.520***	*(1.108)*
White collar high skilled	0.335	0.464	0.041	35.601***	*(1.552)*
Books at home					
0-10 books	0.174	0.374	0.026		
11-100 books	0.478	0.493	0.026	30.297***	*(1.908)*
101-500 books	0.276	0.442	0.026	64.817***	*(2.426)*
More than 500 books	0.072	0.255	0.026	73.718***	*(3.433)*

(continued on next page)

Table A1 (continued)

	Descriptive statistics			Basic model	
	Mean	Std. dev.	Share imputed	Coeff.	Std. err.
School characteristics					
Number of students	849.0	696.7	0.093	0.012***	*(0.002)*
Privately operated	0.193	0.383	0.071	7.500*	*(4.396)*
Share of government funding	0.802	0.289	0.106	-16.293***	*(4.596)*
Share of fully certified teachers at school	0.822	0.294	0.274	6.662**	*(2.793)*
Shortage of math teachers	0.202	0.394	0.041	-5.488***	*(1.031)*
Teacher absenteeism					
No	0.337	0.427	0.213		
A little	0.484	0.447	0.213	-0.325	*(1.175)*
Some	0.140	0.310	0.213	-6.089***	*(1.556)*
A lot	0.039	0.173	0.213	-7.715***	*(2.413)*
School's community location					
Village or rural area (<3,000)	0.092	0.281	0.056		
Town (3,000-15,000)	0.208	0.397	0.056	5.238***	*(1.768)*
Large town (15,000-100,000)	0.311	0.451	0.056	9.935***	*(2.148)*
City (100,000-1,000,000)	0.251	0.422	0.056	14.209***	*(2.594)*
Large city (>1,000,000)	0.137	0.336	0.056	17.482***	*(3.447)*
Country characteristics					
Academic-content autonomy	0.597	0.248	-	-11.666	*(8.826)*
Academic-content autonomy x Initial GDP p.c.	5.043	7.578	-	1.871***	*(0.475)*
GDP per capita (1,000 $)	27.30	20.80	-	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 5. 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 A2: Measures of student assessments: Sources and definitions

	Source (1)	Countries (2)	Waves (3)	Definition (4)	Deviation in wording in specific waves (5)
Standardized external comparison					
School-based 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 used 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 monitoring					
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".
Monitor teacher practice by assessments	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 A2 (continued)

	Source (1)	Countries (2)	Waves (3)	Definition (4)	Deviation in wording in specific waves (5)
Internal testing					
Assessments used 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 used 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).	
Internal teacher monitoring					
Assessments used to judge teacher effectiveness	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”.
Monitor teacher practice by school 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”
Monitor teacher practice by external inspector	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 A3: Country observations by wave

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

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

Table A4: Baseline model for separate underlying assessment indicators

	Math (1)	Science (2)	Reading (3)	Observations (4)	Countries (5)	Waves (6)	R^2 (7)
Standardized external comparison							
School-based 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 used 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 monitoring							
Standardized testing in tested grade	15.497**	11.051	19.380***	1,198,463	59	4	0.386
	(7.244)	*(6.901)*	*(7.169)*				
Monitor teacher practice by assessments	-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 testing							
Assessments used to inform parents	7.923	14.664**	4.234	1,705,602	59	5	0.385
	(6.594)	*(6.974)*	*(7.912)*				
Assessments used 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)*				
Internal teacher monitoring							
Assessments used to judge teacher effectiveness	-4.065	3.110	-1.981	1,705,602	59	5	0.385
	(8.249)	*(9.619)*	*(7.810)*				
Monitor teacher practice by school principal	-19.751	-10.893	-14.239	1,588,962	59	4	0.385
	(14.072)	*(10.793)*	*(10.062)*				
Monitor teacher practice by external inspector	-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 3 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 A5: Interacted model for separate underlying assessment indicators

