Testing

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

Abstract

The significant expansion of student testing has not generally been linked to educational outcomes. We investigate how different testing regimes – providing varying information to parents, teachers, and decision makers – relate to student achievement. We exploit PISA data for two million students in 59 countries observed from 2000-2015. Removing country and year fixed effects, we investigate how testing reforms affect country performance. In low- and medium-performing countries, more standardized testing is associated with higher student achievement, while added internal reporting and teacher monitoring are not. But in high-performing countries expansion of standardized internal testing and teacher monitoring appears harmful.

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

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

September 28, 2021

Annika Bergbauer was a researcher at the ifo Institute at the University of Munich. Eric A. Hanushek is a senior fellow at the Hoover Institution, Stanford University and is affiliated with CESifo and NBER (hanushek@stanford.edu). Ludger Woessmann is a professor of economics at the University of Munich and director of the ifo Center for the Economics of Education and is affiliated with CESifo and IZA. The authors gratefully acknowledge comments from two anonymous referees, Scott Imberman, Joachim Winter, and participants at seminars in Berlin, Maastricht, Madrid, and Moscow, the American Economic Association at San Diego, the European Association of Labour Economists in Lyon, the Association for Education Finance and Policy in Kansas City, the Spring Meeting of Young Economists in Palma de Mallorca, the German Economic Association in Freiburg, the CESifo Area Conference on Economics of Education in Munich, the briq Workshop Skills, Preferences and Educational Inequality in Bonn, the CRC 190 meeting in Ohlstadt, the BGPE Research Workshop in Bamberg, and the Center Seminar of the ifo Center for the Economics of Education in Munich. This work was supported by the Smith Richardson Foundation. The contribution by Bergbauer and Woessmann is part of project CRC TRR 190 of the German Science Foundation. Woessmann acknowledges support by Research on Improving Systems of Education (RISE) which is funded by UK Aid and Australian Aid. The authors do not have relevant or material financial interests that relate to the research described in this paper. An Online Appendix has been included with this article. The data used in this article are available online at https://doi.org/10.7910/DVN/BUID9K (Bergbauer, Hanushek, and Woessmann 2021).

I. Introduction

Student testing 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 testing 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 of testing and alternative school environments, leading to a confused debate.

Understanding the impact of student testing requires consideration of a test's informational content. This paper exploits international comparisons to examine the heterogeneous contribution of different types of testing to overall levels of student achievement. We argue that to varying degrees student assessments (used as a synonym for testing here) create the informational backbone for alternative policies and incentive systems and thus lay the foundation for various behavioral results. Based on the conceptual framework of a principal-agent model, we focus on the information created by a continuum of forms of testing from teacher-developed assessments to standardized external comparisons.

We are interested in the reduced-form effect of testing per se, rather than how the generated information is used in any particular policies or accountability systems. In various applications, such as NCLB in the United States, testing becomes virtually synonymous with its specific use, but we find it useful to separate these.[1] Indeed, with the implementation and the opportunity

[1] In the United States, consideration of testing is mostly restricted to the specific accountability systems exemplified by No Child Left Behind (NCLB), the 2001 federal law that required states to test student outcomes

costs of testing, the net effect of testing may turn negative if the created information does not induce positive behavioral changes.

Our empirical analysis uses data from the Programme for International Student Assessment (PISA) to construct a panel of country observations of student performance. 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 measures of student outcomes, but also rich background information on both students and schooling institutions in the different countries.

From the PISA surveys and other international data sources, we develop measures of usage of different types of student testing based on the nature of the information and the kinds of policies that can be supported. We distinguish four categories of testing: (1) standardized testing with external comparison; (2) standardized testing for internal comparison; (3) internal reporting; and (4) teacher monitoring. While generally abstracting from particular uses, we separate the last category (teacher monitoring) from other forms of internal reporting, because we cannot break apart the informational component from its specific use.

The last two decades are a period of rapid change in student assessment policies across countries, allowing us to link information policies to student outcomes in fixed-effects panel models.[2] Our identification relies on changes in student assessment regimes within countries over time. The basic idea of our approach can conveniently be illustrated when plotting the long-run change in countries' average PISA math score between 2000 and 2015 against the change in the prevalence of standardized external testing. Figure 1 displays this plot for the most stringent

annually in grades 3-8 and to intervene in schools that were not on track to bring all students to state-defined proficiency levels.

[2] 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).

category of testing, changes in standardized testing with external comparison. It clearly shows that countries that expanded the use of this type of testing over the fifteen-year period systematically saw the achievement of their students improve.[3]

The underlying fixed-effects panel model uses the individual student data for estimation at the micro level but, in order to avoid bias from within-country selection of students into schools, measures the informational treatment variables as country aggregates at each point in time. Conditioning on country and year fixed effects allows us to account for unobserved time-invariant country characteristics as well as common time-specific shocks.

The positive average association of standardized external test information with math performance in Figure 1 nonetheless masks important heterogeneity of treatment effects across the four categories of testing and across countries at different performance levels. In addition to separating different kinds of test information, our main model allows for heterogeneous treatment effects based on each country's initial achievement score. This refinement in estimation provides an interesting nuance to the average relationship depicted in Figure 1: the impact of standardized testing with external comparison is significantly positive in initially low-achieving and medium-achieving countries but turns insignificant for high-achieving countries. A similar pattern is evident for standardized testing for internal comparison, where effects even turn negative at very high levels of initial country performance. Similar negative effects for very high-achieving countries are found for teacher monitoring including the oft-touted use of inspectorates, which does not have significant effects even at lower levels of initial achievement.

[3] The variables on both axes of this added-variable plot are conditional on a rich set of student, school, and country controls, based on a long-difference fixed-effect panel model estimated at the individual student level (discussed in detail in Section VI.B). As a result, the depicted values deviate from the raw country data shown in Table A1 in the Online Appendix. While conditioning on changes in other relevant variables provides a cleaner picture of the association of interest, just plotting the raw data gives a very similar result of a strong and significant positive association between changes in standardized external comparisons and changes in PISA scores (not shown).

Internal reporting that simply informs or monitors progress without standardized comparability on the other hand has little discernible association with learning outcomes across countries. The overall pattern of results suggests that positive effects of testing are restricted to standardized forms of testing in settings where schools do not already perform at very high levels. But the effects can even turn negative in high-performing settings when the information does not readily support comparative policies.

The estimated effects of testing are substantial. For example, going from no use to full use of standardized testing with external comparison is associated with an average increase in student achievement of roughly a quarter of a standard deviation (s.d.), implying an impact roughly equivalent to what students learn during an entire school year (Woessmann 2016). With the indicated heterogeneity, effect sizes range from twice as large in the lowest-achieving countries to zero in the highest-achieving countries. In the average OECD country (in terms of initial achievement), the size is 0.13 s.d. The overall order of magnitude is similar to results found for other institutional policies such as school decision making autonomy (e.g., Hanushek, Link, and Woessmann 2013) or (oversubscribed) charter schools (e.g., Abdulkadiroğlu et al. 2011), but substantially higher than any effects found for most resource policies (see Woessmann (2016) for an overview). For comparison, Lavy (2015) finds that a one-hour increase in weekly instruction time raises student achievement by 0.06 s.d. in developed countries and half as much in developing countries.

A number of specification tests provide evidence against substantial bias from coincidental other policies. A placebo test employing leads of the testing variables confirms that the changes in assessment usage are not systematically linked to a country's prior outcome conditions. The effect of external comparative testing emerges irrespective of whether testing is measured

gradually by principals' reports on test usage or dichotomously by legislated national policy reforms. Robustness tests show that results are not affected by any individual country, by consideration of subsets of countries, by controlling for test exclusion rates, by changes in PISA testing procedures, or by estimating the model collapsed to the country-by-wave level.

Our cross-country approach draws on the substantial variation in forms of testing that exists across countries. Testing policies are often set at the national level, making it difficult to construct an adequate comparison group for evaluation in a within-country setting. By moving to international comparisons, it is possible to study these national policies, to investigate which aspects of assessment systems generalize to larger settings and which do not, and to consider how overall country environments interact with the specifics of assessment systems. Only the comparative perspective allows for an investigation of the richness of the full continuum of different forms of testing by exploiting the counterfactual from countries that did not reform at the same point in time. Of course, these advantages come at a cost, because precisely identifying the separate impact of information across nations offers its own challenges. We are not able to investigate the details of specific national schooling programs and policies that might rely on the information created, and there is uncertainty in separating the changes in information flows from the range of individual programs, policies, and usages developed from them. Through a variety of approaches, we can reduce concerns of substantial bias from the most obvious sources in the cross-country setting, but we cannot completely eliminate any possible biases.

In the literature, as well as in policy discussions, the term testing is frequently taken to be synonymous with accountability. We think it is useful to separate these two concepts. Accountability systems link various learning outcomes to rewards, punishments, and incentives for different actors, and they can differ widely in form and substance. Moreover, any given

student assessment can simultaneously be used in multiple ways for accountability purposes. Testing also enters into educational decision making in broader ways than just accountability. Information from student assessments is used in policy formulation, program evaluations, and regulatory structures. We therefore think of the various student assessments as providing information necessary for implementing different sets of policies, potentially inducing behavioral changes that affect learning outcomes.

From a policy perspective, a focus on testing is useful. Policy makers cannot always fully control how information is used by different actors, but they can in general affect the type of testing information that is provided. We interpret our study as a reduced-form analysis that focuses on how the informational content of different testing regimes can support policies, programs, and actions that lead to altered student outcomes. It does not delve into the structures of any specific policies or accountability systems that are subsequently attached to the assessment.

With this perspective, we expand the more specifically focused perspectives of the literature on various forms of accountability – discussed more fully within our conceptual framework in Section II – such as the NCLB legislation in the US (surveyed in Figlio and Loeb 2011), central exit exams (surveyed in Woessmann 2018), publications of school rankings (e.g., Burgess, Wilson, and Worth 2013), and school report cards (e.g., Andrab, Das, and Khwaja 2017). Our analysis also contributes to recent experimental studies of various forms of information provision to parents (e.g., York, Loeb, and Doss 2019; Bergman 2021; Bergman and Chan 2021) and of linking tested outcomes to incentives in education more generally, either for students (e.g., Angrist and Lavy 2009; Kremer, Miguel, and Thornton 2009; Fryer 2011; Bettinger 2012) or for teachers (e.g., Lavy 2009; Glewwe, Ilias, and Kremer 2010; Muralidharan and Sundararaman

2011). None of these experimental studies addresses the effects of testing per se, even though testing undergirds each. Importantly, our cross-country approach allows investigation of the full continuum of testing policies from internal assessments to externally benchmarked comparisons. The estimated reduced-form effects of information on student achievement implicitly include various forms of general equilibrium effects and of behavioral responses by different actors, but those details are beyond the scope of this study.

