Testing with accountability improves student achievement

Annika B. Bergbauer, Eric Hanushek, Ludger Woessmann 18 September 2018

School systems increasingly use student assessments for accountability purposes. By combining accountability reforms with international student achievement data over the past 15 years, this column shows that the expansion of standardised testing with external comparisons has improved student achievement in maths, science, and reading, while internal testing or teacher inspectorates without external comparisons have not.

Related

A different approach to assessment-based accountability
Derek Neal, Gadi Barlevy

Making the grade: Equity and efficiency in education
Richard Freeman, Stephen Machin, Martina Viarengo

Education and growth: Quality, not quantity
Eric Hanushek, Ludger Woessmann

Testing of students is expanding rapidly in many countries. For example, according to the Education, Audiovisual and Culture Executive Agency of the European Commission, between 2000 and 2015, eight of the 18 European countries covered by the agency introduced national tests to make decisions about students' schooling after primary or lower secondary school (Eurydice 2009, 2017). In 23 of the 59 countries in our analyses, the share of schools that use standardised assessments for external comparison increased by more than 20 percentage points in the same period.

The increase in student testing, and the way it is used in schools, have caused heated debate. Proponents argue that increased use of testing and accountability systems are essential to improve educational outcomes. They argue that measuring how students and schools perform, and where they stand compared to others, creates incentives to improve. For example, in its World Development Reportthat focuses on learning, the World Bank (2018)explicitly calls for expansion of student evaluations and concludes that "[t]here is too little measurement of learning, not too much" (p.17).

In contrast, critics view high-stakes tests with reward and incentive systems as inappropriate (Koretz 2017). They argue that increasing the use of student testing damages schooling (Hout and Elliott 2011).[1]

Differing dimensions of student assessments

In our view, much of this debate is confused. It fails to differentiate among alternative forms and uses of testing. Think of the discussion in the US, where consideration of testing is mostly restricted to accountability systems such as No Child Left Behind (NCLB). This standardised testing is normed to a large population to provide external comparisons with consequences for schools. It is completely different to teacher-generated tests, used to assess the pace of classroom learning.

Similarly, evaluating teachers on the basis of student performance is very different to evaluating which students should go to university.[2] Thus, in reality, there are many dimensions of student
nents. Understanding the overall impact of student testing requires that we consider what
essments are used for, and the incentives they create.

Annika B. Bergbauer
Junior Economist, ifo Institute

Eric Hanushek
Paul and Jean Hanna Senior Fellow at the Hoover Institution of Stanford University

Ludger Woessmann
Professor of Economics, University of Munich; Director, ifo Center for the Economics of Education

Don't Miss

IMF reform: The never-ending quest
De Gregorio, Eichengreen, Ito, Wyplosz

Income inequality in France
Garbinti, Goupille-Lebret, Piketty

Trust and the lending activities of banks and fintech firms

These different ways to configure assessments are likely to vary the strength of performance-conducive incentives for different stakeholders, in different school environments. The impact on affect student learning depends on how the data they create is translated into incentives for the actors, and how those incentives change behaviour.

While there have been previous evaluations of the impact of accountability systems, largely within the US (Figlio and Loeb 2011 provides a review), it is not clear how to generalise from them. Policies operate in a specific institutional environment of national school systems, and so the evaluations neglect features that are common across a nation. Testing policies are often set at the national level, so there is often no adequate comparison group that we can use to evaluate outcomes. Therefore most of the applications of these expanded student assessments, as used for accountability purposes, have not been adequately evaluated.

Using over-time variation in international student achievement tests

In a new study (Bergbauer et al. 2018), we use international contrasts to estimate the effects of different types and dimensions of student assessments on overall levels of student achievement. This expands research studying determinants of student achievement in a cross-country setting (Hanushek and Woessmann 2011 and Woessmann 2016 provide reviews). International comparisons make it possible to consider how overall institutional structures interact with the specifics of student assessments and school accountability systems. This cross-country approach allows us to investigate which aspects of student assessment systems generalise to larger settings, and which do not. Of course, identifying the impact of schooling policies across nations is challenging.

Our empirical analysis exploits the increasing amount of international student assessment data. The Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development (OECD), tests the mathematics, science, and reading skills of representative samples of 15-year-old students. It provides a panel of country observations of student performance. We pool the micro data of more than two million students in 59 countries who participated in six PISA waves between 2000 and 2015.

PISA also includes rich background information on students and schooling institutions in these countries. We derive measures of different types of student assessments from the survey data, and from other international data sources. We use 13 indicators, observed at the country-by-wave level, to create measures of four different categories of test usage that corresponded to different incentive patterns: standardised external comparisons, standardised monitoring without external comparison, internal testing, and internal teacher monitoring.