	Math		Science		Reading	
	Main effect	X initial score	Main effect	X initial score	Main effect	X initial score
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized external comparison						
School-based external comparison	39.945***	-0.456***	43.605***	-0.484***	47.018***	-0.481***
	(10.118)	*(0.078)*	*(10.441)*	*(0.117)*	*(9.023)*	*(0.098)*
National standardized exams in lower secondary school	50.625**	-0.464**	50.720***	-0.434**	39.186	-0.273
	(18.887)	*(0.206)*	*(13.905)*	*(0.162)*	*(31.246)*	*(0.301)*
National tests used for career decisions	21.890***	-0.081	11.309	-0.002	20.983**	-0.119
	(5.524)	*(0.077)*	*(6.728)*	*(0.083)*	*(8.517)*	*(0.102)*
Central exit exams	24.550	-0.254	58.473***	-0.542***	54.899	-0.540
	(31.796)	*(0.322)*	*(18.255)*	*(0.156)*	*(46.933)*	*(0.543)*
Standardized monitoring						
Standardized testing in tested grade	46.491***	-0.460***	42.679***	-0.427***	54.278***	-0.509***
	(9.608)	*(0.108)*	*(9.829)*	*(0.105)*	*(9.918)*	*(0.104)*
Monitor teacher practice by assessments	15.863	-0.384***	44.530***	-0.508***	25.154*	-0.391***
	(14.109)	*(0.116)*	*(14.908)*	*(0.174)*	*(12.715)*	*(0.130)*
Achievement data tracked by administrative authority	28.970*	-0.417***	38.054**	-0.419**	43.775**	-0.631**
	(14.631)	*(0.129)*	*(18.191)*	*(0.198)*	*(19.113)*	*(0.242)*
Internal testing						
Assessments used to inform parents	-8.895	0.233***	-10.140	0.314***	-6.900	0.151
	(6.714)	*(0.047)*	*(8.012)*	*(0.079)*	*(10.352)*	*(0.103)*
Assessments used to monitor school progress	6.106	-0.065	2.356	0.065	6.433	-0.115
	(8.812)	*(0.115)*	*(13.376)*	*(0.177)*	*(13.825)*	*(0.177)*
Achievement data posted publicly	15.898	-0.197	22.711	-0.264*	-8.159	-0.123
	(15.782)	*(0.133)*	*(15.355)*	*(0.144)*	*(19.472)*	*(0.236)*
Internal teacher monitoring						
Assessments used to judge teacher effectiveness	0.387	-0.063	0.220	0.037	1.141	-0.043
	(14.989)	*(0.153)*	*(16.015)*	*(0.202)*	*(14.510)*	*(0.163)*
Monitor teacher practice by school principal	0.807	-0.239	31.735	-0.514**	1.358	-0.186
	(26.483)	*(0.208)*	*(21.136)*	*(0.201)*	*(20.928)*	*(0.222)*
Monitor teacher practice by external inspector	18.086	-0.370**	17.783	-0.365*	-6.485	-0.134
	(12.412)	*(0.145)*	*(17.744)*	*(0.207)*	*(16.606)*	*(0.189)*

Notes: Two neighboring cells present results of one separate regression, with "main effect" reporting the coefficient on the variable indicated in the left column and "X initial score" reporting the coefficient on its interaction with the country's PISA score in the initial year (centered at 400, so that the "main effect" coefficient shows the effect of assessments on test scores in a country with 400 PISA points in 2000). 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 A4 for numbers of observations, countries, and waves and Table 3 for the included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

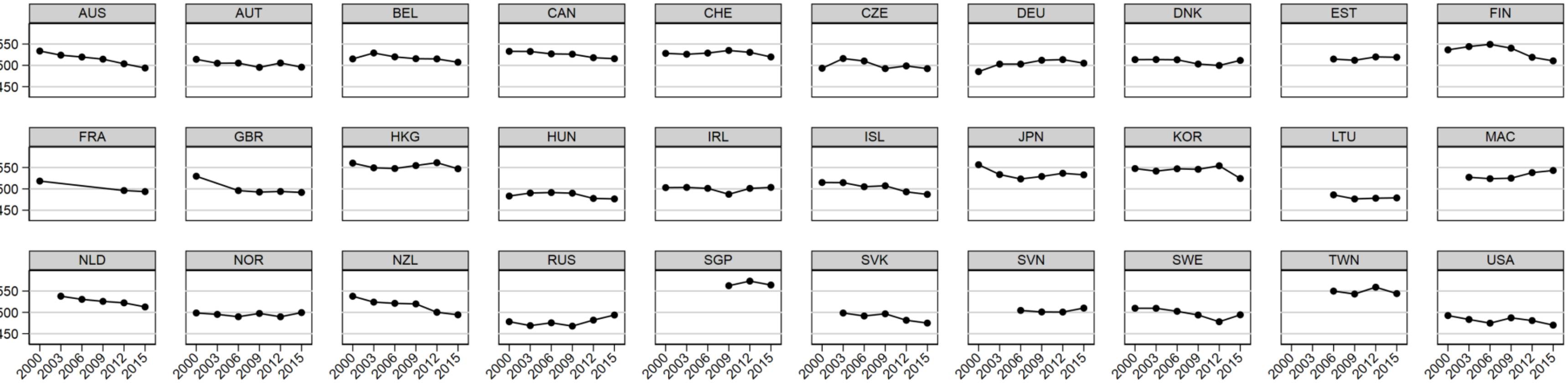
Table A6: Correlation of computer indicators in 2012 with change in PISA score from 2012 to 2015 at the country level