The next section develops a conceptual framework highlighting the importance of different forms of student assessments. Section III introduces the data, and Section IV develops the empirical model. Section V presents our results, concentrating on analyses of the heterogeneous treatment effects inherent in information provision. Section VI reports a placebo test and other specification tests, and Section VII shows a series of robustness analyses. Section VIII concludes.

II. A Framework for Evaluating Testing

We begin with the conceptual framework of a principal-agent structure and identify the continuum of information from internal to external forms of testing as the key motivation for our empirical modeling.[4]

A. The Principal-Agent Framework

A useful way to characterize the structure of educational systems is as a tree of principal-agent problems (Laffont and Martimort 2002).[5] Parents care about their child's achievement of

[4] For a more extensive discussion of the conceptual framework that covers the underlying value functions and the technology of student assessment in greater detail, see the working-paper version of this paper (Bergbauer, Hanushek, and Woessmann 2018).

[5] See Bishop and Woessmann (2004) and Pritchett (2015) for related analyses of education systems as principal-agent relationships.

knowledge and skills, which directly affects their long-run economic outcomes (Card 1999; Hanushek et al. 2015). Parents, however, cannot directly choose the effort level of their children. Instead, they may offer short-term rewards for learning to their child and try as best as possible to observe and control child effort. Similarly, parents cannot fully control the production of the child's achievement in schools, where a key element is the effort levels of teachers and other school personnel.

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 agents. Teaching in the classroom and studying at a desk involve asymmetric information, where the respective principal cannot fully monitor the behavior of the respective agent. Because of the incomplete monitoring and the specific objective functions of parents, teachers, and students, one cannot simply assume that the actions of children and teachers will lead to the optimal result for parents.

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 also parent-administrator, administrator-administrator, and administrator-teacher information and monitoring problems that are relevant to incentive design questions.

If parents had full information about the effort levels of students, teachers, and administrators, they could effectively contract with each to maximize their own objective function. However, actually obtaining and monitoring effort levels is generally costly, and the differing preferences may lead to suboptimal effort levels by students, teachers, and administrators from the perspective of parents.

A common solution is to introduce outside assessments of the outcomes of interest. By creating outcome information, student assessments provide a mechanism for developing better incentives to elicit increased effort by students, teachers, and administrators, thereby ultimately raising student achievement levels to better approximate the desires of the parents.

Nonetheless, a number of issues related to the type and accuracy of information that the tests generate makes the impact of testing a complicated empirical question. There is a classical identification problem of separating the joint effort levels of teachers and students in order to provide the right incentives. Additionally, imperfect measurement technologies may not provide complete information on achievement.[6] Here, we highlight that the internal vs. external character of the information generated by the test is a major source of its ability to solve the underlying principal-agent problems, with important implications for the potential impact of testing.

B. The Continuum from Internal Reporting to Standardized External Comparison

Testing is a ubiquitous component of schooling, but not all tests create the same kind of information. By far the most common type of testing is teacher-developed tests, a form of internal testing that is used both to guide instruction and to provide feedback to students and parents. The key feature of teacher-developed tests is that their results are very difficult to compare across teachers, implying they do not provide the kind of information that would mitigate the principal-agent problem between parents and teachers even if it helps solve the teacher-student problem. More generally, if not standardized across schools, the information

[6] 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. 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), gaming of 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 is the subject of a much larger literature.

generated by internal testing does not directly allow parents and administrators to monitor school performance.[7] At the most extreme, costly tests that have no consequences for any of the actors may be inconsequential for overall performance because nobody takes them seriously.

At the other end of the continuum of testing are standardized tests that allow for external comparisons of student outcomes in different circumstances. These tests are normed to relevant population performance. The comparability of the generated achievement information suggests the possibility of using the tests to support incentives not only for students but also for administrators and teachers by making external information available to parents, policy makers, and the general public.[8] As a general principle, we expect information that is useful for producing stronger incentives will have larger potential impacts on overall achievement.

The information created by standardized testing with external comparison may lead to different incentives across the various actors, and information that helps solve one principal-agent problem may leave others untouched. In some cases, the actions of the individual actors may be plausibly separated. For example, centralized exit exams that have consequences for further schooling of students may be linked to strong incentives for student effort while having limited impact on teacher effort.[9] 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

[7] For example, an extension of teacher-developed tests is periodic content testing provided by external producers (so-called formative assessments). Again, parents generally cannot compare outcomes externally.

[8] For example, school rankings may be published to the general public (see Koning and van der Wiel (2012) for the Netherlands, Burgess, Wilson, and Worth (2013) for Wales, and Nunes, Reis, and Seabra (2015) for Portugal), 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).

[9] By affecting chances to enter specific institutions and fields of higher education and the hiring decisions of employers, central exit exams usually have real consequences for students (see Bishop 1997; Woessmann 2003; Jürges, Schneider, and Büchel 2005; Woessmann et al. 2009; Luedemann 2011; Schwerdt and Woessmann 2017; Woessmann 2018).

students and their efforts.[10] Similarly, differential rewards to teachers based upon test-score growth are high stakes for the teachers, but not for the students. However, even in these cases, strategic complementarity or substitutability in the effort levels of the different actors might produce some ambiguity in responses.[11]

Between the two ends of the information continuum are standardized forms of testing that do not include external comparisons. For example, teachers may regularly use assessments in their classroom that are standardized rather than self-developed but that are not used for a comparison to students in other schools or to the district or national average. In addition, use of standardized tests may support a variety of report card systems without external comparison. It is less obvious that this type of information would solve the described principal-agent problems, and systematic behavioral responses are less likely.[12]

In general, our analysis of testing abstracts from the particular use to which the generated achievement information is put. However, there is one category of internal testing – measures aimed at teacher monitoring – that cannot be separated from a particular use. For example, consider inspections of teacher lesson plans as an element of the monitoring of teacher practices. Or consider principal assessment of teacher's classroom performance based on a standard rubric. We cannot identify whether it is the availability of testing per se or its particular use that is having any impact. Therefore, we will separate information surrounding teacher monitoring from

[10] For analyses of the effects of NCLB and predecessor reforms, see Hanushek and Raymond (2005), Jacob (2005), Neal and Schanzenbach (2010), Rockoff and Turner (2010), Dee and Jacob (2011), Rouse, Hannaway, Goldhaber, and Figlio (2013), Reback, Rockoff, and Schwartz (2014), and Deming, Cohodes, Jennings, and Jencks (2016); see Figlio and Loeb (2011) for a survey.

[11] For a general discussion, see Todd and Wolpin (2003) and De Fraja, Oliveira, and Zanchi (2010). Reback (2008) finds that students do respond in cases where their performance is important to school ratings.

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

other forms of internal reporting in our empirical application below – with the acknowledgment that this is not purely a category of information provision.

These considerations lead us to focus on four categories of testing: (1) standardized testing with external comparison (SCOMP), (2) standardized testing for internal comparison (SINT), (3) internal reporting (IRPT), and (4) teacher monitoring (TMON). The prior principal-agent considerations indicate that SCOMP supports more and stronger incentives than SINT and that both provide stronger information than IRPT. Because of the ambiguity of information for TMON, we do not have strong priors on its incentives. These considerations lead to an expectation that the student achievement impacts are ordered as SCOMP>SINT>IRPT.

Apart from providing incentives, the implementation of most forms of testing also entails some costs, not least the opportunity costs of time of the people involved. For example, testing time in the classroom may take away from students' learning time, and inspections of teacher lessons may create bureaucratic burden in the preparation and evaluation phase and reduce teachers' time and focus in class. If these costs exceed the benefits from improved incentives, the net effect of some forms of testing may even turn negative in certain settings. We do not, however, have direct information on costs of any of the assessments.

While the discussion so far did not differentiate among specific school environments, the policy uses of information from student testing across countries are unlikely to be uniform across systems with different levels of institutional development.[13] For example, a set of high-performing schools might be expected to know how to react to achievement signals and different rewards. Therefore, they may react more strongly to any type of incentive structure created from

[13] Another dimension of heterogeneity may be across parents within a system, in that parents differ in their value functions, discount rates, and/or capacity to drive favorable results. Such differences may lie behind movements such as parents opting out of state-wide testing in the US, as some parents may feel that the measured output does not provide much information about the type of achievement they care about.

student assessments than an otherwise comparable set of low-performing schools. But the results might also just be 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 capacities and be subject to overall political and schooling institutions that already better reflect the desires of parents.

III. International Panel Data

To extract evidence on how test-based information affects student learning, we combine international measures of student achievement with measures of different types of student assessments over a period of 15 years. We describe each of the two components in turn.

A. 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 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 an increasing number of non-OECD countries on a three-year cycle (OECD 2016).[14] 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.[15] 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.

[14] 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.

[15] 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.

In our analyses, we consider student outcomes in all countries that have participated in at least three of the six PISA waves between 2000 and 2015.[16] This yields a sample of 59 countries (35 OECD and 24 non-OECD countries, see Table A1 in the Online Appendix) 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.

PISA student assessments use 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. Testing lasts for up 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.[17] Until 2012, PISA employed paper and pencil tests. In 2015, the testing mode was changed to computer-based testing, a topic we return to in our robustness analysis below.

While the overall test average across all countries was quite stable between 2000 and 2015, achievement moved significantly up in some countries and significantly down in others (see Figure A1 in the Online Appendix). In 14 countries, achievement improved by at least 0.20 s.d. 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 0.20 s.d. in eleven countries (United States,

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

[17] 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.

