

Bergbauer, Annika Barbara; Hanushek, E.A.; Wößmann, Ludger

Article

Extern vergleichende Prüfungen verbessern die Schülerleistungen

ifo Schnelldienst

Provided in Cooperation with:

Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Bergbauer, Annika Barbara; Hanushek, E.A.; Wößmann, Ludger (2018) : Extern vergleichende Prüfungen verbessern die Schülerleistungen, ifo Schnelldienst, ISSN 0018-974X, ifo Institut – Leibniz-Institut für Wirtschaftsforschung an der Universität München, München, Vol. 71, Iss. 20, pp. 16-19

This Version is available at:

<https://hdl.handle.net/10419/198670>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Annika B. Bergbauer, Eric A. Hanushek und Ludger Wößmann

Extern vergleichende Prüfungen verbessern die Schülerleistungen*

Weltweit steigt der Einsatz schulischer Leistungsüberprüfungen. In einer neuen Studie untersuchen wir die Auswirkungen von Reformen von Tests und Prüfungssystemen in 59 Ländern in den Jahren zwischen 2000 und 2015 anhand der PISA-Mikrodaten von über 2 Mio. SchülerInnen. Es zeigt sich, dass die Einführung extern vergleichender Prüfungen die Schülerleistungen in Mathematik, Naturwissenschaften und Lesen verbessert. Im Gegensatz zu den externen Tests führt die Einführung von schulinternen Tests und Inspektionen der Lehrkräfte ohne externen Vergleich nicht systematisch zu besseren Schülerleistungen. Externe Vergleiche sind also wichtig, damit Prüfungen die Schülerleistungen verbessern.

In vielen Ländern haben regelmäßige Leistungsüberprüfungen von SchülerInnen in den letzten Jahren zugenommen. Beispielsweise berichtet die Exekutivagentur Bildung, Audiovisuelles und Kultur der Europäischen Kommission, dass acht der 18 von ihr untersuchten Länder zwischen 2000 und 2015 nationale Tests eingeführt haben, die am Ende der Grundschule oder Sekundarstufe I über die weitere Schullaufbahn der SchülerInnen entscheiden (vgl. Eurydice 2009; 2017). Im gleichen Zeitraum ist in 23 der 59 Länder, die wir in unserer Analyse betrachten, der Anteil der Schulen, die standardisierte Prüfungen für externe Vergleiche einsetzen, um mehr als 20 Prozentpunkte gestiegen.

Über den Anstieg der Schülertests und die Art ihrer Verwendung im Schulbetrieb wird weltweit viel diskutiert. Einerseits vertreten Befürworter die Meinung, dass der zunehmende Einsatz von Tests und Prüfungssystemen zur Verbesserung des Bildungserfolges notwendig sei. Sie argumentieren, dass die Beurteilung von SchülerInnen und Schulen und deren Leistungsvergleich Anreize für bessere Ergebnisse schaffe. Beispielsweise plädiert die Weltbank in ihrem aktuellen Weltentwicklungsbericht zum Thema Bildung explizit für eine Ausweitung von Leistungsüberprüfungen und kommt zu dem Schluss, dass es zu wenig Messung von Lernergebnissen gibt, nicht zu viel (vgl. World Bank 2018, S. 17). Andererseits sehen Kritiker Leistungstests mit Konsequenzen als unangemessen an (vgl. Koretz 2017) und argumentieren, dass ihr vermehrter Einsatz der Bildung schade (vgl. Hout und Elliott 2011).¹

Bei übermäßigem Einsatz von Leistungsüberprüfungen klagen einige Lehrkräfte, Eltern und SchülerInnen gerne über »Testeritis«.

UNTERSCHIEDLICHE FORMEN DER LEISTUNGS-ÜBERPRÜFUNG

Aus unserer Sicht vernachlässigt diese Debatte oftmals die wichtige Unterscheidung zwischen verschiedenen Formen und Verwendungen der Tests. So ist beispielsweise die Diskussion von Leistungsüberprüfungen in den USA weitgehend auf Rechenschaftssysteme von Schulen beschränkt, wie sie im *No Child Left Behind* (NCLB) Gesetz angelegt sind. Solche auf eine große Schülerschaft normierte Tests, deren externer Vergleich Konsequenzen für die Schulen hat, unterscheiden sich grundlegend von Tests, die von den Lehrern selbst erstellt wurden, um die Lernfortschritte ihrer Schulklasse zu überprüfen. Auch ist eine Beurteilung von Lehrkräften auf Basis der Leistungen ihrer SchülerInnen etwas anderes als die Auswahl von SchülerInnen für den Zugang zur Universität.² In der Realität gibt es folglich viele verschiedene Formen von Leistungsüberprüfungen. Um die Auswirkungen von Prüfungssystemen zu verstehen, müssen deshalb ihre jeweilige Verwendung und die dadurch entstehenden Anreize genau berücksichtigt werden.

