Bergbauer, Annika Barbara

Research Report

Conditions and consequences of education: Microeconometric analyses

ifo Beiträge zur Wirtschaftsforschung, No. 86

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Bergbauer, Annika Barbara (2019) : Conditions and consequences of education: Microeconometric analyses, ifo Beiträge zur Wirtschaftsforschung, No. 86, ISBN 978-3-95942-069-3, ifo Institut - Leibniz-Institut für Wirtschaftsforschung an der Universität München, München

This Version is available at:
https://hdl.handle.net/10419/213579

ifo
BEITRÄGE
zur Wirtschaftsforschung

86
2019

Conditions and Consequences of Education – Microeconometric Analyses

Annika B. Bergbauer

ifo INSTITUTE
Leibniz Institute for Economic Research
at the University of Munich

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über
http://dnb.d-nb.de
abrufbar.

ISBN: 978-3-95942-069-3

Druck: ifo Institut, München

ifo Institut im Internet:
http://www.ifo.de

ifo
BEITRÄGE
zur Wirtschaftsforschung

Conditions and Consequences of Education – Microeconometric Analyses

Annika B. Bergbauer

Herausgeber der Reihe: Clemens Fuest
Schriftleitung: Chang Woon Nam

ifo INSTITUT
Leibniz-Institut für Wirtschaftsforschung
an der Universität München e.V.

Preface

This study was prepared by Annika B. Bergbauer while she was working at the Ifo Center for the Economics of Education. It was completed in March 2019 and accepted as a doctoral thesis by the Department of Economics at the University of Munich in July 2019. It consists of four self-contained chapters that empirically analyze educational economics. The first paper relates education measures to economic development. The following three papers contribute to the understanding of student performance focusing on standardized testing, teacher specialization, and political environments.

Chapter 1 motivates the topic and puts the four papers into perspective of the re-search background. Chapter 2 assesses the importance of human capital for regional development in Sub-Saharan Africa. The findings suggest a stronger correlation of development with cognitive skills than with quantitative measures of human capital, such as years of schooling. Chapter 3 investigates the achievement impact of alternative uses of student assessments. In result, the expansion of standardized external comparisons associates with improvements in student achievement. The effect of school-based comparison is stronger in low-performing countries. In contrast, only internal testing without external comparison and internal teacher monitoring including inspectorates do not affect student achievement. Chapter 4 investigates how teacher specialization to a field of subjects during stud-ies affects student achievement. The chapter finds that teacher specialization does raise the academic achievement of boys, but not of girls. Chapter 5 examines the influence of European Union (EU) membership of Eastern European countries on student achievement. The findings suggest a positive and statistically significant correlation of EU membership and reading scores.

Keywords: Education, PISA, assessments, accountability, testing, achievement, development, cognitive skills, human capital, nighttime light intensity, teachers, European Union, Sub-Saharan Africa, Germany, international

JEL-No: I28, H52, L15, D82, P51

Acknowledgements

I am indebted to my supervisor Ludger Wößmann for his constructive and purposeful feedback, for providing an excellent research environment of fascinating people and rewarding projects, and for guiding my scientific thinking.
I thank my second supervisor Amelie Wuppermann for her competent advice and for being a role model.
I am grateful to my third supervisor Simon Wiederhold for his profound research ideas and his cheerful character.

I thank my coauthors Rick Hanushek for his encouraging morale, Marc Piopiunik for taking yet another econometric twist, and Raphael Brade for his statistical intuition and inexhaustible helpfulness.

Thanks to my colleagues at the chair Natalie Obergruber, Katharina Werner, Sven Resnjanski, Lukas Mergele, Larissa Zierow, Bernhard Enzi, Lisa Simon, Philipp Lergetporer, Sarah Kersten, Franziska Kugler, Elisabeth Grewenig, Franziska Hampf, Benjamin Arold, Francesco Cinnirella, Ruth Schüler, Anna Wurm, Pietro Sancassani, and Lavinia Kinne. Thanks to our team assistants Ulrike Baldi-Cohrs and Franziska Binder for being the kindly souls of our team.
Thanks for excellent research assistance to Neysan Khabirpour, Jana Bolvashenkova, Julian Wichert, and Renate Frister.
Many thanks to the ifo colleagues Helmut Hoffmann and Christiane Nowack for their kind support to layout and print this document.

The work on *Testing* was supported by the Smith Richardson Foundation and it is part of the Deutsche Forschungsgemeinschaft project CRC TRR 190.
I am thankful to the friendly and helpful staff of the IQB's FDZ, Aleksander Kocaj and Monika Lacher.

I am indebted to Servaas van der Berg, who initiated my research and who is an outstanding teacher.

I thank my mentor Anne Stelzer for her open ear in career and life questions.

I am deeply grateful to my PhD companions Stefan Lautenbacher, Patrick Reich, Tobias Rossmann, Ilka Gerhards, Daniel Stöhlker, Marvin Deversi, Eleonora Guaneri, Felix Klimm, and Shree Ravi for sharing the joy and the trouble of the past three years.

And I truly appreciate the friends from all around the globe for believing in me!

I am proud of my diligent and successful sister, Lisa. I am grateful to 95-years old grandma Erna, thanks to Bärbel and Neil for hosting me during *Freistunde*, to Annelie and Axel for sending innumerable *Nürnberger Lebkuchen*, and to my cousins, Tina and Jörg, and their families.

I am deeply grateful to my boyfriend Jan for his inexhaustible passion to teach me what he loves, for challenging mountain adventures and bringing me home calmly, for integrating me in his family, for sharing his life with me. Thank you for your patient, courageous, faithful love.

And last but foremost, thanks to my parents Martina and Rolf for their love and care and academic ambition, from early on.

Conditions and Consequences of Education – Microeconometric Analyses

Inaugural-Dissertation

Zur Erlangung des Grades
Doctor oeconomicae publicae (Dr. oec. publ.)

eingereicht an der
Ludwig-Maximilians-Universität München
2019

vorgelegt von

Annika B. Bergbauer

Referent:	Prof. Dr. Ludger Wößmann
Korreferent:	Prof. Dr. Amelie Wuppermann
Promotionsabschlussberatung:	27.07.2019

Content

List of figures

List of tables

1 Introduction

Everyone has an interest in education. Education increases individual wellbeing in many dimensions, such as income, health, and happiness. It also augments a society's wellbeing in national income and decreases inequality and conflicts. Thus, education is worth striving for. Yet, we still do not fully understand which factors determine educational outcomes and how those relate to the economy's wellbeing. The field of the economics of education investigates human capital, i.e., the productive use of education, as a function of inputs including institutions, school resources, family background, and student ability.

This dissertation contributes to the literature in the economics of education by investigating determinants of the human capital production function and the interplay of human capital and economic wellbeing. The first paper examines the role of education for economic development in Sub-Saharan Africa. The second paper studies the impact of institutions on student achievement worldwide and the third paper regards the influence of school resources on student achievement in Germany. The last paper examines the link of economic development and student achievement in the European Union (EU).

This introductory chapter contextualizes the four papers of this dissertation within the economics of education. Section 1.1 presents the interplay of education and economic wellbeing in theory and gives examples of empirical applications. Section 1.2 introduces the determinants of human capital. Empirical strategies for causal inference applied in this dissertation are discussed in Section 1.3. The four papers are summarized in Section 1.4 and Section 1.5 proposes policy implications.

1.1 The role of human capital for economic wellbeing

Education is conducive to economic wellbeing at the micro and at the macro level. The role of education for individuals was formalized by Mincer (1967), explaining wages by years of schooling and years of work experience. Both factors relate positively to wages – more schooling and/or more experience yield higher wages. This general model shows differences in empirical evidence across countries where micro returns seem to vary

considerably (Hanushek & Zhang 2009). Hence, the role of education for individuals seems country-specific.

Education is not only beneficial on the individual level; it also contributes to aggregate wellbeing. There are three broad streams in which growth models theorize on the link between aggregate education and wellbeing. First, the augmented neoclassical theory regards human capital as a production factor, besides physical capital and labor, to describe the evolution of an economy over time towards its steady state (Mankiw, Romer & Weil 1992). Second, endogenous growth models use human capital as a source of technological change (Lucas 1988; Romer 1990; Aghion & Howitt 1998). Third, in growth models of firm-level productivity, human capital facilitates technological diffusion (Nelson & Phelps 1966; Benhabib & Spiegel 1994). Empirical analyses confirm the theory's suggestion of considering education as accomplice of economic wellbeing. Income differences within and across countries seem to be driven by differences in education (Mankiw, Romer & Weil 1992; Acemoglu & Dell 2010). To focus on the role of human capital for aggregate wellbeing, Hanushek & Woessmann (2016) propose a growth model similar to the endogenous growth models:

$$g = \varphi H + \beta X + \varepsilon \qquad (1.1)$$

Cross-country growth g depends on skills H and other factors X. Dependent on data availability, H may be measured by years of schooling, enrollment rates, or test scores. The appropriate measure of education has long been debated (Pritchett 2001). Years of schooling and enrollment rates have been almost ubiquitous as a measure of human capital, despite ignoring differences in school quality and the role of the social environment. Qualitative measures implicitly assume that one year of schooling leads to the same increase in knowledge and skills in every country and that formal schooling is the main, or even only, source of education. Yet, informal education may occur in families and among peers. Ignoring the quality differences across education systems and the role of non-school education factors is a major disadvantage of such a quantitative measure of education. In contrast, using test scores as a qualitative measure of education accounts for those differences. Empirical evidence suggests that qualitative measures of education seem to matter more for economic wellbeing than quantitative measures (Hanushek & Woessmann 2008). One limitation of analysis with qualitative education measures is restricted data availability. Not all countries have been regularly participating in student assessments to generate test score measures. Especially

developing countries have only recently expanded their participation in such assessments. Other countries restrict the access to their test results. Yet, even for the available data only a limited number of studies on economic development uses test scores to measure education. Hence, there is a research gap on the link between qualitative education and economic development at the regional level, especially in less developed countries. Chapter 2 aims at diminishing this research gap by examining the relationship between human capital and economic wellbeing across regions in Sub-Saharan Africa.

As these growth models remain silent on the direction of causation – whether higher human capital causes economic growth or whether economic growth allows for higher human capital – one may also examine the relationship the other way around (Bils & Klenow 2000). Chapter 5 does this by exploring the link of economic wellbeing through EU membership and educational achievement in Eastern Europe.

1.2 Determinants of human capital

The source of skills (H in equation (1.1)) may arise from national institutions, local school resources, the family, and individual ability. The human capital production function considers those factors formalized by Hanushek (1970; 1979) and more recently by Hanushek & Woessmann (2011):

$$Y_i = f(I_i, R_i, F_i, A_i) \qquad (1.2)$$

The education outcome Y_i captures human capital measured by test scores from assessments at the individual level i. The input factors are institutions I_i, school resources R_i, family background F_i, and student ability A_i.

Institutions (I) can be considered as non-resource determinant with most evidence from cross-country analyses (for an overview, see Woessmann 2016). Institutions refer to accountability, school autonomy, school choice and competition (e.g., private or public operation), performance pay, and tracking. This dissertation focuses on accountability in Section 1.2.1 because this relevant determinant of human capital lacks general empirical evidence for a broader geographic area as most studies focus on the United States (Carnoy & Loeb 2002; Hanushek & Raymond 2005; Dee & Jacob 2011; Figlio & Loeb 2011).

While in practice, almost every country around the world has increased school accountability in the past decade (for one form of accountability see the increase in school-based external comparisons between 2000 and 2015 in Figure 3.3). As a result, this institutional determinant of education offers considerable variation across countries and over time. Chapter 2 exploits this cross-country variation over time to examine the influence of accountability on student achievement.

The second determinant of human capital, school resources (R), divide into quantitative and qualitative measures (Hanushek & Rivkin 2006). Quantitative school measures include government expenditure or class size and do not seem to contribute to student achievement in cross-country studies (Hanushek & Woessmann 2011).[1] Qualitative measures focus on teachers as the most important school resource (Hanushek 1970). While research has considered several teacher characteristics, such as gender, age and job experience, certification, and race, it could not name the decisive teacher characteristic which increases student achievement (Hanushek 1997; 2003; Harris & Sass 2011). Despite academic uncertainty, policy reforms throughout the world have targeted teachers offering variation for empirical analysis. As literature is still undecided about the critical teacher characteristic maximizing student achievement this dissertation discusses teacher inputs in sub-section 1.2.2.

The third human capital determinant is family (F) background, usually measured by parental income, by the number of books at home, the level of parental education, and the immigration status (Schütz, Ursprung & Woessmann 2008). Family background is often taken as given and rigorously controlled for (Hanushek & Woessmann 2011).[2] This dissertation takes the same approach as the literature and conditions on family background in empirical analysis but does not further investigate it.

While student ability (A) is an important determinant of student achievement and is correlated with other inputs, such as parental characteristics, it is difficult to measure it

[1] Within-country studies exploiting Maimonides rule find ambiguous effects from class size. While Angrist & Lavy (1999) find a large effect, their revision (Angrist et al. 2017) and Hoxby (2000) do not find a statistically significant link of class size and student achievement.

[2] Empirical evidence from international tests in different countries relates higher socio-economic family background to higher student achievement with considerable variation across family backgrounds within and across countries (for an overview of studies, see Hanushek & Woessmann 2011).

correctly.[3] As a result, ability is often neglected in estimations (Hanushek & Woessmann 2011). Facing the same issues this dissertation does not further examine ability.

Overall, the following sub-sections discuss school accountability and teacher inputs as two central determinants of human capital.

1.2.1 Accountability

The idea of school accountability refers to education systems suffering from principal-agent problems where tasks are not completely observable and asymmetric information prevails. For example, parents give teachers the task to educate their children while teachers give students the task to learn. Yet, parents cannot fully observe teachers' effort and teachers cannot fully observe students' effort. Accountability may operate as a solution to this problem by introducing student testing. Test results provide information on the effort of agents, i.e., teachers and students, when reported to principals or parents. Most cross-country empirical analysis suggests a positive link between student tests in the form of standardized central exams and student achievement (Bishop 1997; 2006; Figlio & Loeb 2011; Deming et al. 2016). Other studies suggest distortions in student achievement from increased accountability (Deere & Strayer 2001; Stecher 2008; Figlio & Getzler 2009; Koretz 2009; Neal & Schanzenbach 2010). As a result, the general effect of accountability is unclear. As most evidence refers to the United States (Carnoy & Loeb 2002; Hanushek & Raymond 2005; Dee & Jacob 2011; Figlio & Loeb 2011), it is uncertain how these findings can be generalized for other countries. Large scale international student assessments, such as the OECD's Programme for International Student Assessment (PISA), allow to compare different accountability measures across countries and over time. One form of accountability are student tests used for different purposes, such as providing external comparisons or evaluating teachers on the basis of student performance. Recently, student tests have experienced major expansions in countries around the world and provide considerable variation for empirical analysis. Evaluating

[3] Some econometric approaches eliminate ability as influence on student achievement using value-added models or student-fixed effects. Both approaches compare within-student variation (over time or across subjects) and eliminate the influence of all student-invariant characteristics on achievement. This concerns all characteristics which do not change between assessments, such as gender or ability, and they do not influence student achievement. Yet, most international student assessments do not allow for value-added calculations due to their cross-sectional structure of resampling another set of students every period while within-student across-subject comparisons are rather possible in international assessments.

these changes, Chapter 3 investigates the effect of different forms of student tests on achievement.