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. We 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. 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). School-level controls include school size (number of students), community location (five categories), share of fully certified teachers, principals' assessments of the extent to which learning in their school is hindered by teacher absenteeism (four categories), shortage of math teachers, private management, 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. We impute missing values in the student and school background variables by using the respective country-by-wave mean and include a set of indicators for each imputed variable-by-observation.[18]

B. Categories of Testing

We derive our measures of different forms of student testing, consistently measured across countries and time, from a combination of the PISA school background questionnaires, regular

[18] The share of missing values is generally very low for the covariates, see Table A2 in the Online Appendix.

data collection by other sections of the OECD, and data compiled under the auspices of the European Commission. This provides us with 13 separate indicators of testing practices, each measured at the country-by-wave level over the period 2000-2015.[19] We collapse this range of testing specifics into the four categories based in our conceptual framework. Here we summarize the constructed categories; details of questions and sources are provided in the data appendix in the Online Appendix.

Standardized Testing with External Comparison (SCOMP). The first category draws on four separate data sources that identify standardized assessments constructed outside of schools and designed 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-focused external comparison"). It also includes indicators of whether central examinations affect student placement at the lower secondary level (two sources) and whether central exit exams determine student outcomes at the end of secondary school (which, together, we term "student-focused external comparison").[20]

Standardized Testing for Internal Comparison (SINT). The second testing category refers to standardized assessments that do not necessarily provide for or are not primarily motivated by external comparison. Three questions in the PISA survey document the prevalence of different aspects of this type of testing: standardized testing in the tested grade, student tests to monitor teacher practices, and tracking of achievement data by an administrative authority.

[19] Table A3 in the Online Appendix provides an overview of the different underlying assessment indicators. Table A4 in the Online Appendix indicates the number of country observations by wave for each indicator.

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

Internal Reporting (IRPT). This category covers testing used for general pedagogical management including informing parents of student progress, public posting of outcomes, and tracking school outcomes across cohorts. The included measures are derived from three separate PISA questions.

Teacher Monitoring (TMON). This final category covers internal assessments that are directly focused on teachers. It combines schools' use of assessments to judge teacher effectiveness and the monitoring of teacher practice by principals and by external inspectorates, again derived directly from the principal surveys in PISA.

C. Sources of Identifying Variation

Our testing measures combine two distinct but closely related elements. One captures national policies as formally legislated. The other captures the share of schools in a country that actually implement the specific form of testing for their students.[21] The variation in both types of measures is informative in its own right. In a sense, the first type captures the most policy-relevant parameter, as legislating specific policies is what policymakers can ultimately do. By contrast, the second type is informative about effects of any actual usage of the different forms of testing. While we combine the different measures into overall indices in our baseline analysis, in Section VI.B we come back to the distinction and show that both types of measures yield similar results, warranting the combined baseline analysis.

As the cross-sectional variation in testing and achievement between schools in a country is particularly prone to endogeneity biases, we never use this variation for identification in our analysis. Rather we aggregate all measures to the country (by wave) level so that they reflect the

[21] In the data, we cannot distinguish whether a change in the proportion of schools using a particular form of testing reflects the (mostly imperfect) implementation of a national policy or an independent policy initiative at the local level.

average share of schools in a country that use a particular form of testing at any given point in time; see Section IV below for details.

D. Aggregation of Separate Indicators

The original 13 indicators of assessment practices consistently available across countries were aggregated into the four main categories by taking the simple average of the observed indicators in each category (see Appendix A.5 in the Online Appendix). For example, for each country-by-wave cell the variable "Standardized testing with external comparison" (SCOMP) is the simple average of its four component variables, namely "School-focused external comparison," "National standardized exams in lower secondary school," "National tests for career decisions," and "Central exit exams."[22] For each country, the average is taken only across those component variables with data available for the country.

Constructing the aggregate categories serves several purposes. In various instances, the survey items are measuring very similar concepts within the same content area, so that the aggregation acts to reduce measurement error in the individual questions and to limit multicollinearity at the country level (which is key in our identification strategy). For example, as discussed more fully in Appendix A.5 in the Online Appendix, the correlation between the two measures of national standardized exams used in lower secondary school 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

[22] The variables in each category are calculated as proportionate usage in terms of the specific indicators for each country and wave. Thus, each variable measures the average share of schools subject to the indicators summarized in the category. 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 Appendix A.5 in the Online Appendix for details). The fact that imputation is not affecting our results is also shown by their robustness to using only the original (non-imputed) observations for each of the underlying 13 separate indicators (see Table 4 and Table A7 in the Online Appendix).

measures of using assessments to inform parents and 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). 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.[23]

The aggregated testing categories are correlated with each other. For the pooled dataset of country-by-wave observations, correlations range between 0.28 and 0.58 across for the categories (Table A5 in the Online Appendix). They are somewhat lower after taking out country and year fixed effects and vary by category: those between SCOMP and other categories are below 0.2; between SINT and other categories are below 0.3; and between IRPT and TMON are 0.48.

E. Descriptive Statistics

Table 1 provides descriptive statistics for the individual indicators of student testing and for the four combined testing categories. The measures derived from the PISA background questionnaires are shares bounded between 0 and 1, whereas the other testing measures are dummy variables.[24] As is evident, some testing practices 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.

For our estimation, the variation over time within individual countries in the different types of testing is key. Figure 2 shows histograms of the 15-year change in the combined measures of

[23] Note that a number of indicators draw on principals' responses about the use of tests in their own schools. Because the PISA sampling involves different schools in each wave, some random error could be introduced. The aggregation also helps to eliminate this sort of measurement error.

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

the four testing categories for the 38 countries observed in both the first and last PISA waves. The implicit policy changes across student assessments in the sampled countries are clearly substantial and support our estimation strategy based on a country-level panel approach.[25] Importantly, there is also wide variation in the changes in usage of the different forms of student assessments across countries, providing the kind of variation used for identification in our analysis. The policy variation is larger for SCOMP than for the other three categories, leading us to expect higher precision (lower standard errors) of the coefficient estimates for this category.

The increasing use of external assessments is quite evident.[26] For example, the share of schools that are externally compared with student assessments increased by more than 50 percentage points in five countries (Luxembourg, Denmark, Italy, Portugal, and Poland) and by more than 20 percentage points in another 18 countries. In three countries, by contrast, the share decreased by more than 20 percentage points (Tunisia, Costa Rica, and Croatia).

Our interest is how different test-based information relates to student outcomes and does not seek to evaluate specific accountability or incentive policies that may be concurrently or subsequently introduced. Some changes in testing regimes have been directly related to more comprehensive (but quite varied) reforms such as in the case of the NCLB in the United States that included various plans for failing schools (Figlio and Loeb 2011), the 2006 *Folkeskole* Act in Denmark that introduced a stronger focus on assessment including national tests (Shewbridge

[25] The exception in this depiction is internal reporting. However, the reduction in this aggregate measure is fully accounted for by a change in the wording of the questionnaire item on assessments to inform parents, where the word "assessments" was replaced by the word "standardized tests" in the 2015 questionnaire (see Table A3 in the Online Appendix). While the mean of this item hardly changed (from 0.98 to 0.97) between 2000 and 2012, it dropped to 0.64 in 2015. Ignoring the 2015 value, the mean of the combined measure of Internal reporting increased by 0.08 from 2000 to 2012. This example indicates the importance of including year fixed effects in our analyses and of taking particular care in considering the question wording. As we will show below, our qualitative results on internal reporting are unaffected by dropping the year 2015 from the analysis.

[26] Figure A2 in the Online Appendix depicts the evolution of using standardized assessments for school-focused external comparison from 2000 to 2015 for each country.

et al. 2011), and the introduction of standardized national assessments to monitor student outcomes in Luxembourg (Shewbridge et al. 2012). In other cases, it appears that testing programs are introduced independent of any prescribed overall incentive or accountability system such as the 2009 introduction of the *Invalsi* national test in Italy.[27] Still, the newly available testing information then plays into a variety of local uses by schools and parents as it provides feedback to schools and teachers on their students' achievement.

As these measures are derived from survey responses by principals, they reflect the combined effect of external policies and the actual implementation of them at the school level. Thus, for example, the introduction of national assessments in Denmark is not accompanied by a discontinuous jump but by a more gradual implementation path.

IV. Testing the Impact of Information: 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 factors such as culture, the general valuation of educational achievement, or other government institutions may introduce bias.

[27] See Figure A2 in the Online Appendix and the description in https://it.wikipedia.org/wiki/Test_INVALSI.

We address leading concerns of bias in cross-country estimation by formulating a fixed-effects panel model of the following form:

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + T_{ct}\beta + \mu_c + \mu_t + \varepsilon_{ict} \quad (1)$$

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 testing *T*. The parameters μ_c and μ_t are country and year fixed effects, respectively, and ε_{ict} is an individual-level error term.

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. In our main specification, we therefore introduce interaction terms between the testing measures T_{ct} and a country's average achievement level when it first participated in PISA, $\bar{A}_{c0}$:

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + T_{ct}\beta_1 + (T_{ct} \times \bar{A}_{c0})\beta_2 + \mu_c + \mu_t + \varepsilon_{ict} \quad (2)$$

In this specification, the parameters β_2 indicate whether the testing effects vary 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.

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, T_{ct}, is aggregated to the country-by-wave level. This specification avoids bias

from within-country selection of students into schools that use student assessments. Second, we include country fixed effects, μ_c, to address any potential 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. Our parameters of interest (β_1 and β_2) will not be affected by systematic, time-invariant differences across countries.[28] This specification implies that countries that do not change their assessment practices over the observation period will not enter into the estimation of the β's. The model also includes time fixed effects μ_t. These capture any global trends in achievement along with common shocks that affect testing in a specific PISA wave (including any changes in the testing instruments).

The key identifying assumption 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.

There are three main sources of potential bias in this kind of identification. First, there may be secular trends that correlate with the treatment variation. For example, if countries whose achievement is trending downwards for other reasons were more likely to enact testing reforms to counteract the downward trend, estimates of the treatment effects of testing reforms would be

[28] Some recent investigations of scores on international assessments have focused on differential effort levels of students across countries (see, for example, Borghans and Schils 2012; Balart, Oosterveen, and Webbink 2018; Gneezy et al. 2019; Zamarro, Hitt, and Mendez 2019). These differences in noncognitive effects related to our outcome variable of PISA scores would be captured by the country fixed effects as long as they do not interact with the incentives introduced by various applications of testing. Note also that other analysis that experimentally investigated test motivation effects in a short form of the very PISA test employed here did not find significant effects of informational feedback, grading, or performance-contingent financial rewards on intended effort, actual effort, or test performance (Baumert and Demmrich 2001).

biased downwards (and vice versa). In Section VI.A, we will report results of a placebo test that includes leads of the treatment variables to test whether achievement trends before any reform are correlated with reform implementation.