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

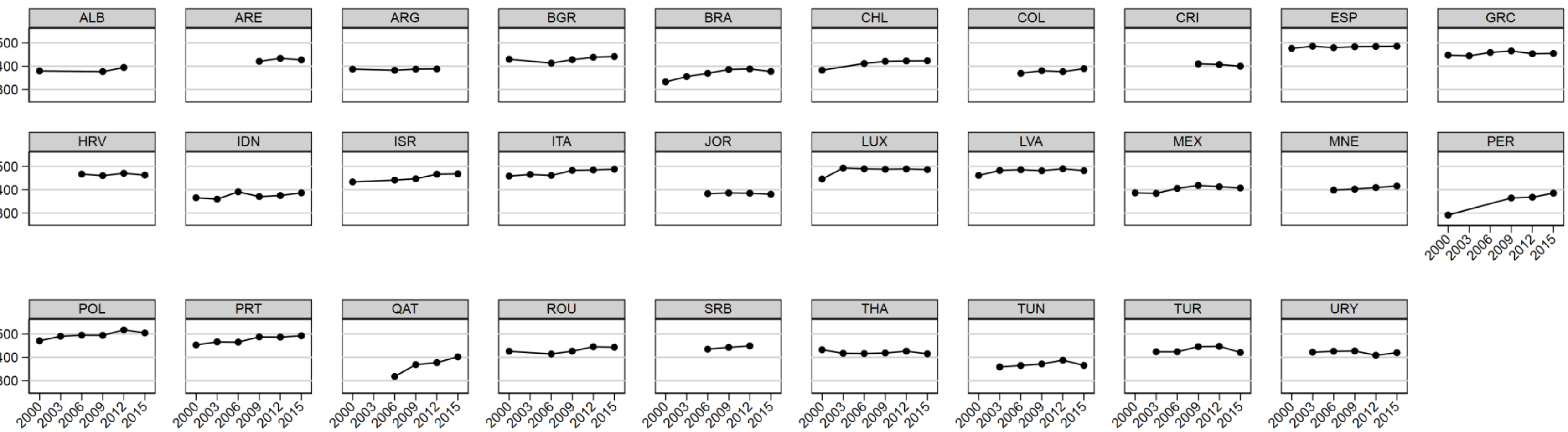
Notes: Correlation between the respective computer indicator (2012) indicated in the first column with the change in PISA test scores (2012-215) in the subject indicated in the header. Sample: 56 country-level observations of countries participating in the PISA waves 2012 and 2015. *p*-values in parentheses.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