1.2.2 Teacher inputs

Teachers seem to be the most important determinant of human capital among the qualitative school resources (Hanushek 1970) and teachers are in the center of policy discussions worldwide (Hanushek 2006). Empirical studies have investigated teaching experience (Hanushek 1997; 2003; Hanushek, Kain & Rivkin 2009; Harris & Sass 2011), certification (Buddin & Zamarro 2009), and teacher subject knowledge (Metzler & Woessmann 2012; Bietenbeck, Piopiunik & Wiederhold 2018; Hanushek, Piopiunik & Wiederhold 2018). Yet, the decisive teacher characteristic which maximizes student achievement remains unclear (Hanushek 1997; 2003; Harris & Sass 2011).

Teacher certification and teaching experience may be regarded as formal pre-service training and as informal in-service training. While teacher certification seems to be unrelated to student achievement and teacher experience seems to increase student achievement in the first three years flattening out afterwards (Harris & Sass 2011), this U.S. evidence may not apply internationally. A potential discrepancy in the effect of teacher certification and experience may be due to teacher training differing considerably across countries. In the United States, candidates with university degrees other than in education can qualify as teachers and may enter the profession many years after university. In Germany, this decision is required upon entering university and there are hardly teacher entrants at a later stage. Hence, teacher certification refers to different teacher training systems across countries and may cause differences in the decisive teacher characteristic which raises student achievement. For example, the U.S. studies examine teacher undergraduate training, college entrance exams, professional development (i.e., in-service formal training), and years of teaching experience and find no significant effect on student achievement (Harris & Sass 2011). This result may be different for German teacher candidates as their initial training integrates content knowledge with pedagogy and prepares specifically for on-the-job tasks. Thus, the effect of teacher certification on student achievement may be stronger and the effect of teaching experiences may be weaker in Germany than in the United States. Yet, this explicit comparison using profound empirical methods remains underexplored in the economic literature.

The third teacher characteristic considered by research is subject knowledge which seems to increase student achievement (Metzler & Woessmann 2012; Hanushek, Piopiunik & Wiederhold 2018; Bietenbeck et al. 2018). Teacher subject knowledge may reinforce through specialization on a field of similar subjects allowing for a deeper content understanding and freeing pedagogical capacities. Consequently, students of specialized teachers may perform better than students of non-specialized teachers. Empirical evidence on the impact of teacher specialization grows. For example, U.S. studies use value-added measures from in-service teachers and find a positive link between teacher specialization and student achievement (Jacob & Lefgren 2008; Condie, Lefgren & Sims 2014; Fryer 2018). Yet, there is little empirical evidence on teacher specialization in Germany. Addressing this gap, Chapter 4 investigates the effect of teacher specialization in a field of subjects on student achievement.

Overall, the four dissertation chapters aim at establishing causal links between human capital and its determinants and at relating human capital and economic development. The following section presents empirical strategies to do so.

1.3 Empirical strategies

Researches in the economics of education aim at finding causal relationships between a treatment and educational outcomes. Yet, the human production function lacks exogeneity of inputs, which prevents causal interpretations. Core concerns originate from omitted variables, sample selection, and reverse causation. For example, high-ability parents tend to hold higher-education degrees and have well-paid occupations. Those parents of high socio-economic status probably choose a high-quality school for their potentially high-ability children. Those children are likely to yield high academic achievement. Yet, it remains unclear, whether the inherited high ability, the parental high socio-economic status, or the high-quality school caused high achievement. Hence, exogenous variation is necessary to derive causal interpretation of a treatment. “Exogenous” refers to variables that determine outside of the system (Angrist & Pischke

2009), i.e., the regarded population has no influence on the appointment of the treatment and there is no spill-over or attrition once the treatment was assigned.[4]

The gold standard of causal estimation are randomized control trials (RCTs) which assign treatment randomly to one population and maintain a control population for comparison (Angrist & Pischke 2009). In a field like education, this is often impossible because of ethical concerns on the determination of life outcomes through education and due to strong resistance of teachers or parents. In absence of random treatment, researches retreat to quasi-experimental methods exploiting natural experiments or political reforms, where treatment is as good as random because it follows externally-set rules which the target population could not influence.

This dissertation employs three different empirical strategies. First, Chapter 2 uses cross-regional correlations, Chapters 3 and 5 exploit policy reforms of countries over time, and Chapter 4 exploits the variation between subjects within students.

1.3.1 Cross-regional analysis

The simplest form of empirical analysis without exogenous variation are ordinary-least-squares (OLS) regressions. Chapter 2 applies this approach to quantify the correlational link between skills and economic development at the regional level within countries. This approach builds on within-country between-regions variation using country-fixed effects. A country-fixed effects approach requires panel data, i.e., repeated observations on the same country. Chapter 2 builds on 112 regions in 15 countries. The fixed effects eliminate country-specific unobserved determinants of national economic development, such as the national education system or property rights. The remaining variation of economic development at regional level may be influenced by factors at the regional level other than skills which may be controlled for, such as population density, geography, nature,

[4] For example, exogenous variation can be achieved by assigning a treatment, e.g., a seat in a high-quality school, to one group of the regarded population randomly chosen by a lottery. Another group is retained for comparison without the treatment, i.e., they attend another school. Random appointment of the treatment should be independent of the students' background, e.g., family characteristics and individual ability are equally distributed across treatment and control group. As a result, comparing student achievement across treatment and control group yields the treatment effect of the high-quality school independent of family background. In this case, higher-school quality may have increased student achievement. In summary, causal interpretation necessitates a force independent of the affected population.

health care, and conflicts. This empirical strategy strongly depends on avoiding omitted variable bias (OVB) which is an unobserved determinant of regional economic development correlated with skills.

Lacking exogenous variation comparable across countries prevents causal interpretation of this empirical strategy and offers to quantify the correlation between skills and regional economic development.

1.3.2 Difference-in-differences

This empirical approach allows to compare potential educational outcomes in the treated and untreated state. The most evident exogenous treatment on national level arises from policy reforms. It is exogenous to the student population because students have no direct influence on the reform's introduction and they could not select into or out of the treatment. National reforms may be compared across countries because their outcomes differ from the common trend. The difference-in-differences approach necessitates cross-country panel data which includes multiple countries over multiple periods. Similar to the previous empirical strategy, one applies country- and time-fixed effects. Country-fixed effects control for unobserved time-invariant country characteristics. For example, people of one country prefer more education than in another country and the first country's national spending on education is higher than of the second country. Country-fixed effects absorb this cross-country difference. Country-fixed effects also absorb the cross-country difference in the intensity of the national treatment. Time-fixed effects eliminate all events specific to a period common across all countries. For example, a general trend towards more education across all countries is absorbed by the time-fixed effects. Hence, the remaining variation after including country- and time-fixed effects, exploits changes over time within a country. Thus, the treatment is defined by a deviation from the common trend while the level of treatment and controls may differ (Angrist & Pischke 2009).

Key assumption to this approach is the parallel trends assumption. This means that in absence of the treatment the educational outcome follows a common trend across all countries given the country- and time-fixed effects. This assumption may be investigated using multiple periods – pre and post treatment. One option is to visualize outcome trends graphically. Another option to validate causality is to examine whether the treatment happens before the outcome or vice versa. Using leads and lags of the outcome

compared to the treatment should yield no effects prior to the treatment, an increase in effects at the time of the treatment, and the flattening out of effects after the treatment.

Country-level analysis has three main advantages. First, it circumvents selection problems. For example, students are unlikely to move to another country to attend school. Yet, they are likely to move to another neighborhood within a city to attend another school and benefit from higher school quality. Hence, an analysis of country-level treatments suffers from unobserved heterogeneity within countries because it fails to capture variation below the national level. Second, country-level analysis allows to condition on time-varying factors at the sub-national level to avoid bias from within-country changes over time. For example, the number of students at a school may increase. With a constant number of school resources available, i.e., teachers, the individual student receives less support and her or his achievement may decrease. As a result, student achievement decreases due to changes on the sub-national level but not due to the national reform. Hence, conditioning on such potential mechanisms at a lower level than the treatment increases precision of the estimates. Third, comparing national policies across countries provides an adequate comparison group which is absent within single countries.

One example of a policy reform comparable across countries is the intensification of accountability. Chapter 3 exploits a cross-country panel in which some countries increase accountability more than other countries and estimates the effect of accountability on student achievement. Another example is the effect of EU accession of Eastern European countries on student achievement compared to always and never members of the EU, as investigated in Chapter 5.

1.3.3 Student-fixed effects

Another empirical strategy which offers causal interpretation is a student-fixed effects approach used in Chapter 4. This approach exploits differences in a characteristic observed several times of the same student – either over time or across subjects. Longitudinal data of the same student allows for a value-added model which isolates the influence of the variable of interest on the increase in skills from one period to the next. Yet, large scale student assessment studies, such as PISA, resample their population in each period which prevents observing the same student over time. In contrast, those studies provide multiple observations per student in different subjects at the same point

in time. Comparing skills between subjects of the same student, i.e., using student-fixed effects, holds constant subject-invariant determinants of achievement at student, family, school, and country level. Hence, the difference in achievement arises from the subject-specific variable of interest.

The student-fixed effects literature usually investigates the effect of teacher characteristics on student achievement (see e.g., Dee 2005; 2007; Schwerdt & Wuppermann 2011; Metzler & Woessmann 2012; Bietenbeck 2014; Bietenbeck, Piopiunik & Wiederhold 2018; Falck, Mang & Woessmann 2018; Hanushek, Piopiunik & Wiederhold 2018).

Key assumption of this approach is the random assignment of teachers to the treatment. This means that teachers must not select into treatment based on an unobserved characteristic which causes differences in student achievement. For example, Chapter 4 examines the effect of teacher specialization on student achievement. Possibly, more diligent teachers specialize rather than lazy teachers. If diligent specialists prepare their classes better which increases student achievement, the effect of teacher specialization is due to teacher diligence and not due to specialization. Thus, the estimated effect is biased by an unobserved teacher characteristic determining selection into treatment. One approach to mitigate concerns on omitted variable bias, is to verify that treatment and control group do not differ in observable characteristics. Overall, student-fixed effects across subjects offer causal interpretation of estimated treatment effects under the same assumption as other empirical methods – no selection into treatment which influences outcomes – while other assumptions are relaxed, such as subject-invariant school resources or family background.

1.4 Outline of the dissertation

This dissertation studies the interplay of economic development and human capital and examines the determinants of human capital.

Chapter 2 (in collaboration with Marc Piopiunik and Simon Wiederhold) assesses the importance of human capital for regional development in Sub-Saharan Africa. We aggregate geo-coded skill data for more than 120,000 students to 112 regions in 15

countries. Our within-country models show that cognitive skills of the population are strongly and robustly associated with economic development, measured by nighttime luminosity. Cognitive skills are more important for development than quantitative measures of human capital, such as years of schooling. Results are robust to accounting for various other determinants of regional development, including population density, geography, nature, health care, and conflicts.

Chapter 3 (in collaboration with Eric Hanushek and Ludger Woessmann) investigates the achievement impact of alternative uses of student assessments. Our dataset covers over two million students in 59 countries observed over six waves in the international PISA from 2000 to 2015. Our empirical model exploits the country panel dimension to investigate assessment reforms over time, accounting for country and year fixed effects. The expansion of standardized external comparisons, both school-based and student-based, is associated with improvements in student achievement. The effect of school-based comparison is stronger in low-performing countries. In contrast, only internal testing without external comparison and internal teacher monitoring including inspectorates do not affect student achievement.

Chapter 4 (in collaboration with Raphael Brade) investigates how teacher specialization to a field of subjects during studies affects student achievement. We estimate the effect of teacher specialization on adolescents' skills in languages and sciences using within-student between-subject variation applied to the German National Assessment Studies of 2012 and 2015. We find that teacher specialization raises the academic achievement of boys, but not of girls. The finding is constant across school tracks, student achievement, and socio-economic status. We find that teacher specialization influences students' attitude towards the subjects, and that attitudes are a mechanism to increase boys' language skills. Teacher-student gender match, and further teaching characteristics seem to be no mechanisms.

Chapter 5 examines the influence of EU membership of Eastern European countries on student achievement. This chapter builds on a panel of six PISA waves over 15 years covering more than one million students in 32 countries. Using a difference-in-differences approach – like Chapter 3 –, Chapter 5 finds a positive and statistically significant correlation of EU membership and reading scores by 0.1 standard deviations (SD). Exploring mechanisms that may transmit EU membership to academic achievement,

school efficiency, family wealth, and family structure seem to be key. Results are robust to negatively selected emigration, endogeneity, and sample composition.

1.5 Policy implications

Societies aim at improving student achievement using policy. Empirical evidence could inform policy-making. There are two approaches to increase student achievement. First, a policy may provide more resources while their effective use is crucial (Hanushek 2003; Hanushek & Woessmann 2011). Second, a policy may change incentives formed by the institutions. This dissertation suggests five potential policy interventions to enhance student achievement. Suggestions from Chapters 2 and 4 speak to the resource approach and Chapters 3 and 5 speak to the incentive approach.

This dissertation suggests three resource related policies. First, increasing educational resources may include the generation of educational information. Policy makers could consider expanding data generation to receive a basis for informed decision-making. As of Chapter 2, analysis of student achievement and economic development in Sub-Saharan Africa is complicated by limited data availability. While standardized assessments of student skills are reliable and available only in certain years and countries, data on economic development measured by the Gross Domestic Product (GDP) per capita has limited reliability (Jerven 2013). This suggests a lack of quality and quantity in regard to education and economic data. More and better data could help to inform policy-making. Hence, developing countries could consider to strengthen their statistical systems by training statisticians, creating positions in statistical offices, and conducting surveys.

Second, the effective use of resources relates to Chapter 2's finding that school quality matters for economic development in Sub-Saharan Africa. After the United Nations *Millennium Development Goal* (MDG) No. 2 aimed at achieving universal primary enrollment by 2015, the succeeding program proposes the *Sustainable Development Goal* (SDG) No. 4 aiming at minimum proficiency standards in reading and mathematics by 2030. National governments may contemplate on joining this initiative.

Third, teachers are a key school resource and Chapter 4 indicates a persistent impact of teacher specialization during studies on student achievement. Hence, initial teacher training may be taken into consideration by politicians and researchers.

In addition to the resource-related proposals, this dissertation suggests two incentive policies. First, information on the education system could be published and connected with consequences. Chapter 3 suggests a positive link of student assessments and student achievement when test results are used for external comparison. Hence, education systems may reflect on adopting accountability measures with external reporting to strengthen incentives to perform. Chapter 3's results indicate that this is particularly effective in poorly performing school systems.