Second, there may be secular shocks that correlate with the timing of treatment. While it is impossible to completely rule out any bias from co-movement of other unobservable factors in this type of identification, one way to explore the potential severity of remaining bias is to explore the extent to which estimates of the treatment effect vary when including different sets of observed controls (Altonji, Elder, and Taber 2005). In Section VI.B, we will therefore test how sensitive results are to relevant covariates.

Third, there may be other policies that are instituted contemporaneously with the treatments. For example, testing reforms may sometimes be implemented together with reforms in schools' autonomy that may have independent effects on student achievement (Hanushek, Link, and Woessmann 2013). In Section VI.B, we test for the robustness of our estimates to the inclusion of controls for changes in school autonomy, as well as in a broad set of additional school features that may be the subject of (or correlated with) other school reforms. In addition, we can control for the various categories of testing reforms simultaneously, meaning that variations in the other types of testing – and any other policies that may correlate with them – are held constant.

For interpretation, we think of our specification as a reduced-form model characterizing the impact of different kinds of performance information on the overall level of learning (A). Information per se does not change student outcomes unless it triggers different behavior from parents, students, and teachers. Any altered behavior could be the result of various specific incentive programs or it could reflect an array of local and family responses to the information. Our purpose, however, is not to trace these different potential mechanisms but to understand the

role of different kinds of assessment information. Sometimes test information is explicitly linked to specific incentives (as with student exit exams), but more generally this is not the case.

Still, the remaining concern then is whether estimated parameters truly pick up the effect of arbitrary bundles of jointly introduced policies or downstream reforms that come quasi automatically with testing, or whether they mainly pick up the effects of specific but unmeasured concomitant or downstream policies. We provide prima facie evidence in Section VI.B that the estimated testing effects are not driven by systematic but correlated policies.

V. Information and Achievement: Basic Results

We begin with estimates of average impacts of various kinds of information across our sampled countries before moving to our main specifications that allow for treatment heterogeneity. 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.[29] Regressions are weighted by student 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.

The average impacts in Table 2 suggest that different information from the four categories of student testing have very different effects on student achievement. Among the four assessment categories, only changes in standardized testing that is used for external comparisons (SCOMP) have a strong and statistically significant positive relationship with changes in student outcomes on average. The coefficients on SINT and IRPT are insignificant and close to zero, whereas there

[29] Table A2 in the Online Appendix shows the coefficients on all control variables for the specification of the first column in Table 3. The estimates for control variables are quite consistent across specifications.

is quite a sizeable negative coefficient on TMON.[30] This pattern of impacts is consistent with the predictions on differing strengths of potential incentives from the conceptual discussion. The point estimate for SCOMP suggests that a change from no usage to complete standardized external comparison is related to an increase in math achievement by more than one quarter of a standard deviation.[31] Results for science and reading achievement are very similar to those for math, although the negative coefficients on TMON are not statistically significant (columns 2 and 3).[32]

However, these average effects mask clear heterogeneity by each country's initial level of achievement. The first three columns of Table 3 present estimates of the interacted model of equation (2) for the three subjects. 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.

There is clear evidence that the marginal value of more information declines as the overall performance of a nation's schools is better. The exception comes from the uniform insignificance of differences in internal reporting (IRPT) for overall student outcomes. While the drop in impact varies across different categories of testing, the greater value of increased information for low-achieving countries is very evident.

[30] Note that, consistent with the larger within-country variation of SCOMP over time documented in Section III.E, the standard error associated with this coefficient estimate is smaller. Still, even with the smaller standard error of this variable, the coefficient estimates on SINT and IRPT would be far from statistical significance.

[31] The point estimates and the statistical significance of the category impacts are very similar (except that the coefficient on TMON is slightly lower at -23.5 and not significant) when each of the four testing categories is included individually (not shown), indicating that there is enough independent variation in the different testing categories for estimation and that the effect of SCOMP does not reflect reforms in other assessment categories.

[32] The hypothesis that the effect of SCOMP is the same as the effects of the other three testing categories is jointly strongly rejected in each of the three subjects. Individually, the coefficient on SCOMP is significantly different from SINT in math and reading, from IRPT in reading, and from TMON in all three subjects.

First, the impact of SCOMP 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 0.37 s.d. in math. With each 10 initial PISA points, this effect is reduced by 0.025 s.d. At 500 PISA points (the OECD mean), the effect of standardized external comparison is still statistically significantly positive at around 0.13 s.d. in all three subjects.

Second, SINT shows a similar pattern. It creates significantly positive impact in initially low-achieving countries. However, effects disappear at higher achievement levels, i.e., for countries with initial scores of roughly above 490 in all subjects. In fact, in contrast to SCOMP, at very high levels of initial country experience the effect of SINT turns significantly negative.

Third, the estimates for TMON also show a declining pattern with initial country scores in math. They are insignificant for most of the initial-achievement distribution but also turn significantly negative at high levels of initial achievement.

The category of SCOMP actually aggregates two quite distinct components – that related to schools and that related to students. Information permitting comparisons of schools to district or national performance puts a spotlight school performance and potentially has its greatest effect on administrators and teachers. The three measures of testing to determine school and career placement decisions for students on the other hand moves the focus to the students themselves. [33]

School-focused comparisons follow a similar heterogeneous pattern as overall SCOMP but go to zero for a somewhat larger set of initially high-achieving countries (columns 4-6 of Table 3). By contrast, the positive impact of student-focused external comparisons does not vary

[33] The measure of student-focused external comparison we use is the simple average of the three underlying indicators of SCOMP except for the one on school-focused external comparison. Note that the estimates of Table 3 are based on smaller student samples from fewer countries, because data on student-focused external comparison are available almost exclusively for OECD and European Union countries.

significantly with initial achievement levels.[34] The results emphasize that focusing information on different actors encourages different responses and leads to separate effects on outcomes.

To establish that our aggregation is not suppressing important heterogeneity within the separate categories, Table 4 presents individual results for each of the 13 underlying country-level indicators of student assessment, where each combination of main and interacted effect represents a separate regression.[35] The disaggregated underlying individual indicators of SCOMP consistently show the pattern of significantly stronger effects in initially poorly performing countries.[36]

Similarly, all three underlying indicators of SINT show the pattern of significant positive effects at low levels of achievement and significantly decreasing effects with initial achievement. Thus, the positive 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 reporting. This finding supports the World Bank attention to testing for low-achieving countries (World Bank 2018). None of the three indicators of IRPT show a significant impact pattern.[37]

[34] Estimates of the specification without interactions with the initial achievement level in Table A6 in the Online Appendix show that the average impact of both school and student assessments is strongly positive and statistically significant, with estimates for the school-focused testing being somewhat larger than for the individual student testing.

[35] The separate regressions of Table 4 do not employ any imputation of the separate treatment variables. Thus, the numbers of countries and waves included in each estimation (reported in columns 5 and 6) vary and are determined by the availability of the specific testing indicator. The fact that these results confirm the previous results of the four combined categories shows that the latter are not driven by the aggregation procedure or by the interpolated imputations required for the aggregation of the separate indicators. Estimates of the separate indicators for the specification without interactions with the initial achievement level are shown in Table A7 in the Online Appendix.

[36] 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. The negative interaction effect for central exit exams reaches statistical significance only in science.

[37] An interesting outlier in the individual-indicator analysis are assessments to inform parents, which show the opposite type of heterogeneity (significantly so in math and science): The expansion of 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

The disaggregated components of TMON provide suggestive heterogeneous impacts. The negative interaction effect of TMON with initial achievement is driven by the two subjective components – monitoring by the school principal and by external inspectorates. On the other hand, negative marginal impacts are not apparent when teacher effectiveness is judged objectively by assessments.

VI. Specification Tests

Our fixed-effects panel model identifies the effect of assessment policies on student achievement from usage changes in testing within countries over time. The specification tests here provide additional information about the validity of the underlying identifying assumptions.

A. A Placebo Test with Leads of the Testing Variables

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. Here the largest concern is that countries that are on a downward trend turn to expanded testing to reform the system. Note that, if generally true, this would tend to bias our estimated effects downward.

Our panel structure 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. 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. As such a

is a significantly positive effect on science achievement of 0.37 s.d. 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.

specification is very demanding in short panels, we first implement it for the restricted model with average effects.

As is evident in Table 5, none of the lead variables of the four testing categories is significantly related to student achievement (i.e., in the wave before reform implementation).[38] At the same time, the results of the contemporaneous testing measures are fully robust to conditioning on the lead variables: SCOMP 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 yet implemented. Moreover, the estimated coefficients for the testing categories are qualitatively similar to those in Table 2.[39]

The fact that the leads of the testing variables are insignificant also indicates that prior 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 testing categories and their leads interacted with initial achievement is overly demanding to the data. Nevertheless, an interacted model just for SCOMP gives confirmatory results: SCOMP 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).[40]

[38] The coefficients on the lead variables are somewhat imprecisely estimated. However, in models with leads for just SCOMP, the lead coefficient is statistically significantly different from the base coefficient at the 5 percent level in science, at the 10 percent level in reading, and at the 20 percent level in math.

[39] By construction, the placebo regression with leads excludes the 2015 PISA data, so the most direct comparison would be the baseline model without the 2015 wave. As indicated in Table 9 below, results are very similar in that specification.

[40] Note that including a policy lag does not provide a similar placebo test because a lagged testing policy may in fact partly capture the effect of previously implemented reforms whose effects have not been completely felt. In a specification that includes the contemporaneous, lead, and lagged variable, both the contemporaneous and the lag of the SCOMP variable are statistically significant while the lead remains insignificant (not shown)

There is no evidence of the introduction of different testing regimes in response to prior educational circumstances. At the same time, it is clearly difficult to estimate time patterns reliably given at most six time-series observations for each country. Thus, while highly suggestive, definitive testing of the key identifying assumptions such as common trends across countries is not possible.[41]

B. Coincidental Other Policies, Long Differences, and other Specification Tests

Another possible concern is that countries may introduce other policies coincidentally with the use of alternative testing policies. Although we cannot consider all such potential policy changes, we can directly analyze what is the most likely synchronized policy – expanded local autonomy in school decision making. Local schools have greater knowledge both of the demands they face and of their own capacities, making them attractive places for much decision making. But for just the reasons discussed in the conceptual model, with asymmetric information about their actions and results, they might not operate in an optimal way from the viewpoint of either the higher-level policy makers or even of the parents – suggesting that information on student outcomes could be generally useful in any moves toward more autonomy in decision making.