Figure 1: PISA math achievement in 2000-2015

Panel A: Countries above initial median achievement

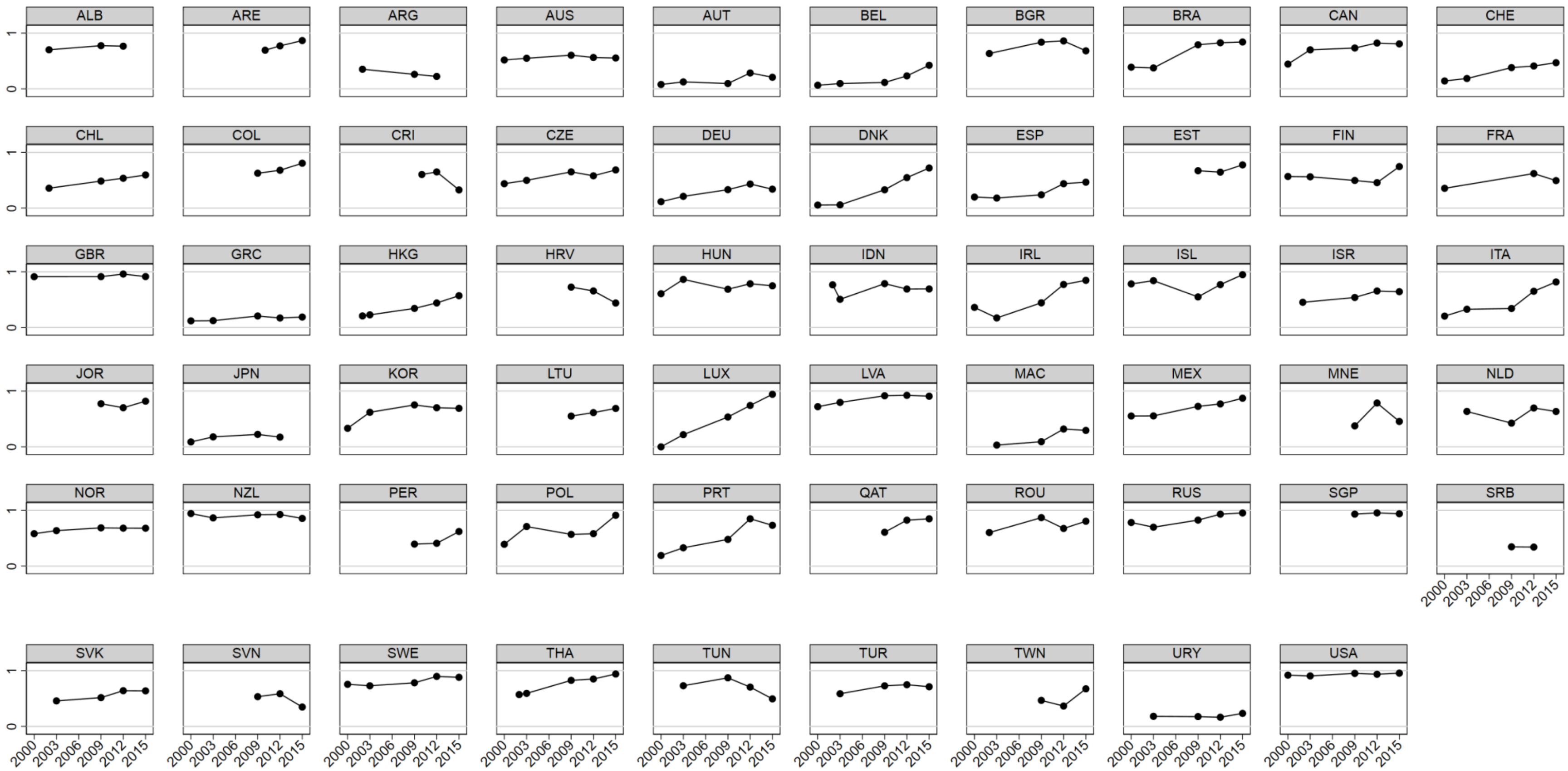


Panel B: Countries below initial median achievement



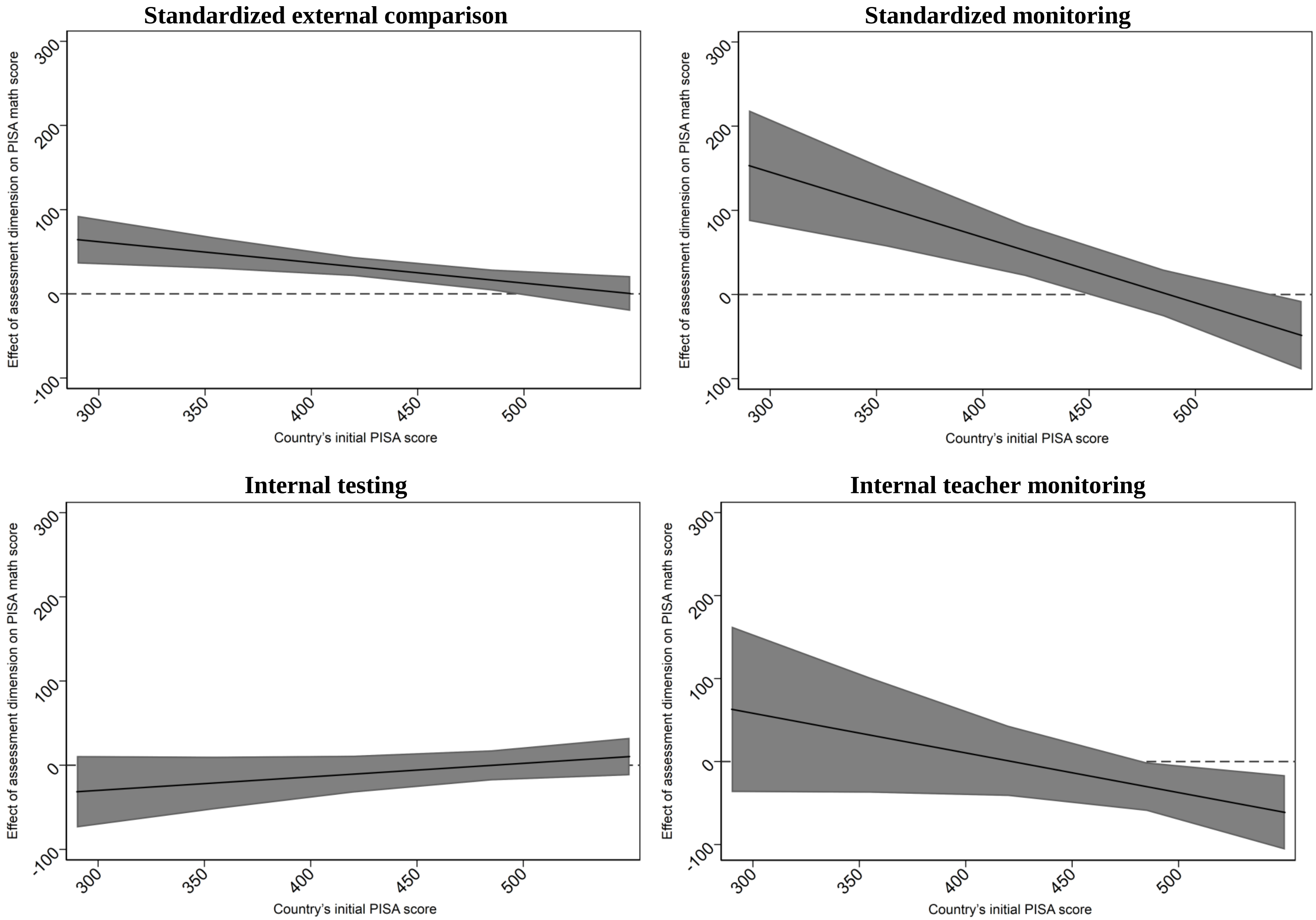
Notes: Country mean achievement in PISA math test. Country sample split at median of initial achievement level for expositional reasons. Country identifiers are listed in Table 1. Own depiction based on PISA micro data.

Figure 2: School-based external comparison in 2000-2015



Notes: Country share of schools with use of assessments for external comparison. Country identifiers are listed in Table 1. Own depiction based on PISA micro data.

Figure 3: Effect of student assessments on math performance by initial achievement levels



Notes: Average marginal effects of student assessments on PISA math score by initial country achievement, with 95 percent confidence intervals. See first column of Table 5 for underlying model.

Table 1: Selected indicators by country