Second, Chapter 5 suggests a positive impact of EU membership on student achievement. At the same time, returns to schooling increased, providing an incentive to perform better at school. This research may be a kind reminder for politicians that international collaboration appears to enhance economic and educational wellbeing.

2 Let Africa shine [5]

Sub-Saharan Africa is the economically least developed region in the world (Easterly & Levine 1997). The poor economic development has dramatic consequences as it comes along with high rates of infant mortality, low life expectancy, and low-calorie intake, among others. Despite the low overall development, there is also huge variation across countries within Sub-Saharan Africa. For example, GDP per capita is $ 778 in Mozambique, but $ 19,756 in the Seychelles, a difference by a factor of 25 (World Bank 2019).[6] Several studies highlighting the role of human capital in economic development would suggest that the huge cross-country differences could be reduced by increasing the human capital of the poorly developed countries (e.g., Mankiw, Romer & Weil 1992; La Fuente & Doménech 2006; Cohen & Soto 2007). While these studies measure human capital as average years of schooling of the population, other research indicates that years of schooling matter for development only to the extent to which they improve the cognitive skills of the population (Hanushek & Woessmann 2008). Therefore, focusing exclusively on formal education (such as years of schooling) – and ignoring differences in cognitive skills – distorts the picture about the relationship between human capital and economic development.

In this chapter, we explore the relationship between human capital and economic development across subnational regions in Sub-Saharan Africa. The primary innovation of our study is to combine region-level data on test scores of more than 120,000 students – a proxy for the cognitive skills of the adult population – with satellite data on nighttime luminosity – our main measure of regional economic development. In a within-country analysis of 112 regions in 15 Sub-Saharan African countries, we account for various other potential determinants of regional development, including quantitative measures of human capital (e.g., years of schooling), geography, nature, health care, and regional conflicts.

[5] This chapter is joint work with Marc Piopiunik of ifo Institute at the University of Munich and CESifo, and with Simon Wiederhold of Catholic University Eichstätt-Ingolstadt, ifo, CESifo, and ROA.

[6] Figures refer to 2007 (i.e., our year of analysis), while the comparison includes only countries in our sample. GDP per capita is measured in PPP-$.

In Sub-Saharan Africa, economic development also differs vastly across *subnational* regions, with GDP per capita varying by a factor of up to 2,500.[7] Similarly, student test scores vary strongly across regions and countries (Bietenbeck, Piopiunik & Wiederhold 2018).[8] We exploit the substantial regional variation in both economic development and cognitive skills in Sub-Saharan Africa. We examine the determinants of regional economic development in cross-sectional specifications that include country fixed effects to avoid picking up unobserved country-specific determinants of economic development. Among others, country fixed effects control for differences in national economic factors such as a country's industry specialization as well as for differences in national institutions, in particular, education systems, openness of the economy, and security of property rights (Acemoglu, Johnson & Robinson 2001).[9] Moreover, our within-country estimates are not affected by cross-country cultural differences in the use of nighttime luminosity versus daytime activities, public versus private lighting, and national conditions for generating electricity.[10]

We measure regional economic development with nighttime luminosity rather than with GDP per capita since GDP has been shown to suffer from substantial measurement error

[7] The richest region in our sample is Gauteng Province in South Africa with a GDP per capita of 14,634 PPP-$ and the poorest region is Mashonaland West in Zimbabwe with 5.8 PPP-$ in 2005 (Gennaioli et al. 2013).

[8] Student achievement is substantially lower in Sub-Saharan Africa than in any developed country. The average performance of students in Sub-Saharan Africa is even dismal when compared to that of students in other developing countries (for a comparison with students in India, see Hanushek & Woessmann 2012).

[9] Appendix Table A 2.1 shows that the education systems vary in many aspects across countries (e.g., students absent from school, duration of compulsory education, share of GDP spent on education, pupil-teacher ratio). Similarly, the economic structure, as measured by the share of agriculture, industry, and services, differs enormously between countries (Appendix Table A 2.2).

[10] While we control for any differences between countries by including country fixed effects, we do not claim that our cross-sectional estimates can be interpreted causally. Our estimates might be biased due to reverse causality and omitted regional factors (see the discussion in Section 2.2).

in Sub-Saharan African countries (Jerven 2013).[11] Therefore, many existing studies on developing countries, particularly those considering subnational levels, have used nighttime luminosity recorded by U.S. Air Force weather satellites (Elvidge et al. 2009; Chen & Nordhaus 2011; Henderson, Storeygard & Weil 2011; Henderson, Storeygard & Weil 2012).[12] The usefulness of nighttime luminosity as a proxy for economic development has been demonstrated in previous work, which established a strong within-country correlation between nighttime luminosity and GDP levels and growth rates (Henderson, Storeygard & Weil 2012). Furthermore, nighttime luminosity is strongly associated with access to electricity and public-goods provision, especially across low-income countries (Min 2008), as well as with regional indicators of household wealth in Africa (Bruederle & Hodler 2017). Moreover, while region-level data on GDP per capita exist in some countries in Sub-Saharan Africa, nighttime luminosity is more widely available, allowing us to investigate the relationship between human capital and regional economic development for a larger set of countries. Importantly, cross-sectional comparisons will work best across regions with similar cultural uses of lights, geography, population density, and extent of top-coding (Ghosh et al. 2010). This likely holds for our setting since we compare only regions within countries. Since the distribution of nighttime luminosity in our sample is strongly right-skewed with a concentration of 69 percent of regions with less light than 1 DN (see also Appendix Figure A 2.3), we follow Henderson, Storeygard & Weil (2012), Michalopoulos & Papaioannou (2013; 2014) and

[11] In the Penn World Tables (PWT), one of the standard compilations of cross-country data on GDP and income, countries are given data quality grades of A, B, C, and D. Chen & Nordhaus (2011) report that the margins of error (root mean squared error) corresponding to these grades are 10 percent, 15 percent, 20 percent, and 30 percent, respectively. All 43 countries in Sub-Saharan Africa receive either grade C or D. There is also the problem of politically motivated misreporting of output in African countries (Jerven 2013). The severity of the problem is prominently illustrated by the case of Nigeria, where after an extensive revision of the statistical office's methodology GDP numbers were revised upwards by 60 percent overnight (Roger 2018). Furthermore, the calculation of GDP is more difficult in developing countries because a larger fraction of economic activities takes place outside the formal sector and because the government statistical infrastructure is weaker.

[12] U.S. Airforce satellites detect light emission on the Earth's surface emitted either naturally, such as sun light and moonlight, or man-made, such as campfires or streetlights. The National Oceanic and Atmospheric Administration (NOAA) provides satellite data cleaned from natural nightlight, aggregated to annual frequency and with 1-kilometer resolution. Light intensity is measured by an integer value between 0 (unlit) and 63 (top-coded maximum). In our sample, no pixels are top-coded. For our analysis, we aggregate pixel-level night light data to the level of the first administrative unit in each country by averaging all pixel values in a region.

Hodler & Raschky (2014) in log transforming the variable. However, results are not sensitive to the log transformation.

When investigating determinants of economic development, human capital is typically assessed by quantitative measures such as years of schooling or share of tertiary-educated workers. However, the appropriate measurement of education has long been debated (Pritchett 2001). For example, using years of schooling as a human-capital measure implicitly assumes that one year of schooling increases knowledge and skills by the same amount in every region. While this assumption may be more critical for cross-country comparisons, regions within the same country likely also differ with respect to the quality of education. Furthermore, the years-of-schooling measure assumes that formal schooling is the primary, or even only, source of education and that differences in non-school factors (e.g., families) have a negligible effect on human capital. Ignoring regional differences in producing human capital is a major drawback of quantitative measures of human capital.[13] Regional differences in the quality of education seem to be particularly important in Sub-Saharan Africa, where education inputs are often missing (World Bank 2018b) and teachers are poorly qualified (Bold et al. 2017; Bietenbeck, Piopiunik & Wiederhold 2018) and often absent from the classroom (Duflo, Hanna & Ryan 2012).

When investigating the role of human capital for economic development, it therefore seems essential to not only consider the amount of schooling, but rather to focus directly on how much students have learned. Following Hanushek & Woessmann (2008; 2012), we obtain information on the *quality* of human capital from student achievement tests, which are interpreted as a proxy for the cognitive skills of the labor force. This measure of human capital includes knowledge and skills acquired both inside and outside formal schooling (e.g., through family, peers, and society). More specifically, we measure cognitive skills by the math and reading test scores of more than 120,000 sixth-grade

[13] Individuals in Sub-Saharan Africa acquire less skills in a school year than individuals in other countries (Glewwe, Maïga & Zheng 2014). Therefore, the lack of cognitive skills in Sub-Saharan Africa is even more severe than the lack of formal education compared to other world regions (Hanushek & Woessmann 2008).

primary-school students, who have been assessed by the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) in 2007.[14]

We find that cognitive skills are a significant determinant of regional economic development. An increase in math test scores by 1 percent is associated with an increase in nighttime luminosity by 5.8 percent.[15] All models account for population density as an important determinant of nighttime luminosity. The test-score estimate suggests that if the worst-performing region would catch up to the best-performing region in terms of math test scores, the luminosity gap between both regions would decrease by almost one-half. When additionally including years of schooling as a quantitative dimension of human capital, we find that both math test scores and years of schooling are significantly related to economic development. However, the elasticity of nighttime luminosity with respect to test scores is four times larger than with respect to years of schooling. This suggests that cognitive skills are more important than years of schooling for explaining economic development in Sub-Saharan Africa. Cognitive skills also remain to be a relevant determinant of regional development when additionally controlling for various variables of geography, nature, health care, and fractionalization.[16]

Several robustness checks support our finding that cognitive skills are significantly related to regional economic development. Using alternative measures of regional development – nighttime luminosity per capita and GDP per capita – yields qualitatively similar results. Furthermore, our findings are robust to using alternative quantitative

[14] SACMEQ is a collaborative network of 15 Sub-Saharan African ministries of education and the UNESCO International Institute for Educational Planning (IIEP). There have been three SACMEQ assessments: the first assessment (conducted in 1995) covered seven countries, the second assessment (2000) 15 countries, and the third assessment (2007) 16 countries. Unfortunately, we cannot link changes in student test scores over time to changes in economic development to estimate growth models because regional nighttime luminosity has been very persistent during this short time period (see Figure A 2.4).

[15] Section 2.2 shows that using dependent and independent variables in logarithm follows from a standard Cobb Douglas production function framework. When using a log-linear specification (with nighttime luminosity in logarithm and test scores linearly), we find that an increase in math test scores by 1 standard deviation (measured at the regional level) is associated with an increase in night light intensity by about 60%. We also show that the implied effect size of the log-linear specification is very similar to that of our preferred log-log specification.

[16] Unfortunately, no regional data on physical capital are available. However, Acemoglu and Dell (2010), who also investigate developing countries, argue that disparities in physical capital across regions within countries are unlikely to be a primary determinant of economic differences because of the relatively free mobility of capital within national boundaries.

measures for human capital, such as primary and secondary enrollment rate, the share of tertiary educated, and the literacy rate. Results are also robust to excluding those regions within each country with the most light (e.g., those containing capital cities) or the least light and are not driven by any single country. Regional development is also significantly related to the reading skills of the population. However, in line with existing research, the association is somewhat stronger with math test scores than with reading test scores (5.8 percent vs. 5.2 percent).

This chapter contributes to the literature on the importance of education, geography, ethnic diversity, and institutions for development at the regional level (e.g., Acemoglu & Dell 2010; Gennaioli et al. 2013; Michalopoulos & Papaioannou 2013; Michalopoulos & Papaioannou 2014; Gershman & Rivera 2018). Several studies have already connected education and nighttime luminosity at the subnational level. In India, the population share with a completed secondary school or college degree is positively related to nighttime luminosity (Castelló-Climent, Chaudhary & Mukhopadhyay 2017). In rural Indonesia, location-specific human capital, i.e., farming skills for rice production proxied by agroclimatic similarity, is positively associated with nightlight luminosity decades later (Bazzi et al. 2016). Acemoglu and Dell (2010) find that years of schooling and experience of the labor force can explain a significant fraction of income disparities across and within countries in the Americas, but unexplained, residual factors are also significant and generally of comparable magnitude. Our findings indicate that (part of) these residuals likely reflect differences in cognitive skills across regions, an important component of human capital that has not been accounted for thus far.[17]

Our study is the first to investigate the importance of the quality of human capital (cognitive skills) in addition to the quantity of human capital (years of schooling) for regional economic development in Sub-Saharan Africa. Consistent with previous studies at the country level, we find that cognitive skills are more important for economic development than years of schooling. Our results suggest that cross-regional studies that include only quantitative human-capital measures vastly underestimate the role of human capital in explaining differences in economic development.

[17] Using data on several thousand firms, Gennaioli et al. (2013) find that region-level education influences regional development through the education of workers, the education of entrepreneurs, and perhaps regional externalities. Since we do not have firm-level data, we cannot separate those channels.

The chapter proceeds as follows. Section 2.1 describes the data, in particular the qualitative and quantitative measures of human capital and the satellite data on nighttime luminosity. Section 2.2 lays out the empirical model. Section 2.3 presents the main results and robustness checks. Section 2.4 concludes.

2.1 Data

This section describes our region-level data. Our main outcome measure of economic development is nighttime luminosity. Our main variable of interest is the level of cognitive skills in a region obtained from a large-scale student assessment. We contrast this quality-based human capital measure with years of schooling and other quantity-based measures. We also account for further determinants of economic development, such as population density, geography, nature, health, and regional conflicts.

2.1.1 Math and reading test scores

We draw on test scores from the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), a collaborative network of 15 Sub-Saharan African Ministries of Education and the UNESCO's International Institute for Educational Planning (IIEP).[18] The network periodically conducts international assessments of the math and reading knowledge of sixth-grade primary-school students. By means of student, teacher, and principal questionnaires, it also collects detailed background information on student and teacher characteristics, as well as on classroom and school resources. The first wave of this assessment took place in 1995 and covered seven countries; the second wave, in 2000, covered 15 countries; and the third wave, in 2007, covered 16 countries.[19] We use the third wave of SACMEQ (SACMEQ III) as it is the most recent and covers the largest number of countries.

[18] Participating countries are Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, the Seychelles, South Africa, Swaziland, Tanzania, Uganda, Zambia, Zanzibar, and Zimbabwe. Angola has observer status.

[19] A fourth wave has already been conducted between 2012 and 2014, but the data is not yet available.

SACMEQ employs a two-stage clustered sampling design to draw nationally representative samples of sixth-grade students for each participating country. Schools are sampled within predefined geographical strata in the first stage, and a simple random sample of students is drawn from each selected school in the second stage. In the third wave, 25 students per school were sampled randomly. The SACMEQ student assessments are designed to reflect the elements common to the math and language curricula in the participating countries. The multiple-choice tests contain items developed by SACMEQ itself, as well as items from other international student assessment such as the Trends in International Mathematics and Science Study (TIMSS). Students in all participating countries are administered the same tests at the end of sixth grade, with tests translated into the local language of instruction if it is different from English. We use the test performance of students in math and reading, which have been scaled to a mean of 500 points and a standard deviation (SD) of 100 points across students participating in the second SACMEQ wave using Item Response Theory (IRT).