Unterschiedliche Ausprägungen der Testverwendung können zu leistungsfördernden Anreizen unterschiedlicher Stärke für unterschiedliche Akteure in unterschiedlichen Umfeldern führen. Wie Leistungs-

der Länder schädigen würden (vgl. Andrews et al. 2014). Aktuelle Analysen weisen dieses potenzielle Problem allerdings zurück (vgl. Ramirez, Schofer und Meyer 2018).

² Eine Übersicht der Literatur zu zentralen Abschlussprüfungen findet sich in Wößmann (2018).

* Dieser Beitrag ist auch auf www.oekonomenstimme.de, 25. September 2018, erschienen.

¹ Auch die internationalen Schülertests selbst, die ja auf freiwilliger Basis ohne Folgen für die Schullaufbahn durchgeführt werden, sind kritisiert worden, weil sie möglicherweise die Bildungssysteme

überprüfungen das Lernen der SchülerInnen beeinflussen, wird also davon abhängen, wie sich die durch die Tests entstandene Information in Anreize und Verhaltensweisen der Akteure umsetzt.

Deshalb ist es unklar, inwieweit sich Erkenntnisse bestehender Analysen der Effekte von Prüfungssystemen, die sich größtenteils auf die USA beziehen (vgl. Figlio und Loeb 2011 für einen Überblick), verallgemeinern lassen. Politikmaßnahmen wirken im jeweils spezifischen institutionellen Rahmen nationaler Schulsysteme; insofern vernachlässigen die Studien allgemeine Bedingungen, die das ganze Land betreffen. Außerdem werden Prüfungssysteme oft auf nationaler Ebene eingeführt, was die Bildung adäquater Vergleichsgruppen zur Evaluierung der Ergebnisse der Reformen erschwert. Folglich wurden die meisten Anwendungen ausgeweiteter Leistungsüberprüfungen bisher nicht angemessen evaluiert.

AUSNUTZUNG VON VERÄNDERUNGEN INTERNATIONALER SCHÜLERTESTS ÜBER DIE ZEIT

In einer neuen Studie (Bergbauer, Hanushek und Wößmann 2018) verwenden wir internationale Vergleiche, um den Einfluss verschiedener Formen und Anwendungen von Leistungsüberprüfungen auf das Leistungsniveau der SchülerInnen zu untersuchen. Unsere Analyse trägt zu der wachsenden Literatur bei, die die Determinanten von Schülerleistungen mit länderübergreifenden Daten erforscht (vgl. Hanushek und Wößmann 2011; Wößmann 2016). Die international vergleichende Analyse ermöglicht es zu betrachten, wie generelle institutionelle Strukturen mit den jeweiligen Besonderheiten von Leistungsüberprüfungen und schulischen Rechenschaftssystemen interagieren. Dieser länderübergreifende Ansatz erlaubt uns zu untersuchen, welche Aspekte von Prüfungssystemen sich auf weite Anwendungsfälle verallgemeinern lassen und welche nicht. Natürlich stehen diesen Vorteilen auch Kosten gegenüber, denn die Untersuchung des Einflusses verschiedener schulpolitischer Maßnahmen im Ländervergleich bringt ihre eigenen Herausforderungen mit sich.

Unsere empirische Analyse nutzt die zunehmend reicheren internationalen Schülerleistungsdaten. Der PISA-Test (Programme for International Student Assessment), in dem die OECD die Mathematik-, Naturwissenschafts- und Leseleistungen repräsentativer Stichproben von 15-jährigen SchülerInnen testet, bietet eine regelmäßig wiederkehrende Beobachtung der Schülerleistungen in vielen Ländern. Konkret kombinieren wir die Mikrodaten von über 2 Mio. SchülerInnen in 59 Ländern aus sechs PISA-Wellen in den Jahren 2000 bis 2015.