All of our estimation includes information on the time pattern of autonomy reforms for each country. Consistent with prior work (Hanushek, Link, and Woessmann 2013), our results confirm that the effect of school autonomy on student achievement is negative in developing countries but positive in developed countries.[42] Importantly, the results on assessment effects are

[41] E.g., adding a linear time trend for each country renders coefficients too imprecise for clear inference.

[42] With six rather than four PISA waves and with 303 rather than 155 country-by-wave observations, we show here that the previous results about autonomy are also robust to the consideration of the effects of student assessment reforms (see Table A2 in the Online Appendix).

not confounded by the potentially coincidental introduction of policies that alter school decision making and autonomy.

In fact, the simultaneous introduction of comparative testing and school autonomy is an exception rather than the rule when considering whether testing reforms correlate with the introduction of other school policies. Table A8 in the Online Appendix shows correlations (taking out country and year fixed effects to reflect the identifying variation in our model) of the four categories of testing policies with measures of other school policies that are consistently available across countries and waves – namely school autonomy, school size, teacher certification, shortage of teachers, private vs. public school management, and share of government funding of schools. The results indicate that there is indeed a small but statistically significant correlation between standardized external testing and school autonomy of 0.16. An additional small but statistically significant correlation of 0.21 exists between the introduction of teacher monitoring and reported shortage of math teachers at school, which may be an indication that teacher monitoring reduces the supply of teachers. All other correlations of the four categories of testing with the other policies are even smaller, and only very few reach statistical significance. These results reduce concerns of bias from other common correlated policies.

As a further indication against the potential concern that other contemporaneous correlated policy changes might affect our results, note that results do not change when the four different testing categories are entered individually or jointly. That is, other forms of testing – and their potentially coinciding other policy changes – are controlled for in the simultaneous model. Only other policies that are coincidental just with the specific form of testing and not with the other ones could potentially still introduce bias. Furthermore, all models control for several time-varying school features including the schools' share of government funding, private/public

management, and size. The school-level covariates also include several variables related to teachers – the share of fully certified teachers, teacher absenteeism, and shortage of math teachers. Contemporaneous policy reforms in these school features are thus also controlled for.

In fact, some of the school-level variables – in particular, those capturing the composition of teachers – could potentially be endogenous to the testing reforms. However, Table 6 shows that qualitative results are unaffected by leaving the teacher controls out of the model (column 1).[43]

Another approach to gauge the potential likelihood of unobserved factors to affect our results is to look at the extent to which the inclusion of the entire set of observed factors changes our estimates. Dropping all covariates from the model does not change the qualitative results (column 2). This invariance holds despite the fact that the explained variance of the model increases substantially by the inclusion of the control variables, from 0.256 to 0.391. The fact that testing results are insensitive to the included set of relevant covariates reduces concerns that our estimates are strongly affected by any omitted variable bias from unobserved characteristics (in the sense of Altonji, Elder, and Taber 2005).

Analysis of effect heterogeneity across countries can also provide some indication on the interpretation of results. Apart from the significant interactions with initial achievement levels contained in our main model, we do not find evidence of consistent heterogeneities in several other environmental dimensions (not shown). In particular, the effects of the four testing categories do not significantly interact with countries' initial level of GDP per capita (contrasting with the heterogeneous effects found for school autonomy in Hanushek, Link, and Woessmann (2013)). Similarly, there are no significant interactions of the testing categories with the level of school autonomy in a country. In addition, SCOMP does not significantly interact with the other

[43] The same is true for achievement in science and reading (not shown). See Table A9 in the Online Appendix for the model with average effects.

three categories of student assessments. In a sense, this lack of heterogeneity is confirmation that our treatment effects do in fact capture consistent reduced-form effects of providing testing information, rather than effects of various concomitant or downstream policies that differ across countries.

Our fixed-effects panel model is identified from changes that occur from one PISA wave to the next, i.e., from three-year changes. This strategy has the advantages of incorporating several changes per country and implicitly of allowing for differences in treatment dosages. The disadvantages are that any measurement error is amplified in the first-differenced changes and that any impact of testing may take time to emerge fully. By restricting identification to changes across all sample periods, we can both reduce the potential influence of measurement error and gauge the long-run relevance of the policy reforms.

We estimate our models in long differences that consider just the total 15-year change from the first to the last PISA wave. The main findings of our model with average effects, shown in column 3 of Table A9 in the Online Appendix, are robust in this long-difference specification.[44] Consistent with larger measurement error in shorter-frequency change data, the estimate of the positive average effect of SCOMP is larger when considering only long-run changes. The estimates of average effects of the other three testing categories remain insignificant. While obviously less precise, the pattern of heterogeneity by initial achievement is also evident in the long-difference specification when the analysis is restricted to the category of SCOMP (column 4 of Table 6).[45]

[44] This is the model depicted graphically for SCOMP in Figure 1 in the introduction.

[45] Similarly, a model restricted to the category of SINT yields a significantly positive main effect and a significantly negative interaction (not shown).

The fact that the long-difference estimate of the average effect of SCOMP is larger than the higher-frequency panel estimate may also indicate that the treatment effect grows over time. A model identified from three waves that imply differences of six to nine years yields estimates that are between these two estimates. Furthermore, considering a lag (and lead) in addition to the contemporaneous reform variable in a model restricted to SCOMP, both the contemporaneous and the lagged reform variable (but not the lead) enter the model significantly positive. The estimated coefficients (not shown) suggest that introducing SCOMP has a positive effect of 0.12 s.d. after one period (three years) which increases to 0.35 s.d. after two periods (six years).

There is a difference between legislated testing reforms and the actual implementation of testing in schools. The latter is particularly relevant for understanding the impacts of actual testing usage (i.e., treatment on the treated), whereas the former may carry particular interest from a policy perspective. As discussed in Section III.E, the implementation path of test usage may be more gradual than any formal policy reform at the national level. Most of our testing measures are derived from reports of school principals on the use of testing in their schools, measured as the country share of schools using the specific testing application. But some are also dummy measures based on dichotomous coding of whether a country has formally legislated a specific testing policy or not, representing partial but well-measured policy changes. In particular, the separate OECD measure of national standardized testing represents coding by country specialists of the changes in assessment policies – i.e., the kinds of accurately observed policy changes that would enter into micro policy evaluations.

While – for the reasons discussed in Section III.D – we prefer the combined testing measures in our baseline specification, it is important to note that consideration of just the dummy measure of SCOMP provides significant estimates of both the main effect on student

performance and of the interaction with the initial achievement score (see second line in Table 4).[46] Thus, the more gradual measure of usage of external comparison in schools and the discontinuous reform indicator of formal national policies yield very similar results, indicating that our results do not depend on adopting one of the specific perspectives.

To check that the negative effects of SINT and TMON at high levels of initial achievement (indicated in Figure 3) are not simply an artefact of the imposed linearity of the interaction model, columns 5-8 of Table 6 report results of a specification that interacts each of the four testing categories with four dummies reflecting the four quartiles of initial country achievement. There is no indication of strong nonlinearity.[47] In particular, the negative effects at high levels of initial achievement are also visible in this specification, indicating that they are not driven by the imposition of linearity. Introducing SINT and TMON in systems that are already performing at a high level may in fact distract teacher attention from more productive forms of instruction.

VII. Robustness

Our results are robust to a number of potentially contaminating factors. In particular, we consider possible peculiarities of the set of countries in our sample, possible effects of student and school exclusions from PISA testing, possible interactions with changes in PISA testing, and alternative estimation approaches. For ease of exposition, we present robustness results with

[46] The Eurydice measure of national testing is also an expert-based dummy measure of national testing. While its average effect on achievement is also significantly positive (see third line of Table A7 in the Online Appendix), the negative interaction with the initial score does not capture statistical significance, likely reflecting the limited initial achievement range captured by this measure which is available for 18 European countries only.

[47] TMON also has a rather steady pattern when entered without the other three testing categories (92.3, -3.7, -36.6, and -102.5), suggesting that the joint specification with four interactions of four testing measures may be rather demanding to depict precise patterns. The separately estimated patterns for the other three measures also indicate rather linear relationships (not shown).

heterogeneity by country achievement level in Table 7 in the text and the results for averages without heterogeneity, which yield similar conclusions, in Table A10 in the Online Appendix.

To ensure that our results are not driven by the peculiarity of any specific country, we re-estimated all of our main models (columns 1-3 of Tables 2 and 3) 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 by level of development, we split the sample into OECD and non-OECD countries. Qualitative results on the average effects are similar in the two subgroups of countries, although the positive effect of SCOMP is larger in OECD countries (columns 1-2 of Table A10 in the Online Appendix). Patterns of heterogeneity by achievement level are less precisely identified within the two more homogeneous subgroups (Table 7). In the OECD countries, the significant effect of SCOMP 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 SCOMP is significant, and the heterogeneous effect of SINT is somewhat more pronounced in OECD countries. But overall, the patterns do not differ substantively between the two country groups.

While PISA has stringent sampling standards, there is some variation across 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, column 4 in Table 7 (and column 3 in Table A10 in the Online Appendix) controls for the

country-by-wave exclusion rates reported in each PISA wave. As is evident, results are hardly affected.

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 in 2015 affect our results.

First, we drop the 2015 wave from our regressions. As is evident from column 5 in Table 7 (and column 4 in Table A10 in the Online Appendix), 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 combination of changes in PISA testing.

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 column of Table A10 in the Online Appendix, we run our model with average effects using 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.

Third, 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 by students and schools in 2012 that we derive from the PISA background questionnaires. As indicated by Table A11 in the Online Appendix, 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.