Importantly, SACMEQ contains identifiers of the education regions of participating schools.[20] We adapt the education regions used in SACMEQ to match the official boundaries of the first administrative divisions below the national level.[21] In robustness checks, we show that results hold when using only countries where SACMEQ education regions perfectly match the official administrative boundaries (Column 4 of Appendix

[20] For some countries, SACMEQ III shapefiles are readily available at StatSilk (2016). For the remaining countries, shapefiles with regional borders can be obtained from the Database of Global Administrative Areas (GADM 2016).

[21] SACMEQ education regions and administrative regions perfectly overlap in Kenya, Lesotho, Namibia, Swaziland, South Africa, Zimbabwe, and Zambia (i.e., for 78 out of 112 regions in our sample). For the remaining eight countries (Botswana, Malawi, Mauritius, Mozambique, the Seychelles, Uganda, Tanzania, and Zanzibar), we manually assign education regions to official regions as follows: We assign the separately sampled, semi-autonomous island of Zanzibar to Tanzania (as a separate region). In mainland Tanzania, we reassign “MWA” (Mwanza) and “NEA” (North East) to “NOR” (North). A decentralization policy in Malawi created six out of formerly three regions. We assign “CEA” (Central East) and “CWE” (Central West) to “CEA” and “SEA” (South East), “SHI” (Shire Highlands), and “SWE” (South West) to “SEA”. In Botswana, we assign “GAB” (Gaborone) and “SOC” (South Central excluding Gaborone) to “SOU” (South). In Mauritius, due to incomplete survey documentation, we could not match the education regions “EBB” (East and Bassin) and “WEV” (Vacoas and West) to official regions and therefore dropped these two education regions. In Mozambique, “CID” (Citadelle) does not have an official regional counterpart and is therefore also dropped. For the Seychelles, we assume that all schools on the main island fall in the region “Central”. In contrast, “EAS”, “ISL”, and “WES” cannot be identified in administrative boundaries or survey documentation, so excluded these three education regions. In Uganda, we combine “SOW” (South West) and “WES” (West) to “WES” and assign “NEA” (North East) to “NOR” (North).

Table A 2.11). In total, we obtain an estimation sample with 112 regions in 15 countries (see Appendix Figure A 2.1).

Using our regional classification, we aggregate SACMEQ student test scores (separately for math and reading) to obtain region-level estimates of the quality of human capital of the population. On average, we observe 1,295 students in a region, ranging from 354 students in the South of the Seychelles to 2,972 students in South Africa's KzwaZulu-Natal (see Appendix Table A 2.6). The average math test score is 505 points with an overall SD across regions of 48 points (see Appendix Table A 2.4). Figure 2.1 shows that math test scores differ widely both across countries and regions. The between-country SD is 16.7, with students in Kenya scoring on average more than 1.2 international SD higher than students in Zambia. However, the within-country test scores differences are even larger (within-country SD: 47.9). For instance, the regional variation is largest in Mauritius, with a differential of 2.1 international SD. We investigate whether regional differences in student test scores – a proxy for the cognitive skills of the adult population – are a relevant determinant of within-country differences in regional economic development.[22]

2.1.2 Regional economic development

Our main measure for regional economic development is nighttime luminosity. Nighttime luminosity is recorded by U.S. Airforce weather satellites. We use Version *4 DMSP-OLS Nighttime Lights Time Series in Average Visible, Stable Lights, and Cloud Free Coverages* composed by the US Air Force Weather Agency (NOAA's National Geophysical Data Center 2016).[23] We use nighttime luminosity from 2007, the year in which student math and reading achievement has been assessed. The data report nightlight from persistent human settlement including gas flares while excluding temporary events, such as fires. All natural light was removed, such as sun- and moonlight, glare, and aurora lighting in the northern hemisphere. Clouds obscuring the earth's surface are also discarded.

[22] See Appendix Table A 2.3 for a definition of all variables used in our empirical analysis. Table A 2.3 also contains information on the sources from which we obtained the data (including the year the data stem from). Appendix Table A 2.4 provides descriptive statistics. Appendix Table A 2.5 shows the number of regions by country and variable.

[23] We use nightlight data with a resolution of 30 arc second grids.

Background noise was coded with values of zero.[24] Nighttime luminosity is measured by integer values, which range from 0 to 63 in digital numbers (DN) proportional to light radiance.[25] Figure 2.2 shows that nightlight luminosity differs substantially across regions in our sample, both at the level of pixels and regional averages. Average nightlight luminosity is 2.7 DN with a standard deviation of 6.4 DN; values range from 0.008 in West Botswana, which contains the Kalahari Desert, to 33.37 DN in Harare, the capital of Zimbabwe (see also Appendix Table A 2.4).[26] This implies that Harare emits about 4,000 times as much light as West Botswana. These dramatic differences in luminosity are partially due to regional differences in population. If we correct for differences in population and compare luminosity per capita, Karas in Namibia is the brightest region and the North of Uganda emits the least light, but the ratio between the two reduces to 567:1.The observed values of luminosity in Sub-Saharan Africa are low even compared to other middle-income and low-income countries, where country-level luminosity typically ranges between 3 and 5 DN (Henderson, Storeygard & Weil 2012).[27] Judging from worldwide comparisons of luminosity, Sub-Saharan Africa is one of the world's least developed regions.[28]

To check the robustness of our results, we alternatively use GDP per capita, a more standard measure of economic development, measured in PPP 2005 dollars and taken from Gennaioli et al. (2013). A drawback of using this measure is that these data do not cover Botswana, Mauritius, the Seychelles, and Zanzibar, reducing our GDP sample to 90 regions in eleven countries. Appendix Table A 2.4 shows that regional differences in GDP per capita are even more pronounced than differences in luminosity per capita. In fact,

[24] Some areas may emit low-intensity man-made luminosity, such as camp fires, which was wrongly treated as noise.

[25] This seemingly arbitrary, linear scale comes from averaging overlapping pixels Henderson, Storeygard & Weil 2012).

[26] This reveals that censoring nighttime luminosity at a value of 63 DN is not an issue in our setting.

[27] In poor, scarcely populated countries like Mozambique, more than 99 percent of pixels are unlit. However, also in highly developed countries a large fraction of pixels is unlit. For instance, in the United States and Canada, 69.3 percent and 93.9 percent of pixels, respectively, are unlit (Henderson, Storeygard & Weil 2012).

[28] The low level of luminosity indicates low economic development in Sub-Saharan Africa, matching other studies: Cervellati, Esposito & Sunde (2017) report 2.59 DN for the whole of Africa and Pfeifer, Wahl & Marczak (2017) report 3.12 DN for South Africa (we obtain a value of 4.06 DN for South Africa).

GDP per capita in the richest region (South Africa's Gauteng Province) is 2,500 times higher than GDP per capita in the poorest region (Zimbabwe's Mashonaland West).

Based on our entire sample, there is a surprisingly low correlation between luminosity per capita and GDP per capita of just 0.40. At first glance, this seems to challenge the idea that both nighttime luminosity and GDP measure the level of economic development. However, Appendix Figure A 2.2 shows that this low correlation is driven just by the ten regions in Zimbabwe, which are much brighter, given their GDP per capita values. This outlier has been detected previously (Ghosh et al. 2010). According to Roger (2018), the discrepancy between luminosity and GDP in Zimbabwe might be explained by severe political tensions, including wide-spread government violence, which took place in Zimbabwe in the 2000s.[29] When excluding Zimbabwe, luminosity per capita and GDP per capita are strongly correlated (r=0.78), with no apparent outliers (right panel in Appendix Figure A 2.2).[30]

2.1.3 Control variables

Population and area

To account for the fact that more light tends to be emitted when people are more clustered, all regressions control for a region's population. All estimations also control for the region's area as nighttime luminosity, holding constant the population, is higher when these people live in a smaller area. We extract the total number of people (in ten-thousands) for the reference year 2005 from the gridded Africa Continental Population Datasets (CIESIN 2016). A region's area is measured in arc degrees directly using our geoprocessing software.[31] We also control for population density (i.e., a region's population divided by its area) as robustness check.

[29] Roger (2018) observed a similar disagreement between GDP and luminosity in Burundi and the Democratic Republic of Congo, which are other conflict-torn countries not included in our sample. For instance, warfare may generate substantial amounts of luminosity, while decreasing regular economic activity.

[30] In Appendix Table Table A 2.11 Panel B, we show that our results are not sensitive to excluding Zimbabwe (or any other country) from the sample.

[31] We use the WGS 1984 coordinate system based on the units of arc degrees. Our resolution of 30 x 30 arc second cells corresponds to a surface of 860 square kilometers at the equator. Wrapping a grid net around the globe results in the largest grids at the equator with decreasing grid size approaching the poles and reaching zero.

Population density varies widely across countries in Sub-Saharan Africa: Namibia is one of the countries with the fewest inhabitants per square kilometer in the world, while Mauritius is one of the countries with the highest population density worldwide. The huge variation in population density can also be seen when comparing the regions in our sample (see Appendix Table A 2.4). Cross-regional differences in population and area are smaller, but also substantial.

Quantitative measures of human capital

Our main quantitative measure of human capital is average years of schooling. Gennaioli et al. (2013) derived years-of-schooling information from household surveys, Integrated Public Use Microdata Series (IPUMS), and the Demographic and Health Surveys (DHS), which asked the household head about primary schooling and further education.[32] On average, the population aged 15 years and older has attended school for 4 years (see Appendix Table A 2.4), which is very low compared to developed countries and other developing countries.[33]

Alternative quantitative measures of human capital are primary and secondary school enrollment, measured as net or gross enrollment ratio provided by the Sub-national African Education and Infrastructure Access Data of the CCAPS Strauss Center, based on the household surveys DHS, Multiple Indicator Cluster Surveys (MICS), and UNICEF SA (South Africa). In these surveys, household members report whether they attended primary or secondary school during the past year and the grade they attended during the survey year. Net enrollment is defined as the share of children in primary-school age (secondary-school age) attending primary (secondary) school. Gross enrollment is defined as the number of persons attending primary (secondary) school relative to the number of children in primary-school (secondary-school) age. Net enrollment is capped at 100 percent, while gross enrollment may exceed 100 percent. In fact, average primary gross enrollment in our sample is 107.7 percent, reflecting frequent grade repetitions in Sub-Saharan Africa (Appendix Table A 2.4). Average primary net enrollment is 80.4 percent, indicating that the Millennium Development Goal of universal primary

[32] The authors used the available household survey closest to 2005.

[33] In the worldwide sample of Gennaioli et al. (2013) covering 110 (developed and developing) countries, the average number of years of schooling is 7.1.

enrolment is still not accomplished in Sub-Saharan Africa.[34] Other quantitative education measures are the share of the population with a college degree (Gennaioli et al. 2013)[35] and the literacy rate of adults, which measures basic education (Gershman & Rivera 2018).[36]

Appendix Table A 2.7 shows the correlations between the various human capital measures. Unsurprisingly, math test scores are positively correlated with all quantitative human capital measures. However, the strength of the correlation varies quite considerably. The correlation is strongest with the population share holding a college degree ($r = 0.35$), years of schooling ($r = 0.28$), and literacy rate ($r = 0.28$). Correlations of math test scores with primary and secondary net enrollment are small and statistically insignificant.

Geography

We use the common spatial measures latitude (distance to equator) and longitude (distance from prime meridian). Development theory links closer proximity to the equator with lower economic wellbeing due to adverse living conditions in the tropics (Gallup, Sachs & Mellinger 1999; Easterly & Levine 2003). A second reason for conditioning on a region's latitude arises from geoprocessing the data. When projecting a map of the globe on a two-dimensional coordinate system, the number of pixels per area changes when moving away from the equator.[37]

Moreover, access to the sea may facilitate trade and thus affect development. Therefore, we create a landlocked dummy, which equals 1 for landlocked regions and 0 for regions with access to the sea. To do so, we use information from Gershman and Rivera (2018), but add Lesotho, Mauritius, and the Seychelles, which are missing in their data.

[34] Enrollment data are not available for Botswana and Zanzibar (see Appendix Table A 2.5).

[35] The measure refers to the population aged 15 years and older with ISCED level 5 or 6. Data refer to the survey year closest to 2005.

[36] The literacy rate is the share of adults aged 15 to 49 years who are able to read at least part of a standard sentence or have attended secondary school. Data refer to 2010. The data collected by Gershman & Rivera (2018) do not include Lesotho, Mauritius, and the Seychelles, reducing the sample to 90 observations (see Appendix Table A 2.5).

[37] In our sample, most countries lie south of the equator and thus have negative values for latitude. Since we estimate a log-log specification (see Section 2.2), we use the absolute distance to the equator to measure latitude.

Finally, we also create a dummy variable indicating whether a region contains a country's capital to control for the fact that regions enclosing capital cities (which are often relatively rich and densely populated) may emit more light and also have higher-skilled population, either due to higher-quality education or due to migration.

Nature

Another determinant of economic development in Sub-Saharan Africa are natural endowments. For instance, terrain ruggedness was found to have affected development in Africa because of historic slave raids. Rugged terrain offered protection from slave trade and this advantage persists until today (Nunn & Puga 2012).[38] We use Nunn and Puga's measure of terrain ruggedness and divide by the region's area, following the authors' suggestion.

Second, mineral resources are frequently discussed as an obstacle to development (Sachs & Warner 2001). Lacking data on mineral deposits at the regional level, we use the number of mineral facilities from the U.S. Geological Survey (Eros & Candelario-Quintana 2006).[39] We divide the number of mineral facilities by a region's area to obtain a density measure. As the prevalence of mineral facilities, in contrast to deposits, is potentially endogenous to the level of economic development, adding this variable might lead to an attenuation bias in the estimated coefficient on cognitive skills.

Third, we capture protection against climatic hazards by the CCAPS Strauss Center's Climate Security Vulnerability Model (version 3.0), which covers the average of a number of climatic phenomena between 1970 and 2011 (Busby et al. 2013). Hazards include floods, rainfall anomalies, chronic water scarcity, coastal elevation, tropical cyclones, or wildfires.

Fourth, agriculture is still an important sector of the economies in Sub-Saharan Africa, with an average share of agriculture in GDP in 2007 of about 14.3 percent in Sub-Saharan Africa, as opposed to only 1.5 percent in the European Union (World Bank 2018c). All countries in our sample have a higher share of agriculture in the GDP than the European

[38] In contrast, even terrain enabled trade and productive activities in most other parts of the world.