Die PISA-Studie umfasst nicht nur Leistungsmessungen, sondern auch umfangreiche Hintergrundinformationen zu den SchülerInnen und Schulen in den teilnehmenden Ländern. Aus diesen Befragungsdaten und weiteren internationalen Datenquellen

haben wir eine Reihe von Maßen unterschiedlicher Formen von Leistungsüberprüfungen erstellt. Wir kombinieren 13 einzelne Indikatoren, die auf der Ebene von Ländern und Wellen beobachtet werden, in vier Kategorien der Testverwendung, die unterschiedliche Anreizmuster widerspiegeln: extern vergleichende Prüfungen, standardisierte Tests ohne externen Vergleich, interne Tests sowie interne Beurteilung der Lehrkräfte.

Diese Datenbasis ermöglicht Panelanalysen auf Länderebene, die auf Veränderungen in den Prüfungssystemen innerhalb der Länder über die Zeit beruht. Da im Beobachtungszeitraum viele Länder ihre Prüfungspraktiken reformiert haben, können wir die Politikmaßnahmen mit Testergebnissen in Panelmodellen verknüpfen, die fixe Effekte für jedes Land und jedes Jahr berücksichtigen. Die Schätzungen ignorieren also sämtliche Niveauunterschiede zwischen den Ländern und nutzen ausschließlich Veränderungen in den Prüfungsregimen, die innerhalb der Länder über die Zeit zu beobachten sind.

Aufbauend auf früheren Analysen der Effekte von Schulautonomie (Hanushek, Link und Wößmann 2013), nutzen wir die Individualdaten der SchülerInnen für Schätzungen auf der Mikroebene, während wir die Maße der Prüfungspraktiken in jedem Jahr auf der Länderebene aggregieren, um Verzerrungen durch eine mögliche Selektion von SchülerInnen in bestimmte Schulen innerhalb eines Landes zu vermeiden. Die Verwendung fixer Effekte für jedes Land und Jahr erlaubt es uns, unbeobachtete zeitinvariante Ländermerkmale sowie gemeinsame zeitspezifische Schocks zu berücksichtigen. Darüber hinaus beinhalten unsere Modelle eine Reihe beobachteter Einflussfaktoren auf Ebene der SchülerInnen, Schulen und Länder. Die wichtigste identifizierende Annahme unseres Ansatzes besteht in der Standardannahme von Panelmodellen mit fixen Effekten: Ohne Reform würde die Veränderung der Schülerleistungen (nach Berücksichtigung der Kontrollvariablen) in Ländern, die Prüfungen eingeführt haben, der Veränderung der Schülerleistungen in den Ländern ähneln, die ihre Prüfungssysteme nicht verändert haben.

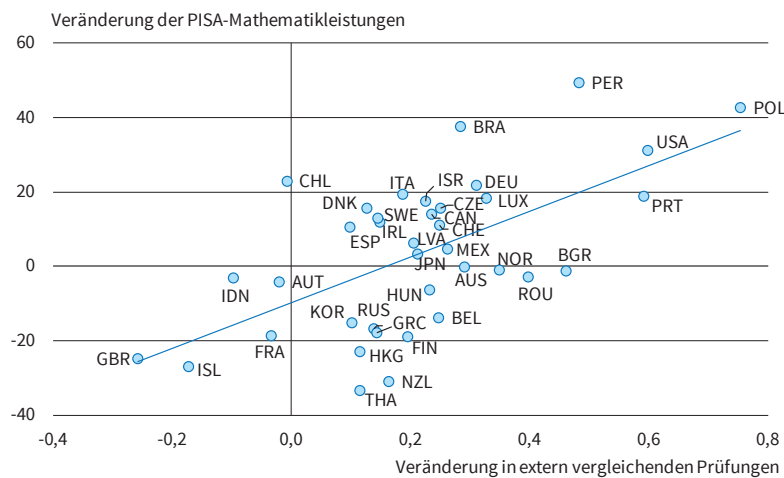
EXTERNE VERGLEICHE SIND ENTSCHEIDEND, DAMIT TESTS DIE SCHÜLERLEISTUNGEN VERBESSERN

Die Ergebnisse zeigen, dass einige Formen der Leistungsüberprüfung das Lernen der SchülerInnen beeinflussen, während andere keinen erkennbaren Einfluss haben. Insbesondere verbessert eine Ausweitung von extern vergleichenden Prüfungen die Schülerleistungen, nicht jedoch eine Ausweitung interner Tests.