Finally, while we estimate all models at the individual student level, the main treatment varies only at the country-by-wave level. An alternative way of estimating our model is thus a two-stage estimation. The first stage is a student-level estimation that regresses test scores on all control variables. After collapsing the residuals of this first-stage estimation to the country-by-wave level, the second stage is a standard panel model that regresses these collapsed residuals on the testing variables, including country and wave fixed effects. Tables A12 and A13 in the Online Appendix show that this two-stage model yields quantitatively very similar results to our main model.[48]

VIII. 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 express the view that high-stakes tests that enter into reward and incentive systems for some individuals are inappropriate

[48] The same qualitative results also emerge when collapsing the original test scores (without residualizing) to the country-by-wave level (not shown), consistent with the insensitivity of our student-level results to the inclusion of controls (see Table 6 and Table A9 in the Online Appendix).

(Koretz 2017). Others argue that increased use of testing is 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 forms of testing and among alternative country environments. Furthermore, 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 counterfactuals.

Our analysis turns to international comparisons to address the key questions of which forms of student testing appear to induce changes that promote higher achievement. The conceptual framework behind the empirical analysis is a principal-agent model that motivates focusing on the strength of potential policies built on the assessment information generated by different forms of testing. The empirical analysis employs international student achievement data to identify the consequential implications of national testing.[49] Specifically, the six waves of the PISA test between 2000 and 2015 permit country-level panel estimation that relies on within-country over-time analysis of country changes in testing practices. We combine data across 59 countries to estimate how varying testing situations and applications affect student outcomes.

Focusing on international comparisons has both advantages and costs. A variety of testing policies that are introduced at the national level cannot be adequately evaluated within individual countries, but moving to cross-country evaluations requires dealing with a range of other

[49] 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).

possible influences on student outcomes. Some issues of measurement error, imprecise wording of questionnaire responses, and, most importantly, other concurrent policies are clearly difficult to address with complete certainty. But the richness of the existing data permits a variety of specification and robustness tests designed to illuminate the potential severity of the most significant issues of coincidental policies or programs.

There are two consistent results from this investigation. First, standardized assessments that provide systematic information supporting comparisons of performance across schools and students (SCOMP and SINT) engenders behavioral responses that improve performance. Second, such information is most valuable for educational systems that are not performing well but – if not used for external comparison – may even be harmful in the highest performing systems. In low-achieving and medium-achieving countries, standardized testing appears to allow for better incentives for performance and for rewarding those who are contributing most to educational improvement efforts. By contrast, it appears that systems that are showing strong results know more about how to boost student performance and are less in need of additional information and accountability systems. While the overall evidence is not as strong, a similar pattern of potential adverse effects of teacher monitoring efforts also appears possible in the highest performing countries. Quite generally, however, systems relying on localized or subjective information that cannot be readily compared across schools and classrooms have little overall impact on student achievement.

References

Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *Quarterly Journal of Economics* 126(2): 699-748.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1): 151-184.

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(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).

Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99(4): 1384-1414.

Balart, Pau, Matthijs Oosterveen, and Dinand Webbink. 2018. "Test Scores, Noncognitive Skills and Economic Growth." *Economics of Education Review* 63: 134-153.

Baumert, Jürgen, and Anke Demmrich. 2001. "Test Motivation in the Assessment of Student Skills: The Effects of Incentives on Motivation and Performance." *European Journal of Psychology of Education* 16(3): 441-462.

Bergbauer, Annika B., Eric A. Hanushek, and Ludger Woessmann. 2018. "Testing." NBER Working Paper 24836. Cambridge, MA: National Bureau of Economic Research.

_____. 2021. "Replication Data." Harvard Dataverse. https://doi.org/10.7910/DVN/BUID9K.

Bergman, Peter. 2021. "Parent-Child Information Frictions and Human Capital Investment: Evidence from a Field Experiment." *Journal of Political Economy* 129(1): 286-322.

Bergman, Peter, and Eric W. Chan. 2021. "Leveraging Parents through Low-Cost Technology: The Impact of High-Frequency Information on Student Achievement." *Journal of Human Resources* 56(1): 125-158.

Bettinger, Eric P. 2012. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." *Review of Economics and Statistics* 94(3): 686-698.

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

Bishop, John H., and Ludger Woessmann. 2004. "Institutional Effects in a Simple Model of Educational Production." *Education Economics* 12(1): 17-38.

Borghans, Lex, and Trudie Schils. 2012. The Leaning Tower of Pisa: Decomposing Achievement Test Scores into Cognitive and Noncognitive Components. Mimeo.

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*, ed. Orley Ashenfelter and David Card, 1801-1863. Amsterdam: North-Holland.

De Fraja, Gianni, Tania Oliveira, and Luisa Zanchi. 2010. "Must Try Harder: Evaluating the Role of Effort in Educational Attainment." *Review of Economics and Statistics* 92(3): 577-597.

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(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(5): 848-862.

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

Figlio, David N., and Joshua Winicki. 2005. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89(2-3): 381-394.

Fryer, Roland G. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126(4): 1755-1798.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2(3): 205-227.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu. 2019. "Measuring Success in Education: The Role of Effort on the Test Itself." *American Economic Review: Insights* 1(3): 291-308.

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(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*, ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 89-200. Amsterdam: North Holland.

_____. 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(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(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(5): 1134-1155.

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(4): 339-355.

Koretz, Daniel. 2017. *The Testing Charade: Pretending to Make Schools Better*. Chicago: University of Chicago Press.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics* 91(3): 437-456.

Laffont, Jean-Jacques, and David Martimort. 2002. *The Theory of Incentives: The Principal-agent Model*. Princeton, NJ: Princeton University Press.

Lavy, Victor. 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics." *American Economic Review* 99(5): 1979-2011.

_____. 2015. "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries." *Economic Journal* 125(588): F397-F424.

Luedemann, Elke. 2011. "Intended and Unintended Short-run Effects of the Introduction of Central Exit Exams: Evidence from Germany." In Elke Luedemann, *Schooling and the Formation of Cognitive and Non-cognitive Outcomes*. ifo Beiträge zur Wirtschaftsforschung 39. Munich: ifo Institute.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1): 39-77.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-based Accountability." *Review of Economics and Statistics* 92(2): 263-283.

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. 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(3): 344-364.

Reback, Randall. 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics* 92(5-6): 1394-1415.

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(3): 207-241.

Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-run Impacts of Accountability on School Quality." *American Economic Journal: Economic Policy* 2(4): 119-147.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. "Feeling the Florida Heat? How Low-performing Schools Respond to Voucher and Accountability Pressure." *American Economic Journal: Economic Policy* 5(2): 251-281.

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.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485): F3-33.

Woessmann, Ludger. 2003. "Schooling Resources, Educational Institutions, and Student Performance: The International Evidence." *Oxford Bulletin of Economics and Statistics* 65(2): 117-170.

_____. 2016. "The Importance of School Systems: Evidence from International Differences in Student Achievement." *Journal of Economic Perspectives* 30(3): 3-32.

_____. 2018. "Central Exit Exams Improve Student Outcomes." *IZA World of Labor* 2018: 419.

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.

York, Benjamin N., Susanna Loeb, and Christopher Doss. 2019. "One Step at a Time: The Effects of an Early Literacy Text-Messaging Program for Parents of Preschoolers." *Journal of Human Resources* 54(3): 537-566.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez. 2019. "When Students Don't Care: Reexamining International Differences in Achievement and Student Effort." *Journal of Human Capital* 13(4): 519-552.

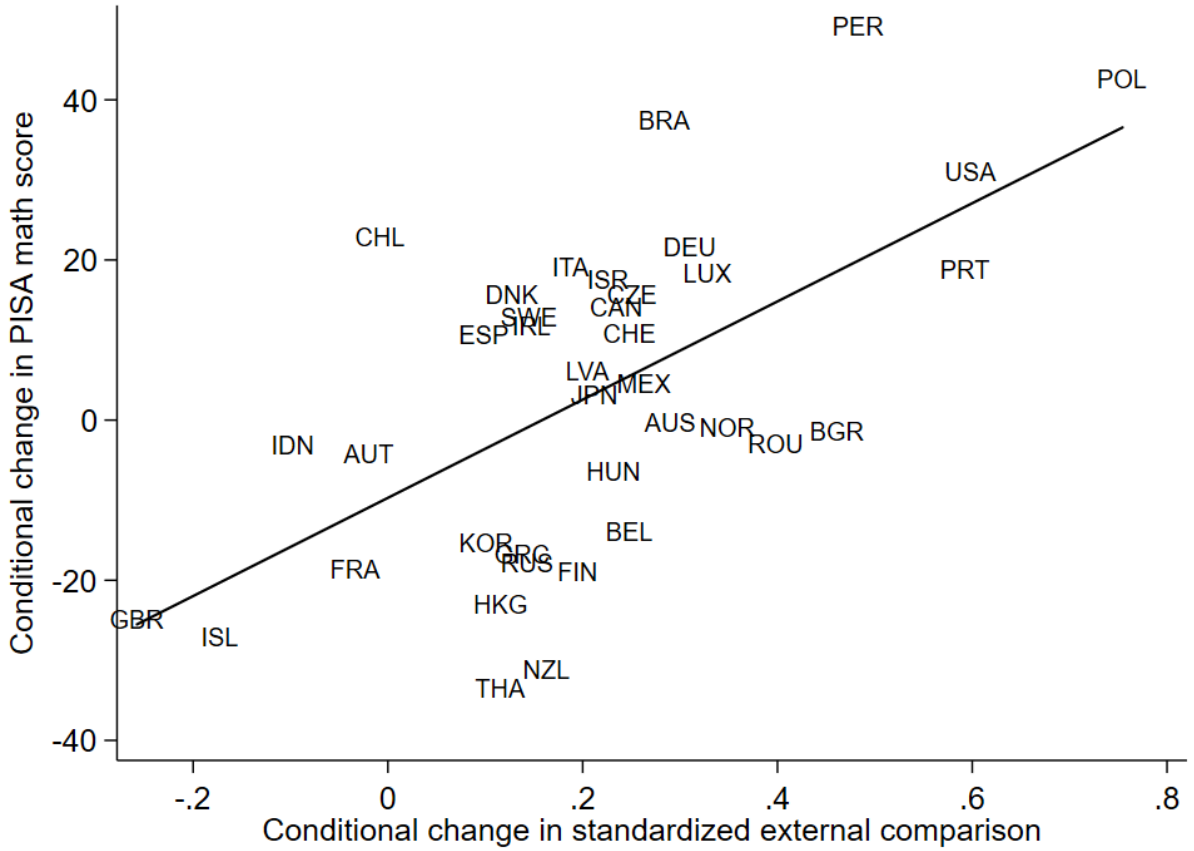

Figure 1

Fifteen-year changes in standardized external comparison and in student achievement

Notes: Added-variable plot of the change in countries' average PISA math score between 2000 and 2015 against the change in the prevalence of standardized testing with external comparison (SCOMP), both conditional on a rich set of student, school, and country controls, based on a long-difference fixed-effect panel model estimated at the individual student level. Mean of unconditional change added to each axis. See column 3 of Table A9 in the Online Appendix for underlying model.