[39] Mineral facilities include mines, plants, mills, and refineries of aluminum, cement, coal, copper, diamond, gold, iron and steel, nickel, platinum-group metals, salt, and silver.

average (up to 35.3 percent in Kenya); see Appendix Table A 2.2. Due to the importance of agriculture, we include land suitability for rain-fed agriculture collected by Gershman and Rivera (2018) as another control variable.[40]

Finally, we add a measure of average temperature in a region during the period 1950 to 2000 in degrees Celsius as collected by Gennaioli et al. (2013) based on the WorldClim database. The average temperature is about 20 degrees, the coldest region being the mountains of Mokhotlong in Lesotho with 8 degrees, and the warmest region is Port Louis in Mauritius with 27 degrees.

Health

A basic factor fostering development is child health and nutrition. Despite considerable improvements in recent years, child nutrition was still poor in the mid-1990s (when most students in our sample were born).[41] To capture child malnutrition, we use the average share of underweight children below five years between 1990 and 2002 from the Center for International Earth Science Information Network (CIESIN) (2017). Underweight is defined as having a weight-for-age z-score that is more than two SD below the median of the international reference population. To capture medical care in a region, we use the share of home births relative to births delivered at a medical facility, taken from Gershman and Rivera (2018).

Fractionalization

Africa is well-known for its ethno-linguistic diversity, which we capture by the number of different tribes in a region, collected by Nunn & Wantchekon (2011) based on Murdock's ethnic atlas from 1959. Second, we use region-level data by Gershman and Rivera (2018) on ethno-linguistic fractionalization, considering the distance between groups, and ethnolinguistic polarization, considering the conflict potential through the relationship of the dominant group to minorities. Gershman and Rivera (2018) argue that the existence of a sizable ethnic minority alongside the dominant group substantially increases the likelihood of ethnic conflict. Furthermore, we use their measures of religious fractionalization and religious polarization, which distinguish between Christianity, Islam, "traditional" religion, and no religion. A third fractionalization indicator is conflict. We use the count of deadly fatalities in 2007 due to conflict, as reported by the Armed

[40] The variable is coded on a scale from 1 (highest suitability for agriculture) to 8 (lowest suitability).

[41] The average age of SACMEQ students tested in 2007 is 13.5 years, implying a birth year of 1993 or 1994.

Conflict Location & Event Data Project (ACLED) (Clionadh et al. 2010). Regions without deadly conflicts are assigned a value of zero.

Preferences

To measure cross-regional differences in preferences, we use data on trust, patience, risk aversion, and altruism elicited by means of survey questions and behavioral experiments by Falk et al. (2018). However, preference measures are available only in 51 out of 112 regions, so we present results in the Appendix.

2.2 Empirical model

As a starting point to derive our empirical specification, we use a Cobb-Douglas production function (Mankiw, Romer & Weil 1992) with multiple inputs:

$$Y = H\beta X\alpha' \qquad (2.1)$$

Here, Y is the output, H is human capital, and $\boldsymbol{X}$ is a vector containing other determinants of productivity (specified below). Applying a logarithmic transformation of both sides of equation (2.1), we obtain

$$\ln(Y) = \beta \ln(H) + \ln(X)\,\alpha' \qquad (2.2)$$

Human capital, H, is of primary importance for us. Given existing research that cognitive skills of the population are more important for economic development than quantitative measures of human capital (Hanushek & Woessmann 2008), our baseline model includes student test scores as a proxy for the cognitive skills of the adult population. To investigate the importance of qualitative *vis-à-vis* quantitative measures of human capital, we also add years of schooling and alternative attainment measures (such as primary and secondary enrollment) in some specifications.

Based on equation (2.2), we estimate the following log-log model at the regional level: [42]

$$\ln(light_{rc} + 0.01) = \beta_1 \ln(test\ scores_{rc}) + \beta_2 \ln(population_{rc}) + \beta_3 \ln(area_{rc}) + \ln(X_{rc})\,\alpha' + \lambda_c + \varepsilon_{rc}, \qquad (2.3)$$

where $light_{rc}$ is the nighttime luminosity in region r in country c; $test\ scores_{rc}$ is the math (or reading) test score from the SACMEQ student assessment, a proxy for the cognitive skills of the adult population; $population_{rc}$ is the population in 10,000 inhabitants and $area_{rc}$ the size of the region in arc degrees. Note that all specifications condition on population and area because more densely populated areas, *ceteris paribus*, tend to have higher nighttime luminosity. The vector $\boldsymbol{X}_{rc}$ contains several other region-level determinants of economic development, such as geography, nature, health, and regional conflicts. Due to data limitations, we cannot explicitly account for regional differences in physical capital, which is included in most cross-country growth models. However, Acemoglu & Dell (2010) point out that differences in physical capital across regions within countries are unlikely to be a primary determinant of economic differences because of the relatively free mobility of capital within national boundaries. ε_{rc} is an idiosyncratic, region-specific error term.

We follow the literature in adding 0.01 to nighttime luminosity since reported absence of luminosity does not reflect actual darkness and underreporting in low-lit areas may be a relic of the nightlight data-generation process, which assigns a value of zero to background noise (Michalopoulos & Papaioannou 2013; Hodler & Raschky 2014; Michalopoulos & Papaioannou 2014; Bruederle & Hodler 2017; Bitzer & Gören 2018; Gershman & Rivera 2018).[43] We also add 0.01 to explanatory variables that have values of zero before applying the logarithmic transformation.[44]

[42] Taking the logarithm of nighttime luminosity implies that the strongly right-skewed distribution becomes more similar to a normal distribution (see Appendix Figure A 2.3). Using a log-linear specification, which is derived from the individual-level Mincer earnings function (e.g., Hanushek, Ruhose & Woessmann 2017), leads to qualitatively similar results. See Appendix Table A10.

[43] However, man-made luminosity could have just been too low for satellites to detect. Henderson, Storeygard & Weil (2012) note that even highly developed countries have unlit pixels.

[44] Note that we do not log-transform binary variables such as indicators for landlocked regions or capital regions.

By including country fixed effects, λ_c, we relate differences in nighttime luminosity to differences in human capital only across regions within the same country. Among others, country fixed effects control for differences in national educational institutions (e.g., duration of compulsory schooling) and national economic factors such as sectoral composition.[45] Moreover, country fixed effects account for differences in national institutions, in particular, openness of the economy such as a common customs system and quality of property rights due to national laws (Acemoglu, Johnson & Robinson 2001). Within-country estimates also have the advantage that they are not affected by cross-country cultural differences in nighttime versus daytime activities, public versus private lighting, and national conditions for generating electricity.

Despite including country fixed effects and a rich set of control variables, we shy away from interpreting the coefficient on test scores, β_1, causally. One reason why β_1 might be biased is reverse causality. For instance, regions with more economic resources might be able to produce students with higher math and reading achievement, which would bias β_1 upward. However, existing research suggests that in our setting reverse causality is rather unlikely given that the most convincing evidence from randomized interventions shows that resources have, at best, small effects on student achievement (for a survey, see Ganimian & Murnane 2016).[46] The coefficient β_1 might also be biased because of omitted variables at the regional level that are correlated with both economic development and human capital. For example, the quality of property rights may vary across regions because of clans with traditional jurisdiction; the industry specialization may also be region-specific. Finally, β_1 might be biased if individuals move across regions between sixth grade, that is, when cognitive skills are assessed, and labor-market entry. This might lead to an upward bias, for example, if high-skilled individuals move to more developed regions after school. At the same time, β_1 might be downward biased since cognitive skills tested in sixth grade is an error-ridden measure of the true skills of the current workforce. Thus, the direction of bias in β_1 is not clear *a priori*.

[45] See Appendix Table A 2.1 and Table A 2.2 for differences in the education systems and industry specialization across our sample countries.

[46] Moreover, we also run our baseline specification with luminosity in 2013 (i.e., the latest available year in our data) as an outcome, while still using test scores in 2007 as the main explanatory variable. If luminosity is not perfectly correlated over time, this should reduce a potential reverse causality problem. Reassuringly, results are very similar as to our main specification (see Appendix Table A 2.8).

2.3 Results

2.3.1 Main results

Table 2.1 reports estimates of the association between nighttime luminosity and math test scores based on the model in equation (2.3). All specifications include country fixed effects. Throughout our analysis, we cluster-bootstrap standard errors at the country level (50 repetitions).[47]

The results in Table 2.1 show a strong positive within-country association between luminosity and math skills. When controlling only for country fixed effects, a one-percent increase in math test scores is associated with a whopping 12.9 percent increase in luminosity (Column 1). However, a considerable part of this correlation is due to the fact that both nighttime luminosity and test scores are higher in more populated areas. When we additionally control for population, the coefficient on math test scores decreases by more than half (Column 2). Population is strongly positively related to luminosity, with an elasticity close to one. In Column 3, we replace population by area, leading to a slight increase in the test-score coefficient; however, it remains substantially smaller than in the unconditional within-country specification. The coefficient on area is negative and close to one, indicating that nighttime luminosity is lower in larger regions.

In Column 4 of Table 2.1, we simultaneously include population and area. This constitutes our main specification. We find that a one-percent increase in math test scores is associated with a 5.8 percent increase in luminosity. To assess the magnitude of the test-score coefficient, we translate the percentage change into a change in absolute values.

[47] Recent research has shown that clustered standard errors can be biased downward in samples with a small number of clusters (for example, Donald & Lang 2007, Cameron, Gelbach & Miller 2008, Angrist & Pischke (2009), and Barrios et al. (2012). Although there is no widely accepted threshold when the number of clusters is "small," the work of Cameron, Gelbach & Miller 2008, Angrist & Pischke (2009), and Harden 2011) suggests a cutoff of around 40 clusters. Due to the small number of countries in our sample, we cluster bootstrap standard errors (using Stata's bootstrap command for implementation). Appendix Table A 2.9 provides the results using Stata's standard sandwich estimator clustered at the country level. As expected, standard errors decrease somewhat compared to the bootstrapping procedure.

Evaluated at the mean, one percent increase in test scores amounts to roughly 5.1 SACMEQ points (see Appendix Figure A 2.4). The difference in SACMEQ scores between the best-performing region (the South and Curipe in Mauritius) and the worst-performing region (the South of Zambia) is 234 points or 45.9 percent. Thus, if the population in the South of Zambia had math skills similar to those of the people in the South and Curipe in Mauritius, regional luminosity would increase by 45.9 * 5.8 percent = 266.2 percent or 7.3 DN. This increase would close the luminosity gap between these two regions by about 44 percent.[48] Similarly, if the region at the 25th percentile of the skill distribution (Omaheke in Namibia; 468 points) would increase the average math skills of its population to the level of the region at the 75th percentile (Gauteng Province in South Africa; 545 points), regional luminosity would increase by 15.1 * 5.8 percent = 87.6 percent or 2.4 DN. Such increase would close the luminosity gap between these two regions by roughly 10 percent.[49] The strength of the estimated relationship between luminosity and math test scores suggests that the cognitive skills of the population are a highly relevant determinant of economic performance in Sub-Saharan Africa.

In our baseline specification in Column 4 of Table 2.1, the negative coefficient on area becomes much smaller in absolute value when population is accounted for, which is due to the fact that larger regions in our sample are often less populated (e.g., deserts). Instead of controlling separately for population and area, we can also combine both measures to population density (Column 5). Population density is strongly correlated with luminosity; the estimated elasticity suggests a half-percent increase in luminosity when population density increases by one percent. However, note that the test-score coefficient barely changes.

Figure 2.3 depicts the relationship between luminosity and math test scores graphically. In Panel (a), only country fixed effects are purged from the variables, thus showing the variation we are using in our empirical analysis. Panel (b) shows the relationship between nighttime luminosity and math skills in our main specification. Both graphs suggest that the relationship between luminosity and test scores is not driven by outliers. Appendix

[48] Luminosity in South and Curipe is 16.5 DN vs. 0.1 DN in South of Zambia.

[49] Gauteng Province has a level of luminosity of 23.2 DN, Omaheke a level of 0.1 DN.

Figure A 2.5 additionally shows that nighttime luminosity and math test scores are positively associated within each country in our sample.[50]

2.3.2 Additional education measures

Thus far, we investigated the relationship between nighttime luminosity and math test scores, a *qualitative* measure of human capital and a proxy for the cognitive skills of the population. In this section, we assess the relevance of cognitive skills for economic development when we additionally control for various *quantitative* measures of human capital.

In Table 2.2, we add seven different measures for the quantity of human capital in a region, which reflect different aspects of the education system.[51] For comparison, in Column 1 we present our baseline specification for the subset of 90 regions for which we have information on years of schooling, the most commonly used quantitative measure of human capital. Adding years of schooling in Column 2 reduces the coefficient on test scores, which is not surprising given that the two human-capital measures are positively correlated ($r = 0.28$; see Appendix Table A 2.7).[52] However, both measures are statistically significant and economically meaningful. This suggests that test scores and years of schooling have a large amount of independent variation (which is not surprising, given that the correlation is far smaller than one). Strikingly, the coefficient on test scores is four times larger than the coefficient on years of schooling, emphasizing the relevance of cognitive skills for economic development.[53]

[50] In Appendix Figure A 2.5, we excluded all countries with fewer than five regions (Malawi, Swaziland, and Uganda).

[51] See Appendix Table A 2.1 for the duration of primary schooling, secondary schooling, and compulsory education in our sample countries.

[52] The cross-regional correlation between test scores and years of schooling is considerably smaller than the cross-country correlation (0.28 with p-value 0.01 opposed to 0.32 with p-value 0.3). This possibly reflects substantial differences across regions in the quality of education, leading to differences in the learning progress made during one school year, or in the importance of education in the society. When using only within-country cross-regional variation, the correlation between test scores and years of schooling is much larger at 0.47.

[53] When replacing math test score by years of schooling in our baseline specification, the coefficient is also substantially smaller than the respective coefficient on test scores: a one-percent increase in years of schooling is related to an increase of 0.443 percent in luminosity.

In Columns 3 and 4, we control for primary gross enrollment and primary net enrollment, respectively, two measures of school enrollment at the early stage of schooling. While the coefficient on primary gross enrollment is insignificant, the coefficient on primary net enrollment is even larger than the respective coefficient on years of schooling in Column 2 (but much less precisely estimated). The estimated elasticity suggests that increasing net enrollment by one percent increases luminosity by 1.5 percent. The difference in the estimated relationship with luminosity between both enrollment rates results from gross enrollment including age-inappropriate students, suggesting that education systems with many students beyond primary-school age (e.g., due to class retention) produce less human capital that is relevant for economic development.

In Columns 5 and 6, we use secondary gross and net enrollment, respectively. The point estimates are very similar to the years-of-schooling estimate, and both are statistically significant. Resembling the pattern for primary enrollment, the coefficient on net enrollment is somewhat larger than the coefficient on gross enrollment. However, the coefficients on secondary enrollment being much more similar than the respective coefficients on primary enrollment can be explained by the fact that gross enrollment and net enrollment rates are more strongly correlated for secondary education (0.94) than for primary education (0.74).