	OECD	PISA math score		School-based external comparison		National standardized exams in lower sec. school		National tests used for career decisions	
	2015	2000	2015	2000	2015	2000	2015	2000	2015
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
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	.	.
Austria (AUT)	1	514	496	0.08	0.21	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	.	.
Chile (CHL) [a]	1	383	423	0.36	0.60	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
Denmark (DNK)	1	514	512	0.06	0.72	1	1	1	1
Estonia (EST) [c]	1	515	519	0.67	0.78	1	1	.	.
Finland (FIN)	1	536	511	0.57	0.75	0	0	.	.
France (FRA)	1	518	494	0.36	0.50	1	1	.	.
Germany (DEU)	1	485	505	0.12	0.34	.	.	0	1
Greece (GRC)	1	447	455	0.12	0.19	0	0	0	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	.	.	.	.
Ireland (IRL)	1	503	504	0.36	0.85	1	1	1	1
Israel (ISR) [a]	1	434	468	0.45	0.64	0	0	.	.
Italy (ITA)	1	459	489	0.21	0.82	1	1	0	1
Japan (JPN)	1	557	533	0.09	0.17	0	0	.	.
Jordan (JOR) [c]	0	384	381	0.77	0.82	.	.	.	.
Korea (KOR)	1	548	524	0.33	0.69	0	0	.	.
Latvia (LVA)	1	462	482	0.72	0.91	1	1	1	1