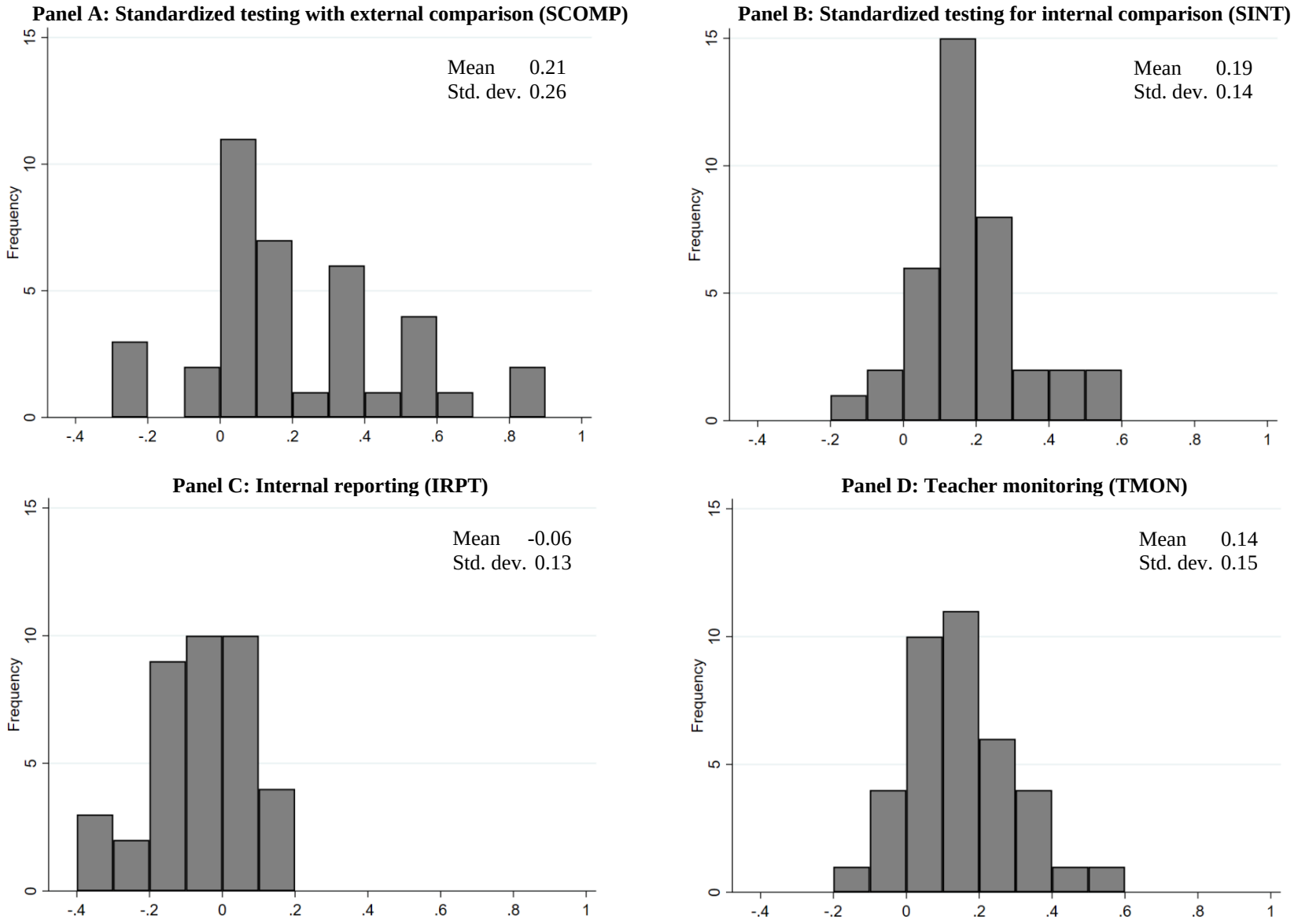

Figure 2

Histograms of change in four categories of student testing, 2000-2015

Notes: Histograms of change between 2000 and 2015 in the four combined measures of student assessment for the 38 countries observed both in the first and last PISA waves.

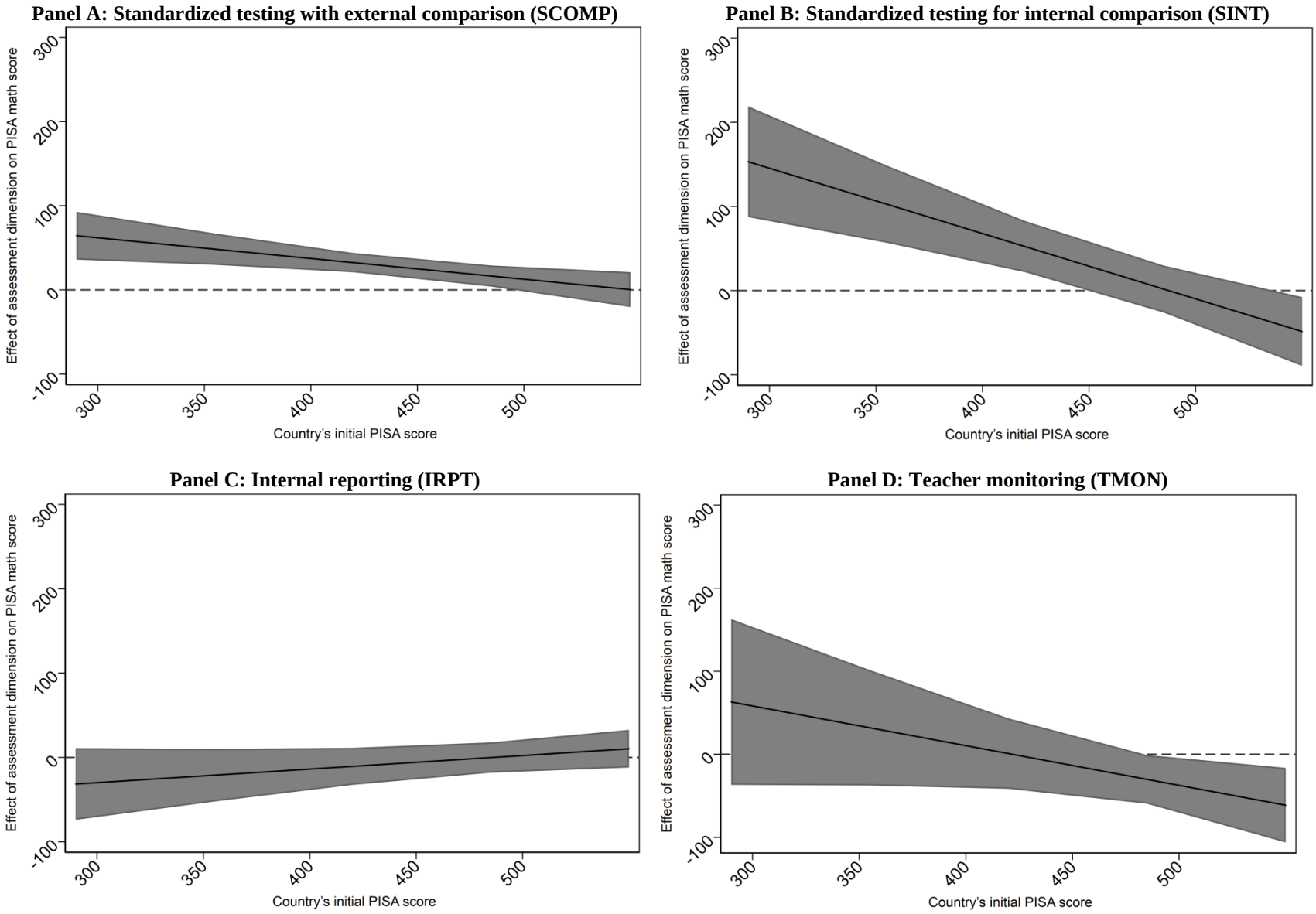

Figure 3

Effect of student testing 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 3 for underlying model.