In Column 7, we include the share of the regional population with a tertiary degree, reflecting higher education participation. While the share of tertiary-educated individuals is significantly related to luminosity, the estimated elasticity is smaller than for primary and secondary (net) enrollment. This suggests that basic education is somewhat more important than tertiary education for economic development in Sub-Saharan Africa (see also Easterly & Levine 1997; Petrakis & Stamatakis 2002).[54] Column 8, which includes the literacy rate as a very basic measure of human capital (defined as the share of individuals who are able to read simple sentences), strengthens this interpretation: The estimated

[54] Note that the coefficient on test scores decreases considerably when we account for the share of tertiary educated in the population. Supposedly, this is due to the fact that the share of tertiary educated more closely reflects the quality of education since it refers to a completed degree and not just to (school) attendance.

elasticity of the literacy rate is about 40 percent larger than the elasticity of the share of tertiary educated.[55]

Across specifications, the coefficient on math test scores remains sizeable and substantially larger than the coefficients on the various measures of human capital quantity. These results are broadly in line with previous work investigating the relationship between economic development and education quality *vis-à-vis* education quantity at the country level (e.g., Hanushek & Woessmann 2008). However, in contrast to existing country-level results, our measures of education quantity remain significant predictors of regional development in Sub-Saharan Africa when the quality of education is accounted for.[56] One potential reason for the differing results is that Hanushek & Woessmann (2008) estimate cross-country growth models, assessing the importance of cognitive skills versus years of schooling for a country's growth rate. In contrast, we assess differences in economic development across regions within countries in a cross-section. Moreover, test scores may be measured with more error at the regional level as compared to the country level, leading to an attenuation bias in the test scores estimate.[57] Furthermore, we focus entirely on countries at a very low stage of economic development, while the sample of Hanushek & Woessmann (2008) also includes many middle-income and high-income countries.

There are various reasons why educational attainment is still relevant for economic development even when conditioning on test scores. Attainment measures may proxy for components of human capital that are important for economic development, for example, non-cognitive skills, which are not captured by test scores (see also Hanushek et al. 2015). Educational attainment may also affect other domains of cognitive skills that are not tested in the math and reading assessments of SACMEQ. However, in terms of magnitude, cognitive skills, as measured by test scores, are considerably more relevant

[55] Since we always condition on cognitive skills, the described differences in the elasticities of the various quantitative human capital measures are not straightforward to interpret, as they partly depend on the correlation between the respective measure and cognitive skills. However, the pattern of results is very similar when cognitive skills are not included in the estimation (not shown).

[56] When we aggregate our regions to the country level, regressions yield very similar results (not shown).

[57] However, we consider the measurement-error explanation less likely, given that we observe on average test scores of almost 1,300 students in a region; at a minimum, we observe 354 students per region (see Appendix Table A 2.6).

than all sorts of quantitative human capital measures in explaining cross-regional differences in economic development in Sub-Saharan Africa.

2.3.3 Other determinants of development

The literature has identified several determinants of economic development beyond human capital. While we are careful in not making any causal claims, we nevertheless seek to investigate whether the estimated coefficients on cognitive skills pick up the impact of other determinants of development. The presentation of the results (in Table 2.3 to 2.6) always follows the same structure: for each group of determinants, we first add each factor individually to our baseline specification, which contains math test scores as well as population and area.[58] In the last column of each table, we include all determinants of a group simultaneously.

In Table 2.3, we add various controls for geography: longitude (distance to prime meridian), latitude (absolute distance to equator), and binary indicators for whether regions are landlocked (i.e., have no access to the sea) or contain the country's capital. Except for a significantly positive relationship between capital regions and economic development, none of the geographical controls is significantly related to luminosity. The coefficient on math test scores remains significant, even when all geographical controls are included simultaneously (Column 5). To check more thoroughly whether our results are driven by densely populated regions in a country, Appendix Table A 2.11 excludes capital regions (Column 1) and the two most densely populated regions (Column 2), respectively, in each country. The test scores coefficient is significant and sizeable in both specifications.

In Table 2.4, we include several features of a region's nature, that is, terrain characteristics, natural resources, and climatic conditions. In particular, we use terrain ruggedness, the presence of mineral facilities, protection against climatic hazards, land suitability for agriculture, and temperature. Only the number of mineral facilities is significantly related to economic development. The positive and significant relationship with economic development is hardly surprising, since the prevalence of mineral *facilities*

[58] Note that across specifications in Tables 2.3 to 2.6, the coefficient on population is significantly positive and the coefficient on area is (almost always) significantly negative, as in the baseline specification.

(not deposits) likely depends on the level of economic development. The coefficient on test scores is largely unaffected by adding the environmental controls. Only in the specification with all controls included simultaneously (Column 6), the test scores coefficient decreases somewhat in size.[59]

Healthcare provision is often insufficient in developing countries, which may hamper economic development. In Table 2.5, we control for infant underweight and home births (*vis-à-vis* hospital births), two important indicators of healthcare quality. Both coefficients show the expected negative relationship with luminosity, but none is statistically significant. The coefficient on math test scores remains sizeable in all specifications, but turns statistically insignificant once we control for both health measures simultaneously (Column 4).[60]

Table 2.6 adds several indicators of fractionalization as measures of regional conflicts. The number of different tribes in a region is not significantly related to luminosity, while the number of conflict fatalities is (albeit small in magnitude). The latter result may be due to the fact that warfare in itself is an activity that generates substantial amounts of luminosity (Roger 2018). Ethno-linguistic fractionalization and religious fractionalization are also significantly positively related to luminosity. However, both fatalities and religious fractionalization lose significance in the full-control model, while ethno-linguistic polarization becomes significantly negative (Column 7). Math test scores remain a strong and statistically significant predictor of luminosity across all specifications.

Since we do not have sufficient statistical power to include all individual determinants from Table 2.4 to 2.6 simultaneously, we compute the first principal component of each group of determinants and include these components alongside years of schooling, the

[59] In Column 6 of Table 2.4, estimates are based on only 80 regions because we have no data on land suitability for Lesotho (reducing the sample compared to Column 4) and no data on temperature for Botswana and Zanzibar (reducing the sample compared to Column 5).

[60] The results that control for home births have to be interpreted carefully, as home birth data are available mainly for the years after 2007 (up to 2013, depending on the country). This potentially renders home births an endogenous control.

most widely-used quantitative measure of human capital.[61] In Table 2.7, when the first principal components are included individually, only years of schooling (positive), poor health care (negative), and the degree of fractionalization (positive) are significantly related to luminosity. When these determinants are included simultaneously, only the coefficient on years of schooling is statistically significant (Column 6). Importantly, math test scores retain a significantly positive coefficient even in this very demanding specification. Overall, these findings show that the relationship between math test scores and luminosity remains robust even when accounting for many other factors of regional development in Sub-Saharan Africa.

2.3.4 Alternative outcomes

In the literature on economic growth, it is standard to use outcome variables in per-capita terms. Column 1 of Table 2.8 shows that our results are very similar when using luminosity per capita instead of absolute luminosity (but still controlling for area). In Column 2 of Table 2.8, we employ a more traditional measure of economic output, namely, real GDP per capita. Results corroborate the importance of cognitive skills for economic development: Test scores are significantly positively related to GDP per capita, although the coefficient decreases somewhat in magnitude.[62] Hence, also when considering a more common output measure, the quality of human capital remains an important determinant of economic development.

2.3.5 Robustness checks

Several robustness checks and specification tests increase confidence in our results. First, we show that our findings are robust to the exact model specification. In Appendix Table A 2.9, we show that our results are not sensitive to the functional form. In Columns 1–4, math test scores, population, and area are included linearly, while the outcome is still in logarithm. In Column 5, both luminosity and math test scores (as well as the other controls) enter linearly. For expositional purposes, math test scores, population, and area are standardized to have mean 0 and SD 1 across countries. Across all specifications, we

[61] Appendix Table A 2.13 shows that the first principal component always explains a reasonable share of variance, ranging from 38 percent (geography) to 64 percent (poor health care). Unsurprisingly, the share of explained variance of the first component tends to decrease with the number of variables in a group.

[62] See Column 1 of Table 2.2 for results on luminosity for the sample with the same 90 regions.

observe a significant and positive relationship between math test scores and luminosity. In the specification in Column 1, an improvement in math test scores by 1 SD (which amounts to 48 SACMEQ points) is associated with an increase in luminosity by 62 percent. Interestingly, this magnitude is very similar to that of the log-log specification in Column 4 of Table 2.1. Recall that the difference in SACMEQ scores between the best-performing region (the South and Curipe in Mauritius) and the worst-performing region (the South of Zambia) is 234 points, or 4.8 SD. Thus, the test scores estimate in the log-linear specification suggests that if the South of Zambia performed as well as the South and Curipe in Mauritius in terms of test scores, regional luminosity would increase by 4.8*62=297.6 percent or 8.1 DN.

Results are similar when we instead use luminosity per capita (Column 2 of Appendix Table A 2.10) and GDP per capita (Column 3) as outcomes. In Column 4, we check for non-linearities in the association between luminosity and math test scores by adding squared test scores. The linear test scores term remains positive and sizeable, while the coefficient on squared test scores is rather small and statistically insignificant. Finally, in Column 5, we use the original, i.e., non-logarithmic, values of all variables. We find that a one-SD increase in math test scores is related to an increase in luminosity by 1.89 DN. This implies an increase in luminosity of 9.1 DN when the worst-performing region in terms of SACMEQ scores would improve the math skills of its population to the level of the best-performing region. These results show that, reassuringly, the implied effect sizes of the log-log, log-linear, and linear-linear specifications are very similar. Thus, irrespective of the precise functional form, the cognitive skills of a region's population are a statistically significant and economically meaningful predictor of economic development.[63]

Next, we check the robustness of our results to several sample restrictions (Appendix Table A 2.11). In Panel A, we show that our results are robust to excluding potential regional outliers. We begin by excluding all capital regions, which typically emit the most

[63] Results are also robust to applying the logarithmic transformation to the original luminosity value (i.e., not adding 0.01 before taking the logarithm).

light (Column 1).[64] In Column 2, we exclude the two most densely populated regions in a country (typically the capital region and the region containing the second-largest city), and in Column 3, we exclude the least densely populated region of each country. In Column 4, we keep only countries whose SACMEQ regions match the official administrative regional boundaries to ensure that our reclassification of regions is not driving the results. In Panel B of Appendix Table A 2.11, we re-estimate our baseline specification when excluding each country individually. Math test scores remain significantly positive and sizeable in all those subsamples.[65] Therefore, we conclude that our main results are not driven by specific regions or specific countries.

In Appendix Table A 2.12, we add four key economic preferences – trust, patience, risk aversion, and altruism – to our baseline specification. However, these results should be interpreted very cautiously, as they are based on only 51 regions. No preference measure is consistently related to luminosity. However, in the specification with all preferences included (Column 6), regions where the population has higher levels of trust (potentially fostering cooperative behavior) experience higher nighttime luminosity, whereas altruism is negatively related to luminosity. The test scores coefficient in this 51-region sample is only marginally significant and decreases substantially in size compared to the full sample (see Column 1). While adding the preference measures does not reduce the test scores coefficient, it turns insignificant in most specifications due to larger standard errors.

Finally, we find a strong positive association between the reading skills of the population and nighttime luminosity, using the reading scores from SACMEQ (Appendix Table A 2.14). However, the test score coefficients are consistently somewhat smaller for reading than for math (see Table 2.1), reflecting either a more important role of math skills for economic development or that reading skills cannot be measured as consistently across languages as math skills, implying more measurement error.

[64] Capitals in our sample are: Gaborone in Botswana, Nairobi in Kenya, Maseru in Lesotho, Lilongwe in Malawi, Port Louis in Mauritius, Maputo in Mozambique, Windhoek in Namibia (which is in the region Khoma Highland), Victoria in the Seychelles, South Africa's three capitals in Pretoria, Cape Town, and Bloemfontein (in the regions Western Cape, Gauteng, and Freestate), Mbabane in Swaziland (in the Hhohho region), Dodoma in Tanzania (in the Central region), Kampala in Uganda (in the Central region), Lusaka in Zambia, Zanzibar City in Zanzibar, and Harare in Zimbabwe.

[65] The only substantial drop in coefficient magnitude occurs when Namibia is excluded from the sample.

2.4 Conclusion

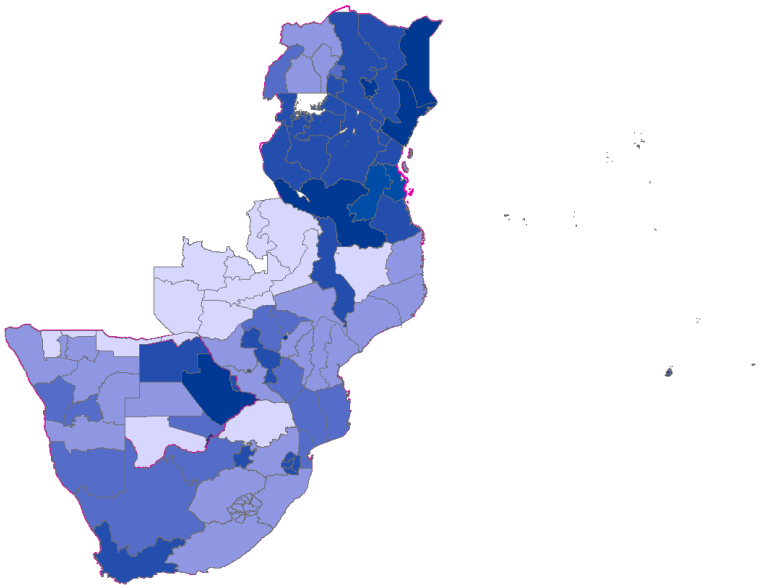
We use cognitive skill data for a large set of subnational regions in Sub-Saharan Africa to investigate the role of human capital for economic development in one of the least developed areas worldwide. Our measure of cognitive skills of the African population is based on test scores from more than 120,000 students, which we aggregate to 112 regions in 15 countries. We use light emissions at night as a measure for economic development, circumventing the problem of low reliability of GDP figures in African countries due to poor statistical capacity and possibly politically motivated misreporting of output indicators.

In within-country cross-regional estimations, we find that a one-percent increase in cognitive skills is associated with an increase in luminosity by 5.8 percent. This magnitude is substantial, as it implies that almost half of the gap in economic performance between the least-skilled region and the most-skilled region in Sub-Saharan Africa could be closed if the least-skilled region could raise the cognitive skills of its population to the level of the most-skilled region. Moreover, the coefficient on cognitive skills is four times as large as the coefficient on years of schooling, the most commonly used (quantitative) measure of human capital. Our cross-sectional estimates are, of course, subject to questions about causality. However, considering a range of alternative influences does not change the pattern of results. In particular, results are robust to including alternative human capital measures and various other determinants of regional development, such as population density, geography, nature, health care, and regional conflicts. Furthermore, the results are not driven by specific regions or countries.