(continued on next page)

Table 1 (continued)

	OECD	PISA math score		School-based external comparison		National standardized exams in lower sec. school		National tests used for career decisions	
	2015	2000	2015	2000	2015	2000	2015	2000	2015
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lithuania (LTU) [c]	0	486	479	0.55	0.69	.	.	0	0
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
New Zealand (NZL)	1	538	494	0.94	0.86	0	0	.	.
Norway(NOR)	1	499	500	0.58	0.68	0	1	0	1
Peru (PER) [a]	0	292	386	0.40	0.62	.	.	.	.
Poland (POL)	1	471	505	0.39	0.91	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	.	.	.	.
Slovak Republic (SVK) [b]	1	499	475	0.46	0.64	0	0	.	.
Slovenia (SVN) [c]	1	505	510	0.54	0.35	0	0	0	0
Spain (ESP)	1	476	486	0.20	0.47	0	0	.	.
Sweden (SWE)	1	510	494	0.76	0.88	0	0	1	1
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	.	.
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
United States (USA)	1	493	470	0.92	0.96	0	1	.	.
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

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 2: Descriptive statistics of assessment measures

	Mean (1)	Std. dev. (2)	Min (3)	Max (4)	Countries (5)	Waves (6)
Standardized external comparison	**0.518**	**0.271**	**0.022**	**0.978**	**59**	**6**
School-based external comparison	0.573	0.251	0	0.960	59	5
National standardized exams in lower secondary school	0.292	0.452	0	1	37	6
National tests used for career decisions	0.601	0.481	0	1	18	6
Central exit exams	0.689	0.442	0	1	30	6
Standardized monitoring	**0.714**	**0.160**	**0.219**	**0.996**	**59**	**6**
Standardized testing in tested grade	0.721	0.233	0	1	59	4
Monitor teacher practice by assessments	0.750	0.191	0.128	1	59	4
Achievement data tracked by administrative authority	0.723	0.201	0.070	1	59	4
Internal testing	**0.684**	**0.147**	**0.216**	**0.963**	**59**	**6**
Assessments used to inform parents	0.892	0.185	0.141	1	59	5
Assessments used to monitor school progress	0.770	0.209	0	1	59	5
Achievement data posted publicly	0.393	0.239	0.016	0.927	59	4
Internal teacher monitoring	**0.553**	**0.216**	**0.026**	**0.971**	**59**	**6**
Assessments used to judge teacher effectiveness	0.532	0.261	0	0.992	59	5
Monitor teacher practice by school principal	0.773	0.262	0.049	1	59	4
Monitor teacher practice by external inspector	0.402	0.255	0.006	0.994	59	4

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

Table 3: The effect of different dimensions of student assessments on student achievement: Fixed-effects panel models

	Math					Science	Reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Standardized external comparison	26.365***				28.811***	23.282***	28.424***
	(6.058)				*(6.126)*	*(6.144)*	*(5.911)*
Standardized monitoring		-4.800			-5.469	1.252	-2.036
		(15.238)			*(14.062)*	*(13.950)*	*(13.148)*
Internal testing			2.093		7.491	17.669	-12.660
			(10.067)		*(11.646)*	*(13.155)*	*(14.736)*
Internal teacher monitoring				-23.478	-35.850**	-27.549*	-25.358
				(14.518)	*(15.680)*	*(14.226)*	*(15.835)*
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student observations	2,094,856	2,094,856	2,094,856	2,094,856	2,094,856	2,094,705	2,187,415
Country observations	59	59	59	59	59	59	59
Country-by-wave observations	303	303	303	303	303	303	303
R^2	0.391	0.390	0.390	0.390	0.391	0.348	0.357

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. Control variables include: 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; year fixed effects. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 4: Disaggregation of standardized external comparisons into school-based and student-based comparisons