Table 1

Descriptive statistics of testing measures

	Mean (1)	Std. dev. (2)	Min (3)	Max (4)	Countries (5)	Waves (6)
Standardized testing with external comparison (SCOMP)	**0.518**	**0.271**	**0.022**	**0.978**	**59**	**6**
School-focused 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 for career decisions	0.601	0.481	0	1	18	6
Central exit exams	0.689	0.442	0	1	30	6
Standardized testing for internal comparison (SINT)	**0.714**	**0.160**	**0.219**	**0.996**	**59**	**6**
Standardized testing in tested grade	0.721	0.233	0	1	59	4
Student tests to monitor teacher practice	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 reporting (IRPT)	**0.684**	**0.147**	**0.216**	**0.963**	**59**	**6**
Assessments to inform parents	0.892	0.185	0.141	1	59	5
Assessments 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
Teacher monitoring (TMON)	**0.553**	**0.216**	**0.026**	**0.971**	**59**	**6**
Teacher effectiveness judged by assessments	0.532	0.261	0	0.992	59	5
Teacher practice monitored by principal	0.773	0.262	0.049	1	59	4
Teacher practice monitored by external inspectors	0.402	0.255	0.006	0.994	59	4

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

Table 2

The average effect of different forms of student testing on student achievement: Fixed-effects panel models

	Math	Science	Reading
	(1)	(2)	(3)
Standardized testing with external comparison (SCOMP)	28.811***	23.282***	28.424***
	(6.126)	*(6.144)*	*(5.911)*
Standardized testing for internal comparison (SINT)	-5.469	1.252	-2.036
	(14.062)	*(13.950)*	*(13.148)*
Internal reporting (IRPT)	7.491	17.669	-12.660
	(11.646)	*(13.155)*	*(14.736)*
Teacher monitoring (TMON)	-35.850**	-27.549*	-25.358
	(15.680)	*(14.226)*	*(15.835)*
Control variables	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Student observations	2,094,856	2,094,705	2,187,415
Country observations	59	59	59
Country-by-wave observations	303	303	303
R^2	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 3

Effects of student testing by initial achievement level: Fixed-effects panel models

	Math (1)	Science (2)	Reading (3)	Math (4)	Science (5)	Reading (6)
Standardized testing with external comparison (SCOMP)	37.304***	28.680***	47.977***			
	(6.530)	*(8.222)*	*(9.005)*			
× initial score	-0.246***	-0.149	-0.345***			
	(0.085)	*(0.101)*	*(0.113)*			
Standardized testing for internal comparison (SINT)	67.772***	86.860***	88.701***	72.689***	77.183**	116.503***
	(17.139)	*(20.263)*	*(21.396)*	*(26.701)*	*(34.691)*	*(31.505)*
× 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 reporting (IRPT)	-13.858	-14.734	-26.214	-14.462	-0.669	-44.234
	(12.216)	*(15.155)*	*(17.261)*	*(21.562)*	*(35.177)*	*(33.433)*
× 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)*
Teacher monitoring (TMON)	10.432	18.210	-22.463	-0.620	2.077	-42.345
	(25.005)	*(25.338)*	*(32.946)*	*(32.969)*	*(42.956)*	*(43.058)*
× 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)*
School-focused external comparison				45.740***	39.343*	49.581**
				(15.067)	*(21.244)*	*(21.699)*
× initial score				-0.385**	-0.347	-0.361
				(0.165)	*(0.229)*	*(0.248)*
Student-focused external comparison				15.138**	7.120	2.535
				(6.518)	*(10.564)*	*(5.975)*
× initial score				-0.019	0.079	0.147
				(0.105)	*(0.160)*	*(0.091)*
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 2 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 4

Estimations for separate underlying testing indicators

	Math		Science		Reading				
	Main effect (1a)	× initial score (1b)	Main effect (2a)	× initial score (2b)	Main effect (3a)	× initial score (3b)	Obser-vations (4)	Coun-tries (5)	Waves (6)
Standardized testing with external comparison (SCOMP)									
School-focused external comparison	39.945***	-0.456***	43.605***	-0.484***	47.018***	-0.481***	1,703,142	59	5
	(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	1,517,693	36	6
	(18.887)	*(0.206)*	*(13.905)*	*(0.162)*	*(31.246)*	*(0.301)*			
National tests for career decisions	21.890***	-0.081	11.309	-0.002	20.983**	-0.119	676,732	21	6
	(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	1,141,162	30	6
	(31.796)	*(0.322)*	*(18.255)*	*(0.156)*	*(46.933)*	*(0.543)*			
Standardized testing for internal comparison (SINT)									
Standardized testing in tested grade	46.491***	-0.460***	42.679***	-0.427***	54.278***	-0.509***	1,198,463	59	4
	(9.608)	*(0.108)*	*(9.829)*	*(0.105)*	*(9.918)*	*(0.104)*			
Student tests to monitor teacher practice	15.863	-0.384***	44.530***	-0.508***	25.154*	-0.391***	1,537,802	59	4
	(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**	1,713,976	59	4
	(14.631)	*(0.129)*	*(18.191)*	*(0.198)*	*(19.113)*	*(0.242)*			
Internal reporting (IRPT)									
Assessments to inform parents	-8.895	0.233***	-10.140	0.314***	-6.900	0.151	1,705,602	59	5
	(6.714)	*(0.047)*	*(8.012)*	*(0.079)*	*(10.352)*	*(0.103)*			
Assessments to monitor school progress	6.106	-0.065	2.356	0.065	6.433	-0.115	1,705,602	59	5
	(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	1,713,976	59	4
	(15.782)	*(0.133)*	*(15.355)*	*(0.144)*	*(19.472)*	*(0.236)*			
Teacher monitoring (TMON)									
Teacher effectiveness judged by assessments	0.387	-0.063	0.220	0.037	1.141	-0.043	1,705,602	59	5
	(14.989)	*(0.153)*	*(16.015)*	*(0.202)*	*(14.510)*	*(0.163)*			
Teacher practice monitored by principal	0.807	-0.239	31.735	-0.514**	1.358	-0.186	1,588,962	59	4
	(26.483)	*(0.208)*	*(21.136)*	*(0.201)*	*(20.928)*	*(0.222)*			
Teacher practice monitored by external inspectors	18.086	-0.370**	17.783	-0.365*	-6.485	-0.134	1,588,962	59	4
	(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 "× 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 2 for the included control variables. Observations 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 5

Placebo test with leads of testing reforms

	Math	Science	Reading
	(1)	(2)	(3)
Standardized testing with external comparison (SCOMP)	25.104***	24.567***	27.787***
	(6.316)	*(5.242)*	*(7.501)*
Standardized testing for internal comparison (SINT)	-16.172	-3.734	4.660
	(18.139)	*(19.288)*	*(18.490)*
Internal reporting (IRPT)	14.305	19.522	-17.675
	(15.367)	*(21.238)*	*(20.325)*
Teacher monitoring (TMON)	-35.785	-38.797*	-31.560
	(22.833)	*(19.796)*	*(19.079)*
Lead (SCOMP)	12.119	4.475	5.746
	(11.045)	*(8.506)*	*(9.351)*
Lead (SINT)	-15.195	-11.138	-17.220
	(13.881)	*(16.216)*	*(19.718)*
Lead (IRPT)	6.965	-7.014	5.567
	(14.408)	*(15.286)*	*(14.069)*
Lead (TMON)	-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 testing 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 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 6

Specification tests

	No teacher controls	No controls	Long difference (2000+2015 only)		Interactions with four quartiles of initial score			
					×Q1	×Q2	×Q3	×Q4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standardized testing with external comparison (SCOMP)	37.340***	53.124***	18.944	69.060***	55.899***	26.505***	9.208	18.278
	(5.986)	*(11.586)*	*(24.016)*	*(17.063)*	*(16.514)*	*(7.515)*	*(11.065)*	*(13.847)*
× initial score	-0.249***	-0.440***	0.211	-0.272				
	(0.080)	*(0.144)*	*(0.222)*	*(0.187)*				
Standardized testing for internal comparison (SINT)	74.378***	54.154***	42.848		60.373**	31.831*	-15.650	-67.691**
	(18.061)	*(17.107)*	*(31.020)*		*(26.276)*	*(17.614)*	*(18.383)*	*(27.897)*
× initial score	-0.845***	-0.525***	-0.510					
	(0.183)	*(0.166)*	*(0.335)*					
Internal reporting (IRPT)	-10.574	-13.016	-106.185**		-25.596	-11.618	0.771	19.721
	(12.230)	*(14.113)*	*(45.672)*		*(21.609)*	*(13.145)*	*(12.970)*	*(15.521)*
× initial score	0.157	0.166	1.119**					
	(0.097)	*(0.121)*	*(0.473)*					
Teacher monitoring (TMON)	-0.187	-1.592	72.304		55.611	-39.794***	-18.496	-57.127***
	(24.352)	*(30.817)*	*(52.716)*		*(40.507)*	*(14.249)*	*(25.776)*	*(20.785)*
× initial score	-0.411*	-0.255	-1.106*					
	(0.245)	*(0.297)*	*(0.551)*					
Teacher control variables	No	No	Yes	Yes	Yes			
Other control variables	Yes	No	Yes	Yes	Yes			
Country fixed effects	Yes	Yes	Yes	Yes	Yes			
Year fixed effects	Yes	Yes	Yes	Yes	Yes			
Student observations	2,094,856	2,094,856	404,344	404,344	2,094,856			
Country observations	59	59	38	38	59			
Country-by-wave observations	303	303	76	76	303			
R^2	0.392	0.258	0.367	0.365	0.393			

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). Model in columns (5)-(8) is estimated as one joined model that interacts each assessment measure with four dummies for the quartiles of initial country scores. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 7

Robustness tests

	OECD countries		Non-OECD countries	Control for exclusion rates	Without 2015	Rescaled test scale
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized testing with external comparison (SCOMP)	51.462	22.346***	26.378***	35.439***	35.085***	60.655***
	(30.820)	*(7.479)*	*(5.872)*	*(7.362)*	*(9.954)*	*(15.693)*
× initial score	-0.359		-0.374***	-0.217**	-0.189	-0.507**
	(0.326)		*(0.106)*	*(0.096)*	*(0.125)*	*(0.196)*
Standardized testing for internal comparison (SINT)	58.619*	64.291*	20.508	61.292***	55.777***	8.894
	(32.496)	*(34.495)*	*(18.675)*	*(20.757)*	*(19.008)*	*(30.447)*
× initial score	-0.547*	-0.636*	-0.319*	-0.716***	-0.703***	-0.152
	(0.321)	*(0.343)*	*(0.185)*	*(0.207)*	*(0.209)*	*(0.274)*
Internal reporting (IRPT)	18.179	6.054	-10.840	-11.153	-1.941	-5.212
	(29.982)	*(11.613)*	*(13.040)*	*(12.372)*	*(31.980)*	*(15.369)*
× initial score	-0.134		0.232**	0.126	0.020	0.076
	(0.262)		*(0.105)*	*(0.105)*	*(0.334)*	*(0.131)*
Teacher monitoring (TMON)	46.444	61.681	0.663	4.894	8.063	-72.152**
	(38.979)	*(40.538)*	*(20.416)*	*(29.938)*	*(40.220)*	*(35.725)*
× initial score	-0.733*	-0.887*	-0.342	-0.402	-0.681	0.666*
	(0.385)	*(0.387)*	*(0.315)*	*(0.292)*	*(0.434)*	*(0.359)*
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	1,434,355	660,501	2,045,454	1,679,250	1,698,971
Country observations	35	35	24	59	59	58
Country-by-wave observations	197	197	106	289	247	223
R^2	0.285	0.285	0.443	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 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.