Our results imply that the cognitive skills of the population are an important determinant of economic development in Sub-Saharan Africa. However, despite considerable progress in increasing school enrollment, children in these countries are still learning remarkably little in school. To illustrate this point, Bietenbeck, Piopiunik & Wiederhold (2018) refer to one item from the SACMEQ test, which asked students to calculate the number of pages remaining in a 130-page book after the first 78 pages have been read. Only 30 percent of sixth-grade students participating in SAQMEQ could answer this question. In comparison, two-thirds of fourth-grade students in OECD countries

answered this question correctly. In fact, existing evidence from international student assessments suggests that Sub-Saharan Africa is the region with the lowest cognitive skills worldwide. Previous work in the African context (e.g., Muralidharan & Sundararaman 2011; Duflo, Hanna & Ryan 2012; Bietenbeck, Piopiunik & Wiederhold 2018) highlights the role of teachers in enhancing cognitive skills, suggesting a potentially effective policy to improve the economic performance in one of the poorest regions worldwide.

Figure 2.1: Regional distribution of math test scores

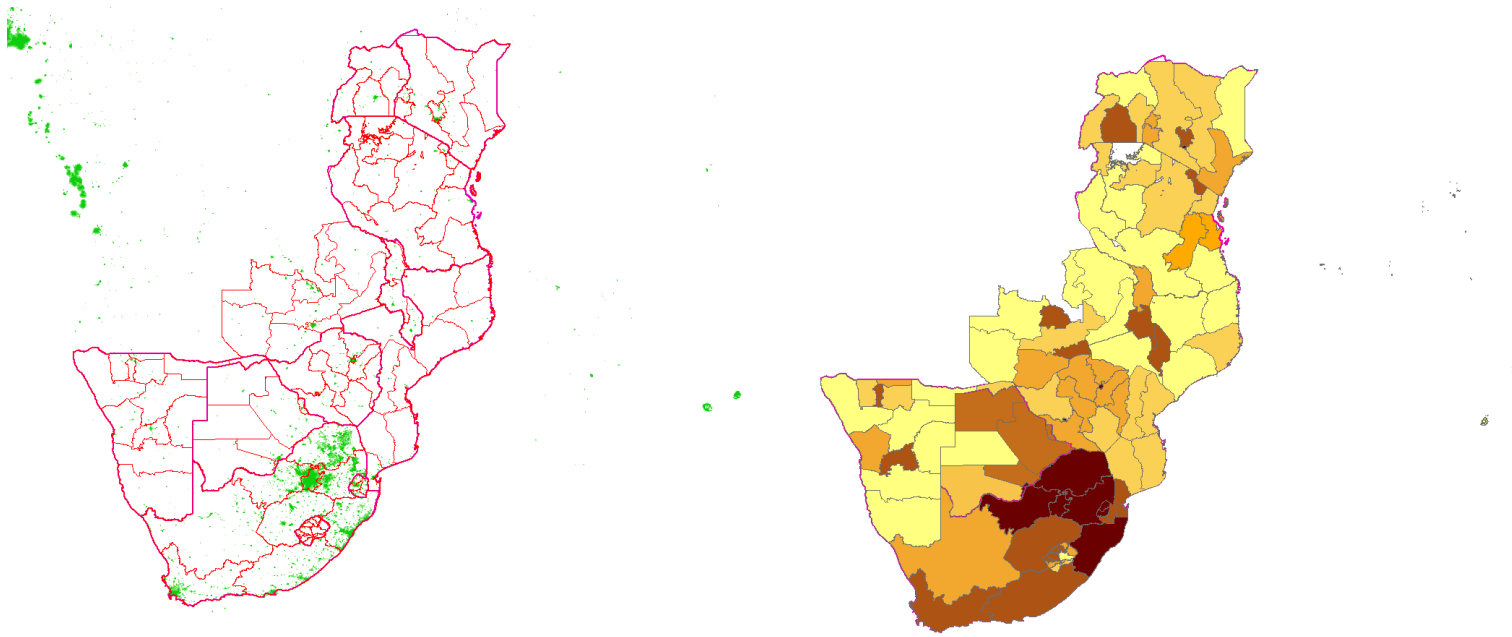


Notes: The map shows average math test scores by region, with colors ranging from light blue (lowest test scores) to dark blue (highest test scores) in five shadings: (i) 1.00 to 1.78 SD below mean, (ii) 0.22 to 1.00 SD below mean, (iii) 0.56 SD above mean to 0.22 SD below mean, (iv) 1.35 to 0.56 SD above mean, and (v) 2.13 to 1.35 SD above sample mean. Data Source: SACMEQ (2007).

Figure 2.2: Regional distribution of luminosity

(a) Pixel-level distribution of luminosity

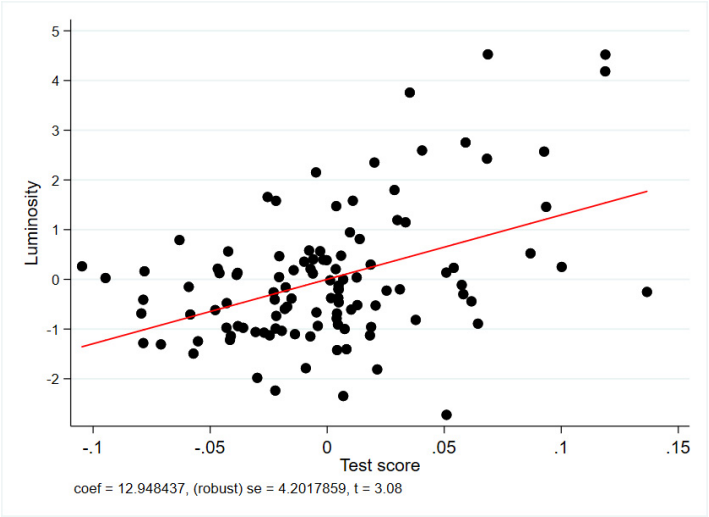
(b) Average luminosity



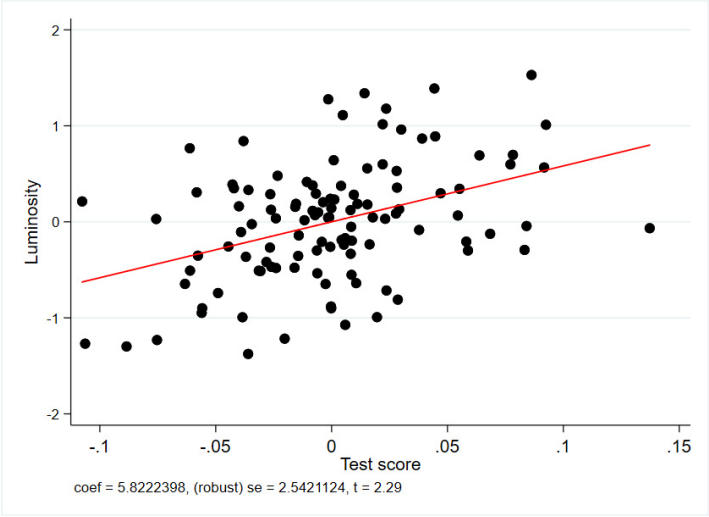
Notes: Panel (a) shows pixel-level luminosity with regional borders indicated by red lines and country borders indicated by thick purple lines. Luminosity is depicted by colors ranging from light yellow to dark brown (lit with a maximum of 63). Unlit areas are in white. Panel (b) shows the average luminosity of a region. Average regional luminosity is depicted in quintiles of the same size by colors increasing from yellow to brown, where darker colors indicate higher luminosity (light yellow is 0.01-0.05 DN, dark yellow is 0.05-0.09 DN, orange is 0.09-0.28 DN, dark orange is 0.32-1.50 DN, and brown is 1.80-33.37 DN). Data Source: V4 DMSP-OLS (NOAA's National Geophysical Data Center 2016) Nighttime Lights Time Series.

Figure 2.3: Relationship between luminosity and math test scores across regions within countries

(a) Within-country variation



(b) Within-country variation conditional on regional population and area



Note: The graph plots our baseline specifications in Table 2.1. To construct the figure in panel (a), we regressed luminosity and math test scores on country fixed effects (see Column 1 of Table 2.1). Panel (b) additionally controls for population and area (see Column 4 of Table 2.1). The solid red line is the linear fit between residualized luminosity and residualized test scores. Both variables are in logarithm. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013).

Table 2.1: Baseline results

	(1)	(2)	(3)	(4)	(5)
Math test score	12.948***	5.539**	7.875***	5.822**	6.259***
	(3.647)	(2.542)	(2.302)	(2.852)	(1.978)
Population		0.972***		0.750***	
		(0.077)		(0.163)	
Area			-1.028***	-0.286*	
			(0.078)	(0.149)	
Population density					0.533***
					(0.027)
Observations	112	112	112	112	112
Adj. R-squared	0.058	0.778	0.689	0.787	0.780

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

Table 2.2: Adding controls for educational attainment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math test score	6.654**	4.907**	6.939**	6.518**	4.485**	5.620***	4.728*	5.964**
	(2.739)	(2.493)	(2.957)	(2.839)	(1.743)	(2.040)	(2.789)	(2.522)
Years of schooling		1.216***						
		(0.222)						
Primary gross enrollment			0.611					
			(0.620)					
Primary net enrollment				1.496*				
				(0.825)				
Secondary gross enrollment					1.119***			
					(0.167)			
Secondary net enrollment						1.249***		
						(0.208)		
Share tertiary educated							0.820***	
							(0.133)	
Literacy rate								1.160***
								(0.336)
Population	0.816***	0.693***	0.802***	0.792***	0.799***	0.738***	0.741***	0.608***
	(0.228)	(0.137)	(0.212)	(0.206)	(0.173)	(0.258)	(0.164)	(0.120)
Area	-0.210	-0.198	-0.202	-0.172	-0.085	-0.129	-0.181	-0.354***
	(0.196)	(0.156)	(0.172)	(0.211)	(0.212)	(0.264)	(0.187)	(0.116)
Observations	90	90	90	90	90	90	90	90
Adj. R-squared	0.778	0.856	0.779	0.794	0.852	0.843	0.840	0.849

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017), CCAPS Sub-national African Education and Infrastructure Access Data (2014), Gershman and Rivera (2018).

Table 2.3: Adding controls for geography

	(1)	(2)	(3)	(4)	(5)
Math test score	5.677**	5.319**	5.665**	4.710*	3.995*
	(2.694)	(2.393)	(2.417)	(2.511)	(2.366)
Longitude	-0.362				-0.493
	(1.292)				(1.302)
Latitude		0.327			0.312
		(0.250)			(0.381)
Landlocked			-0.062		-0.035
			(0.701)		(0.183)
Capital				0.513***	0.489***
				(0.191)	(0.190)
Population	0.759***	0.776***	0.759***	0.684***	0.728***
	(0.204)	(0.157)	(0.128)	(0.158)	(0.187)
Area	-0.280	-0.288**	-0.282**	-0.304**	-0.296*
	(0.171)	(0.134)	(0.116)	(0.136)	(0.170)
Observations	112	112	112	112	112
Adj. R-squared	0.785	0.790	0.785	0.800	0.799

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables except Landlocked and Capital are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017), Gershman and Rivera (2018).

Table 2.4: Adding controls for nature

	(1)	(2)	(3)	(4)	(5)	(6)
Math test score	5.717**	5.688***	5.729**	5.824***	6.611**	4.725*
	(2.371)	(1.933)	(2.574)	(2.094)	(2.849)	(2.497)
Ruggedness	0.099					-0.047
	(0.105)					(0.157)
Mineral facilities		0.079**				0.110*
		(0.032)				(0.064)
Protection against climatic hazards			0.840			0.403
			(1.343)			(1.604)
Land suitability for agriculture				0.441		0.464
				(0.313)		(0.393)
Temperature					-0.138	-0.803
					(1.065)	(1.122)
Population	0.737***	0.713***	0.756***	0.730***	0.820***	0.646***
	(0.196)	(0.150)	(0.129)	(0.106)	(0.187)	(0.223)
Area	-0.306*	-0.292**	-0.276**	-0.333***	-0.203	-0.369
	(0.182)	(0.141)	(0.115)	(0.094)	(0.167)	(0.243)
Observations	112	112	112	90	90	80
Adj. R-squared	0.788	0.803	0.787	0.797	0.775	0.818

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017), Nunn & Puga (2012), Eros & Candelario-Quintana (2006), CCAPS Sub-national African Education and Infrastructure Access Data (2014), Gershman and Rivera (2018).

Table 2.5: Adding controls for health

	(1)	(2)	(3)	(4)
Math test score	6.437**	5.929***	5.146*	4.679
	(2.676)	(2.148)	(3.010)	(3.161)
Infant underweight		-0.392		-0.381
		(0.303)		(0.495)
Home births			-0.249	-0.244
			(0.274)	(0.324)
Population	0.680***	0.641***	0.685***	0.646***
	(0.142)	(0.164)	(0.124)	(0.206)
Area	-0.355*	-0.358***	-0.314*	-0.318
	(0.186)	(0.125)	(0.168)	(0.204)
Observations	85	85	85	85
Adj. R-squared	0.793	0.797	0.799	0.803

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017), Gershman and Rivera (2018).

Table 2.6: Adding controls for fractionalization

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math test score	5.396**	6.042**	5.697***	6.236**	6.254**	6.244**	5.079*
	(2.476)	(2.360)	(1.845)	(2.543)	(2.493)	(2.458)	(2.685)
Tribes	-0.196						-0.269
	(0.277)						(0.277)
Fatalities		0.077***					0.049
		(0.020)					(0.044)
Ethno-linguistic fractionalization			0.215**				0.622**
			(0.088)				(0.300)
Ethno-linguistic polarization				0.108			-0.668*
				(0.133)			(0.391)
Religious fractionalization					0.169*		-1.730
					(0.093)		(2.497)
Religious polarization						0.184	1.850
						(0.137)	(2.518)
Population	0.755***	0.749***	0.753***	0.717***	0.683***	0.688***	0.780***
	(0.180)	(0.151)	(0.113)	(0.098)	(0.100)	(0.134)	(0.186)
Area	-0.408**	-0.272**	-0.305***	-0.327**	-0.378***	-0.374***	-0.418**
	(0.163)	(0.136)	(0.113)	(0.130)	(0.113)	(0.104)	(0.213)
Observations	112	112	90	90	90	90	90
Adj. R-squared	0.787	0.791	0.802	0.793	0.799	0.800	0.815

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017), Nunn & Wantchekon (2011), Clionadh et al. (2010).