	Math	Science	Reading
	(1)	(2)	(3)
School-based external comparison	25.015***	21.317**	23.480***
	(7.667)	*(8.246)*	*(7.291)*
Student-based external comparison	17.309***	15.198***	14.481***
	(3.620)	*(3.883)*	*(3.753)*
Standardized monitoring	-4.658	-8.333	-8.400
	(16.599)	*(15.007)*	*(14.602)*
Internal testing	4.896	13.419	-16.890
	(13.686)	*(15.306)*	*(18.616)*
Internal teacher monitoring	-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 3 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 5: Effects of student assessments by initial achievement level: Fixed-effects panel models

	Math (1)	Science (2)	Reading (3)	Math (4)	Science (5)	Reading (6)
Standardized external comparison	37.304***	28.680***	47.977***			
	(6.530)	*(8.222)*	*(9.005)*			
X initial score	-0.246***	-0.149	-0.345***			
	(0.085)	*(0.101)*	*(0.113)*			
School-based external comparison				45.740***	39.343*	49.581**
				(15.067)	*(21.244)*	*(21.699)*
X initial score				-0.385**	-0.347	-0.361
				(0.165)	*(0.229)*	*(0.248)*
Student-based external comparison				15.138**	7.120	2.535
				(6.518)	*(10.564)*	*(5.975)*
X initial score				-0.019	0.079	0.147
				(0.105)	*(0.160)*	*(0.091)*
Standardized monitoring	67.772***	86.860***	88.701***	72.689***	77.183**	116.503***
	(17.139)	*(20.263)*	*(21.396)*	*(26.701)*	*(34.691)*	*(31.505)*
X initial score	-0.776***	-0.989***	-1.026***	-0.756***	-0.921**	-1.378***
	(0.175)	*(0.255)*	*(0.260)*	*(0.273)*	*(0.387)*	*(0.377)*
Internal testing	-13.858	-14.734	-26.214	-14.462	-0.669	-44.234
	(12.216)	*(15.155)*	*(17.261)*	*(21.562)*	*(35.177)*	*(33.433)*
X initial score	0.161	0.289**	0.082	0.159	0.087	0.219
	(0.100)	*(0.143)*	*(0.185)*	*(0.201)*	*(0.324)*	*(0.337)*
Internal teacher monitoring	10.432	18.210	-22.463	-0.620	2.077	-42.345
	(25.005)	*(25.338)*	*(32.946)*	*(32.969)*	*(42.956)*	*(43.058)*
X initial score	-0.478*	-0.407	0.077	-0.290	-0.191	0.421
	(0.249)	*(0.289)*	*(0.317)*	*(0.355)*	*(0.506)*	*(0.436)*
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Student observations	2,094,856	2,094,705	2,187,415	1,672,041	1,671,914	1,751,351
Country observations	59	59	59	42	42	42
Country-by-wave observations	303	303	303	230	230	230
R^2	0.393	0.349	0.359	0.350	0.316	0.323

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. Initial score: country's PISA score in the initial year (centered at 400, so that main-effect coefficient shows effect of assessments on test scores in a country with 400 PISA points in 2000). Sample: student-level observations in six PISA waves 2000-2015. See Table 3 for included control variables. Complete model of specification in column 1 displayed in Table A1. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 6: Placebo test with leads of assessment reforms

	Math	Science	Reading
	(1)	(2)	(3)
Standardized external comparison	25.104***	24.567***	27.787***
	(6.316)	*(5.242)*	*(7.501)*
Standardized monitoring	-16.172	-3.734	4.660
	(18.139)	*(19.288)*	*(18.490)*
Internal testing	14.305	19.522	-17.675
	(15.367)	*(21.238)*	*(20.325)*
Internal teacher monitoring	-35.785	-38.797*	-31.560
	(22.833)	*(19.796)*	*(19.079)*
Lead (Standardized external comparison)	12.119	4.475	5.746
	(11.045)	*(8.506)*	*(9.351)*
Lead (Standardized monitoring)	-15.195	-11.138	-17.220
	(13.881)	*(16.216)*	*(19.718)*
Lead (Internal testing)	6.965	-7.014	5.567
	(14.408)	*(15.286)*	*(14.069)*
Lead (Internal teacher monitoring)	-5.394	20.922	-15.352
	(17.088)	*(18.269)*	*(17.759)*
Control variables	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Student observations	1,638,149	1,638,084	1,710,196
Country observations	59	59	59
Country-by-wave observations	235	235	235
R^2	0.396	0.350	0.361