This database permits country-level panel estimation that relies on within-country over-time analysis of country changes in assessment practices. Because there was rapid change in student assessment policies across countries between 2000 and 2015, we can link policies to outcomes in panel models that include fixed effects for country and year. That is, the estimation ignores any level differences across countries, and uses only changes in student assessment regimes that happen within countries, over time.

By building on an analysis of school autonomy in Hanushek et al. (2013), we use the individual student data for estimation at the micro level, but measure our treatment variables as country aggregates at each point to avoid bias from within-country selection of students into schools. Conditioning on country and year fixed effects allows us to account for unobserved time-invariant country characteristics, as well as common time-specific shocks. Our models further condition on a rich set of student, school, and country measures. The key identifying assumption of our analysis is the standard assumption of fixed-effects panel models. In the absence of reform, the change in achievement in countries that introduced assessments would have been similar to the change in achievement in countries that did not reform (conditional on the included control variables).

External comparison is crucial for testing to improve student achievement

Our results suggest that some uses of student testing affect student learning, while others have no discernible impact. In particular, expanding standardised testing with external comparisons s student achievement, whereas internal testing does not.

Thakor, Merton

Events

The Economic Consequences of Brexit
19 - 19 September 2018 / London School of Economics / London School of Economics

INFER Workshop
20 - 21 September 2018 / Bucharest, Romania / Faculty of Finance and Banking, Bucharest University of Economic Studies

Economic Forecasting with NIESR
24 - 26 September 2018 / London / the National Institute of Economic and Social Research and the University of Warwick

MIGW 2018
24 - 25 September 2018 / Mülheim a.d.R., Germany / University of applied sciences Ruhr-West, Muelheim a.d.R., Germany

Integrating Europe's AI and Cybersecurity Strategies
26 - 26 September 2018 / Press Club Brussels, Rue Froissart 95, 1000 Brussels, Belgium / Center for Data Innovation

CEPR Policy Research

Discussion Papers | Insights

Homeownership of immigrants in France: selection effects related to international migration flows
Gobillon, Solignac

Climate Change and Long-Run Discount Rates: Evidence from Real Estate
Giglio, Maggiori, Stroebel, Weber

The Permanent Effects of Fiscal Consolidations
Summers, Fatás

Demographics and the Secular Stagnation Hypothesis in Europe
Favero, Galasso

QE and the Bank Lending Channel in the United Kingdom
Butt, Churm, McMahon, Morotz, Schanz

Subscribe

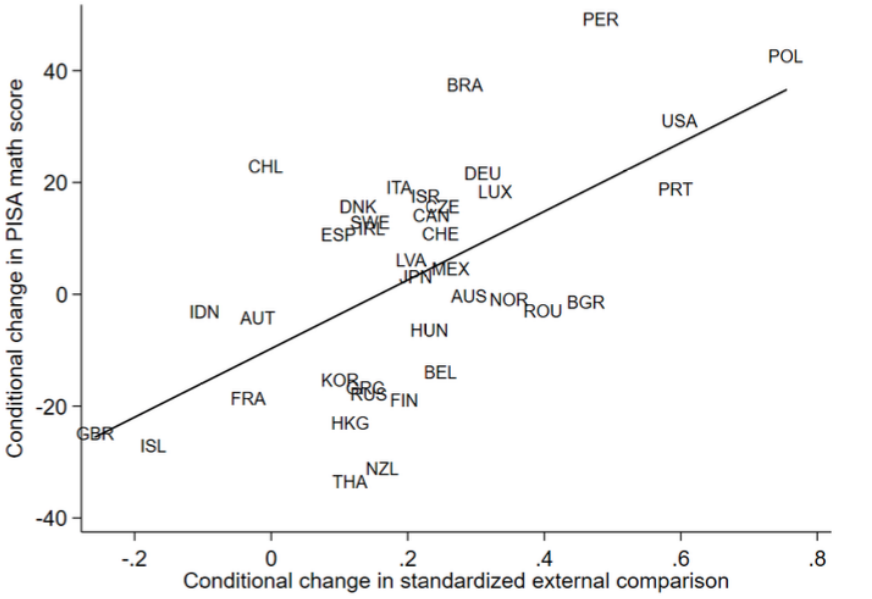
@VoxEU

RSS Feeds

Weekly Digest

Maybe the easiest way to see our results is to look at the long-run changein the use of testing, and in student achievement, between 2000 and 2015. As Figure 1 shows, countries that expanded the use of standardised testing for external comparisons have systematically seen the average math achievement of their students improve, compared to countries that did not expand (or even reduced) this type of testing. By contrast, changes in the use of internal testing are not systematically associated with changes in student achievement across countries.