Am einfachsten lassen sich unsere Ergebnisse durch einen Vergleich der langfristigen Veränderungen in der Verwendung externer Prüfungen und in den Schülerleistungen der Länder zwischen 2000 und 2015 veranschaulichen. Wie Abbildung 1 zeigt,

Abb. 1

Länder, die zwischen 2000 und 2015 extern vergleichende Prüfungen ausgeweitet haben, haben ihre Schülerleistungen verbessert



Anmerkungen: Zusammenhang zwischen der Veränderung der durchschnittlichen PISA-Mathematikleistung eines Landes von 2000 bis 2015 und der Veränderung in extern vergleichenden Prüfungen nach Herausrechnung zahlreicher Einflussfaktoren auf Ebene der SchülerInnen, Schulen und Länder.

Quelle: Darstellung der Autoren auf Basis von Bergbauer, Hanushek und Wößmann (2018, Tab. 7, Spalte 3).

© ifo Institut

haben sich die durchschnittlichen Mathematikleistungen der SchülerInnen in Ländern, die die Verwendung extern vergleichender Prüfungen ausgeweitet haben, über den 15-Jahreszeitraum systematisch verbessert im Vergleich zu Ländern, die die Verwendung dieser Form von Prüfungen nicht ausgeweitet oder sogar verringert haben. Im Gegensatz dazu gehen Veränderungen in der Verwendung von internen Tests nicht systematisch mit Veränderungen der Schülerleistungen der Länder einher.

Diese Ergebnisse finden sich ebenso in unseren detaillierten Regressionsanalysen auf Individualebene, die die vollständige Variation in den Testpraktiken und Schülerleistungen im Panel der sechs PISA-Wellen nutzen. Einerseits bestätigen die Regressionsergebnisse, dass eine vermehrte Verwendung extern vergleichender Prüfungen mit höheren Schülerleistungen in den internationalen Tests einhergeht. Dies zeigt sich für Schülerleistungen in allen drei Fächern – Mathematik, Naturwissenschaften und Lesen. Darüber hinaus trifft es für schulbasierte Formen extern vergleichender Prüfungen, die die Leistungen der Schulen mit den jeweiligen regionalen oder nationalen Leistungen vergleichen, ebenso zu wie für schülerbasierte Formen extern vergleichender Prüfungen wie national standardisierte Prüfungen, die über den weiteren Bildungsweg entscheiden.

Andererseits ergeben sich für interne Tests, die nur informieren oder Fortschritte überwachen, ohne externe Vergleiche zu ermöglichen, kaum erkennbare Effekte auf die Schülerleistungen. Gleiches gilt für interne Prüfungen, die zur Beurteilung der Lehrkräfte eingesetzt werden, einschließlich Inspektionen von Unterrichtsstunden.

Auch die Einführung von standardisierten Tests ohne externe Vergleiche geht im Durchschnitt nicht

systematisch mit Veränderungen in den Schülerleistungen einher. Allerdings hat sie in ursprünglich schwach abscheidenden Ländern positive Effekte, nicht hingegen in Ländern mit bereits starken Leistungen. Ein ähnliches Muster, dass sich stärkere Effekte in leistungsschwächeren Schulsystemen finden, zeigt sich auch für die bereits angesprochenen Effekte schulbasierter Formen extern vergleichender Prüfungen.

Eine interessante Placebo-Analyse zeigt, dass sich die Schülerleistungen in dem Jahr, in dem extern vergleichende Prüfungen tatsächlich eingeführt werden, verbessern, nicht jedoch in der vorherigen Welle, in der die Prüfungen noch nicht eingeführt worden sind.

Dies zeigt auch, dass das vorherige Leistungsniveau eines Landes nicht vorhersagt, ob das Land in Zukunft Prüfungen einführen wird, was gegen die Möglichkeit spricht, dass eine endogene Einführung der Prüfungssysteme in Bezug auf frühere Leistungen ein relevantes Bedenken bei der Interpretation unserer Ergebnisse ist. Weitere Robustheitsanalysen bestätigen, dass die Ergebnisse nicht von einzelnen Ländern oder von Veränderungen im PISA-Testverfahren abhängen und dass sie robust sind in Untergruppen von Ländern und bei Berücksichtigung von Ausschlussraten aus den PISA-Tests.