Notes: Dependent variable: PISA test score in subject indicated in the header. Lead indicates values of test usage category from subsequent period, i.e., before its later introduction. 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 3 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 7: Robustness tests: Base specification

	OECD countries	Non-OECD countries	Long difference (2000+2015 only)	Control for exclusion rates	Without 2015	Rescaled test scale
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized external comparison	29.303***	16.429*	61.184***	27.431***	31.205***	33.247***
	(7.471)	*(8.387)*	*(9.981)*	*(6.160)*	*(5.996)*	*(8.937)*
Standardized monitoring	4.671	-10.835	-16.515	-5.817	-10.664	-10.906
	(15.292)	*(19.542)*	*(19.191)*	*(13.900)*	*(15.272)*	*(15.499)*
Internal testing	1.727	15.001	19.131	5.665	6.381	5.434
	(13.704)	*(14.846)*	*(26.395)*	*(10.619)*	*(16.582)*	*(9.393)*
Internal teacher monitoring	-25.693	-22.625	-13.438	-35.308**	-46.460**	-29.108
	(16.190)	*(21.114)*	*(23.881)*	*(15.460)*	*(20.489)*	*(21.312)*
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Student observations	1,434,355	660,501	404,344	2,045,454	1,679,250	1,698,971
Country observations	35	24	38	59	59	58
Country-by-wave observations	197	106	76	289	247	223
R^2	0.283	0.441	0.365	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 3 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 8: Robustness tests: Interacted specification

	OECD countries		Non-OECD countries	Long difference (2000+2015 only)		Control for exclusion rates	Without 2015	Rescaled test scale
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standardized external comparison	51.462	22.346***	26.378***	18.944	69.060***	35.439***	35.085***	60.655***
	(30.820)	*(7.479)*	*(5.872)*	*(24.016)*	*(17.063)*	*(7.362)*	*(9.954)*	*(15.693)*
X initial score	-0.359		-0.374***	0.211	-0.272	-0.217**	-0.189	-0.507**
	(0.326)		*(0.106)*	*(0.222)*	*(0.187)*	*(0.096)*	*(0.125)*	*(0.196)*
Standardized monitoring	58.619*	64.291*	20.508	42.848		61.292***	55.777***	8.894
	(32.496)	*(34.495)*	*(18.675)*	*(31.020)*		*(20.757)*	*(19.008)*	*(30.447)*
X initial score	-0.547*	-0.636*	-0.319*	-0.510		-0.716***	-0.703***	-0.152
	(0.321)	*(0.343)*	*(0.185)*	*(0.335)*		*(0.207)*	*(0.209)*	*(0.274)*
Internal testing	18.179	6.054	-10.840	-106.185**		-11.153	-1.941	-5.212
	(29.982)	*(11.613)*	*(13.040)*	*(45.672)*		*(12.372)*	*(31.980)*	*(15.369)*
X initial score	-0.134		0.232**	1.119**		0.126	0.020	0.076
	(0.262)		*(0.105)*	*(0.473)*		*(0.105)*	*(0.334)*	*(0.131)*
Internal teacher monitoring	46.444	61.681	0.663	72.304		4.894	8.063	-72.152**
	(38.979)	*(40.538)*	*(20.416)*	*(52.716)*		*(29.938)*	*(40.220)*	*(35.725)*
X initial score	-0.733*	-0.887*	-0.342	-1.106*		-0.402	-0.681	0.666*
	(0.385)	*(0.387)*	*(0.315)*	*(0.551)*		*(0.292)*	*(0.434)*	*(0.359)*
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student observations	1,434,355	1,434,355	660,501	404,344	404,344	2,045,454	1,679,250	1,698,971
Country observations	35	35	24	38	38	59	59	58
Country-by-wave observations	197	197	106	76	76	289	247	223
R^2	0.285	0.285	0.443	0.367	0.365	0.389	0.400	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. Initial score: country's PISA score in the initial year (centered at 400, so that main-effect coefficient shows effect of assessments on test scores in a country with 400 PISA points in 2000). Sample: student-level observations in six PISA waves 2000-2015. Rescaled test scale available for waves 2006-2015 only. See Table 3 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.