Figure 1 The relationship between maths test scores and use of standardised testing for external comparison, 2000–2015



Source: Bergbauer et al. (2018).

Notes: Figure is an added-variable plot of the change in countries' average PISA math score against the change in the use of standardised testing for external comparison, conditional on student, school, and country controls.

These results are replicated when we perform individual-level regression analyses that employ the full panel variation in testing regimes and student achievement over the six PISA waves.

Expanded standardised testing that provides external comparisons is associated with increased performance in the international tests. This is true for student achievement in mathematics, science, and reading. It also applies both to school-based forms of external comparisons to district or national performance, and for student-based forms of external comparisons such as national standardized exams that are used for career decisions.

But internal testing that simply informs or monitors progress without external comparability had little discernible effect on overall performance. The same is true for internal assessments used to monitor teachers, including inspectorates. Introducing standardised monitoring without external comparison has a positive effect in countries that were performing poorly, but not in high-performing countries. This mirrors the pattern for the impact of school-based external comparisons, for which there are larger impacts in poorer performing systems.

We also perform a placebo analysis. The use of standardised external comparisons has a significant positive effect on student achievement in the year in which they are implemented, but not in the previous wave. This also indicates that how a country has performed in the past does not predict whether it will implement an assessment reform. This implies it is not likely that endogeneity of assessment reforms to how a school system is performing is a concern for our results. Further robustness analyses show that results are not affected by any individual country, or by changes in PISA testing procedures, and that they are robust to subsets of countries, and when we control for test exclusion rates.

conclusions

The implications of testing regimes are increasingly important for policy, because it has become easier to expand assessments as testing technologies change. Linking accountability systems with reform and improvement has led to worldwide increases in testing. At the same time, a backlash against testing and monitoring in schools has inspired often contentious public debate.

Our results indicate that accountability systems that use standardised tests to compare outcomes across schools and students improve student outcomes. These systems tend to be consequential and produce higher student achievement than those that simply report the results of standardised tests. They also produce better achievement results than systems relying on localised or subjective information that cannot be readily compared across schools and classrooms, which are found to have little impact on student achievement.

Furthermore, the effects of testing and accountability systems are larger in school systems that are performing poorly. The varying impact of student assessments across countries at different achievement levels shows that broad generalisations from specific country testing systems are not always appropriate.

References

Andrews, P, and co-authors (2014), "OECD and Pisa tests are damaging education worldwide," *The Guardian*, 6 May.

Bergbauer, A B, E A Hanushek, and L Woessmann (2018), "Testing", NBER working paper 24836.

Eurydice (2009), National testing of pupils in Europe: Objectives, organisation and use of results, European Commission Education, Audiovisual and Culture Executive Agency.

Eurydice (2017), "Online platform, ec.europa.eu/eurydice", Education Audiovisual & Culture Executive Agency.

Figlio, D and S Loeb (2011), "School accountability", in *Handbook of the Economics of Education* Vol 3, edited by E A Hanushek, S Machin, and L Woessmann, North Holland: 383-421.

Hanushek, E A, S Link, and L Woessmann (2013), "Does school autonomy make sense everywhere? Panel estimates from PISA", *Journal of Development Economics* 104: 212-232.

Hanushek, E A, and L Woessmann (2011), "The economics of international differences in educational achievement" In *Handbook of the Economics of Education,* Vol 3, edited by E A Hanushek, S Machin, and L Woessmann, North Holland: 89-200.

Hout, M, and S W Elliott (2011), *Incentives and test-based accountability in education*, National Academies Press.

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

Ramirez, F O, E Schofer, and J W Meyer (2018), "International tests, national assessments, and educational development (1970-2012)", *Comparative Education Review* 62(3): 344-364.

Woessmann, L (2016), "The importance of school systems: Evidence from international differences in student achievement", *Journal of Economic Perspectives* 30(3): 3-32.

Woessmann, L (2018), "Central exit exams improve student outcomes", *IZA World of Labor* 2018: 419.

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

Endnotes

[1] Even international testing itself – conducted on a voluntary basis in a low-stakes situation – has come under attack for potentially harming the educational programs of countries (Andrews and rs 2014). Recent analysis, however, argues this is not a problem (Ramirez et al. 2018).

[2] See Woessmann (2018) for a review of the literature on central exit exams.

5

A A

Topics: Education

Tags: schools, education, Education reform, standardised testing, student achievement

Related

A different approach to assessment-based accountability
Derek Neal, Gadi Barlevy

Making the grade: Equity and efficiency in education
Richard Freeman, Stephen Machin, Martina Viarengo

Education and growth: Quality, not quantity
Eric Hanushek, Ludger Woessmann