SCHLUSSEFOLGERUNGEN

Aus politischer Sicht wird es zunehmend wichtiger, die Auswirkungen verschiedener Prüfungssysteme zu verstehen. Mit den sich verändernden Prüfungstechnologien wird es immer leichter, Leistungserhebungen durchzuführen. Darüber hinaus hat die Verknüpfung von schulischen Rechenschaftssystemen mit Reform- und Verbesserungsvorhaben zu einer weltweiten Zunahme von Tests zu Rechenschaftszwecken geführt. Gleichzeitig haben Gegenreaktionen gegen verschiedene Verwendungsformen von Tests und Leistungsüberprüfungen die Prüfungspraktiken in eine oftmals kontroverse öffentliche Debatte geführt. Unsere Analysen können diese Debatte mit wissenschaftlicher Evidenz informieren.

Unsere Ergebnisse zeigen, dass Prüfungssysteme, die standardisierte Tests zum Vergleich der Ergebnisse mit anderen Schulen und SchülerInnen verwenden, die Schülerleistungen verbessern. Diese Prüfungssysteme haben üblicherweise Konsequenzen für die Beteiligten und erzielen bessere Ergeb-

nisse als Systeme, die standardisierte Tests ohne externe Vergleiche verwenden. Sie erzielen auch bessere Leistungen als Systeme, die sich auf lokale oder subjektive Informationen verlassen, die nicht direkt zwischen den Schulen oder Klassen verglichen werden können. Für solche internen Tests konnten kaum Effekte auf die Schülerleistungen festgestellt werden.

Darüber hinaus sind die Effekte der Einführung von Prüfungssystemen tendenziell stärker in solchen Schulsystemen, die bisher schwache Schülerleistungen aufweisen. Dieser ungleiche Einfluss von Prüfungssystemen in Ländern mit unterschiedlichen Leistungsniveaus zeigt die Gefahren einer allzu breiten Verallgemeinerung der Befunde von spezifischen Prüfungssystemen einzelner Länder auf.

LITERATUR

- Andrews, P. et al. (2014), »OECD and Pisa tests are damaging education worldwide«, *The Guardian*, verfügbar unter: <https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics>, aufgerufen am 20. Juni 2018.
- Bergbauer, A. B., E. A. Hanushek und L. Wößmann (2018), »Testing«, NBER Working Paper 24836, National Bureau of Economic Research, Cambridge, MA.
- Eurydice (2009), *National testing of pupils in Europe: Objectives, organisation and use of results*, European Commission; Education, Audiovisual and Culture Executive Agency (EACEA), Eurydice, Brüssel.
- Eurydice (2017), *Online platform, ec.europa.eu/eurydice*, Education Audiovisual & Culture Executive Agency (EACEA), Eurydice Uni, Brüssel.
- Figlio, D. und S. Loeb (2011), »School accountability«, in: E. A. Hanushek, St. Machin und L. Wößmann (Hrsg.), *Handbook of the Economics of Education*, Vol. 3, North Holland, Amsterdam, 383–421.
- Hanushek, E. A., S. Link und L. Wößmann (2013), »Does school autonomy make sense everywhere? Panel estimates from PISA«, *Journal of Development Economics* 104, 212–232.
- Hanushek, E. A. und L. Wößmann (2011), »The economics of international differences in educational achievement«, in: E. A. Hanushek, St. Machin und L. Wößmann (Hrsg.), *Handbook of the Economics of Education*, Vol. 3, North Holland, Amsterdam, 89–200.
- Hout, M. und St. W. Elliott (Hrsg., 2011), *Incentives and test-based accountability in education*, National Academies Press, Washington, DC.
- Koretz, D. (2017), *The testing charade: Pretending to make schools better*, University of Chicago Press, Chicago.
- Ramirez, F. O., E. Schofer und J. W. Meyer (2018), »International tests, national assessments, and educational development (1970–2012)«, *Comparative Education Review* 62(3), 344–364.
- Wößmann, L. (2016), »The importance of school systems: Evidence from international differences in student achievement«, *Journal of Economic Perspectives* 30(3), 3–32.
- Wößmann, 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, Washington DC.