Table 2.7: Adding years of schooling and first principal component of all groups of other control variables

	(1)	(2)	(3)	(4)	(5)	(6)
Math test score	4.907**	5.951**	5.976**	5.006**	5.883**	4.676*
	(2.493)	(2.773)	(2.593)	(2.035)	(2.702)	(2.568)
Years of schooling	1.216***					1.127***
	(0.222)					(0.322)
Geography		0.150				0.046
		(0.282)				(0.145)
Nature			0.114			-0.082
			(0.088)			(0.087)
Poor health care				-0.263***		-0.116
				(0.100)		(0.106)
Fractionalization					0.142***	0.029
					(0.055)	(0.077)
Population	0.693***	0.737***	0.657***	0.635***	0.739***	0.570***
	(0.137)	(0.180)	(0.162)	(0.114)	(0.091)	(0.157)
Area	-0.198	-0.291*	-0.388**	-0.333**	-0.343***	-0.297**
	(0.156)	(0.161)	(0.169)	(0.134)	(0.098)	(0.151)
Observations	90	112	80	85	90	80
Adj. R-squared	0.856	0.787	0.802	0.804	0.803	0.867

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. The controls are indices created with principal component analysis from all controls (except for mineral facilities) in the available field using the first component. Component loading is reported in Table A 2.13. All variables in logarithm except for dummies. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

Table 2.8: Alternative measures for economic development

Dependent Variable	Luminosity per capita	GDP per capita
	(1)	(2)
Math test score	5.470**	4.488**
	(2.678)	(2.065)
Area	-0.098	-0.096
	(0.079)	(0.064)
Observations	112	90
Adj. R-squared	0.0308	0.194

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable indicated in column header. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

Appendix

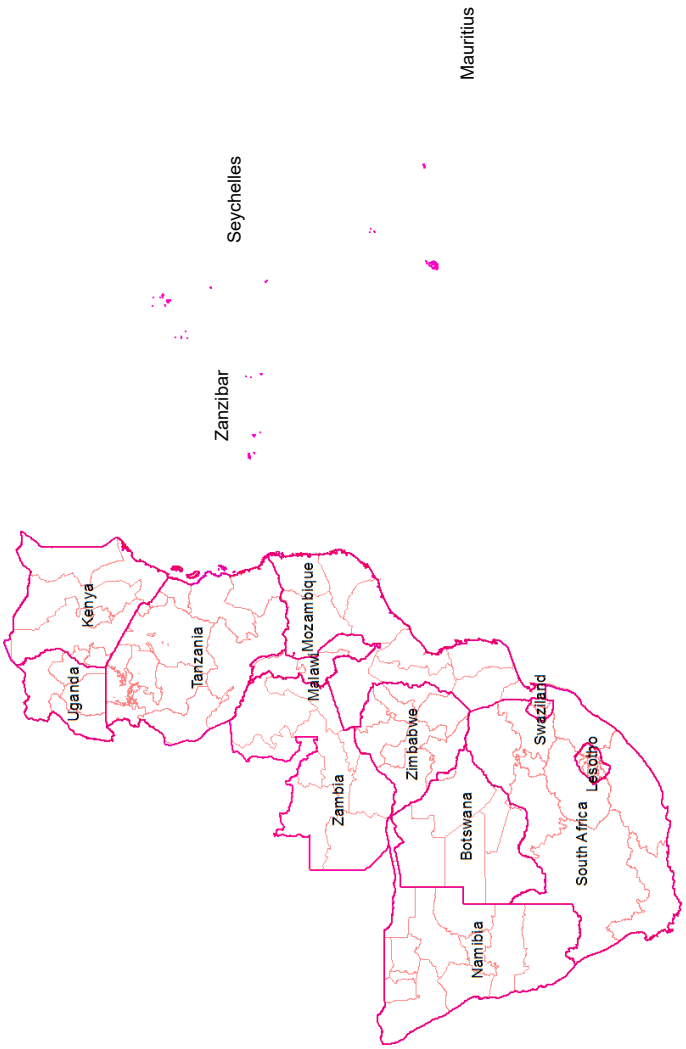
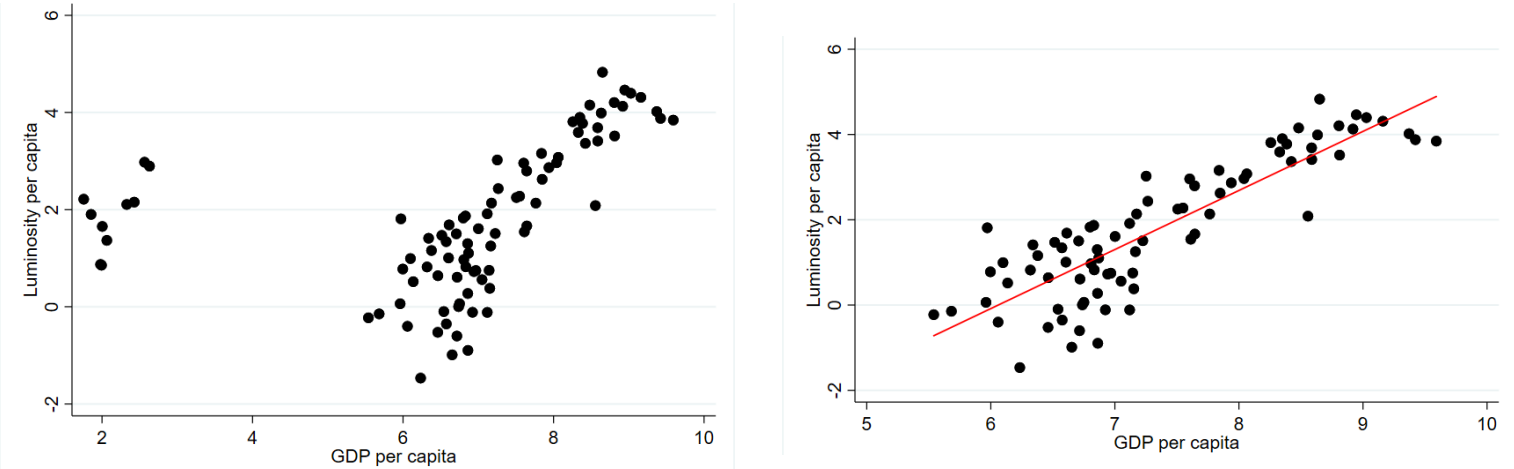


Figure A 2.1: Map of sample countries

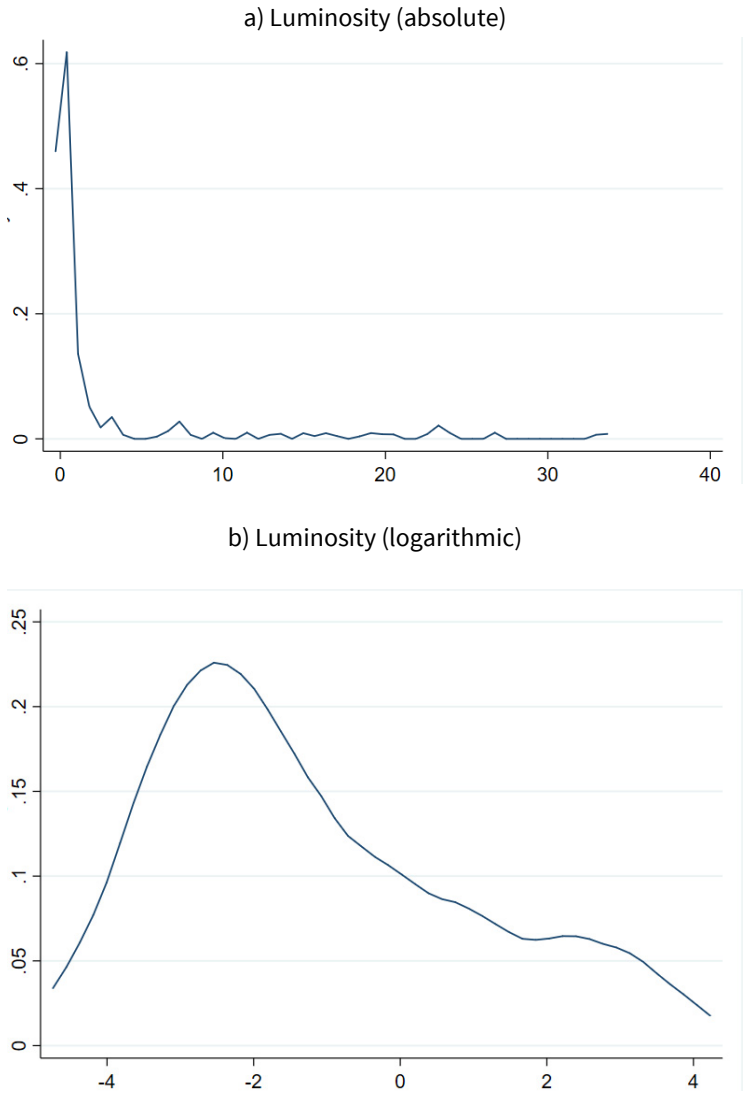
Notes: The map shows the countries included in the sample. Regional borders are indicated by red lines and country borders indicated by thick purple lines. Data Source: StatSilk (2016), Global Administrative Areas (GADM 2016).

Figure A 2.2: Relationship between luminosity per capita and GDP per capita



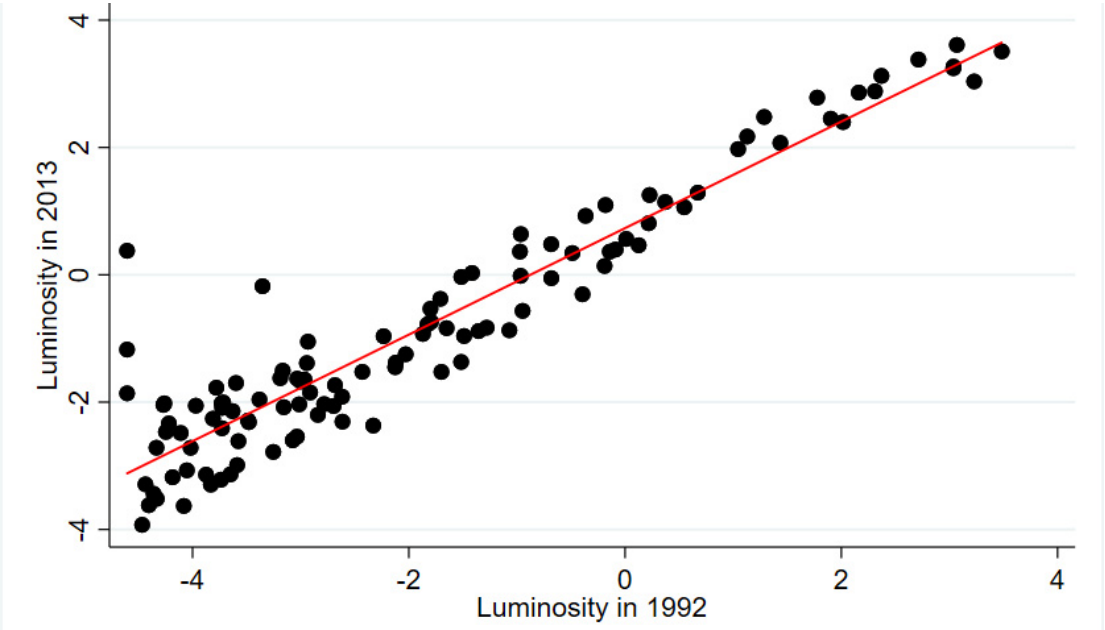
Notes: The graph shows the relationship of luminosity and GDP per capita. The left panel contains all countries in the sample. The right panel excludes the outliers to the left-hand side, i.e., Zimbabwe. All variables are in logarithm. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013).

Figure A 2.3: Distribution of luminosity



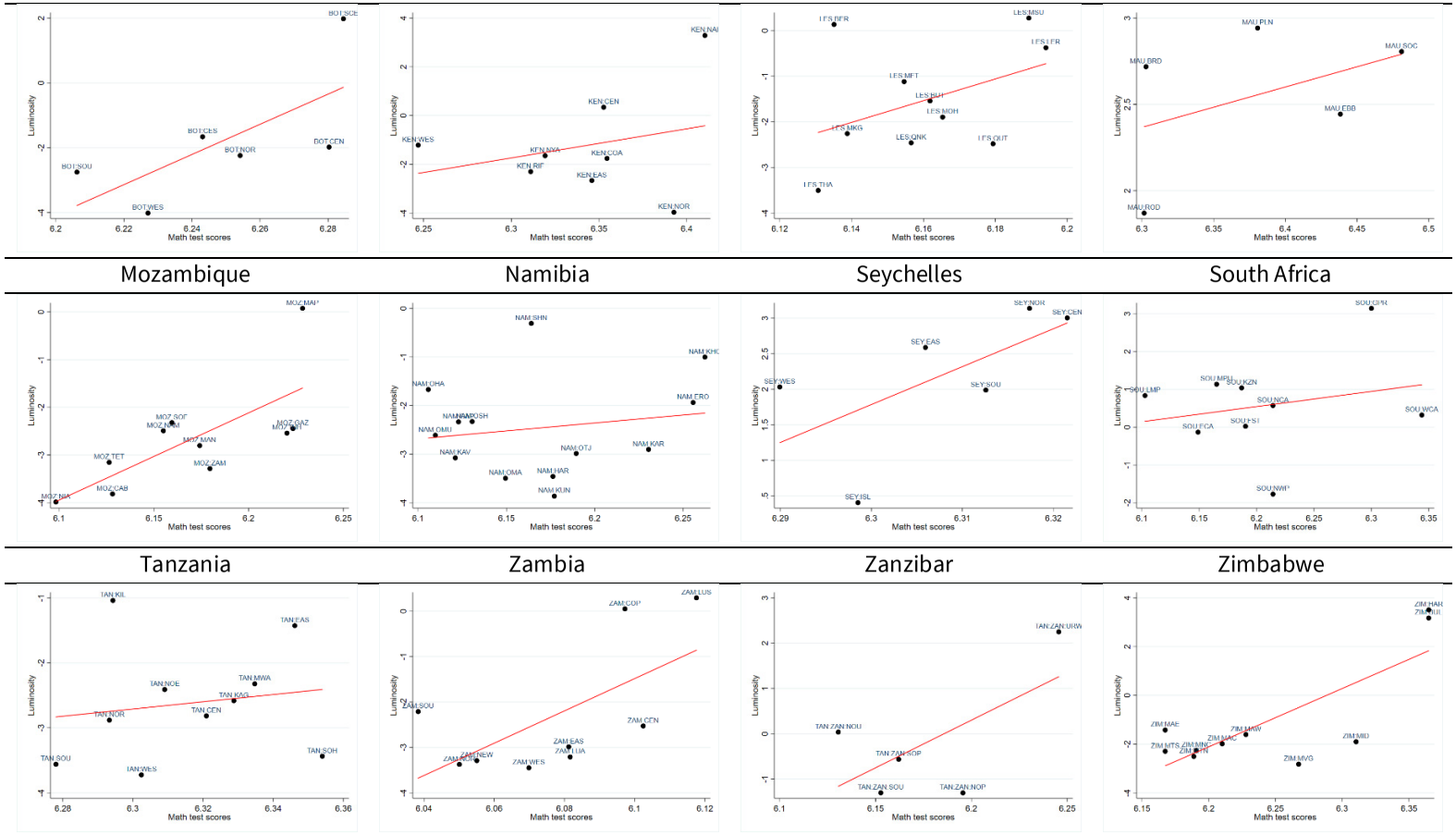
Notes: The graph shows the distribution of a) luminosity and b) log luminosity. *Data sources*: V4 DMSP-OLS Nighttime Lights Time Series.

Figure A 2.4: Persistence of luminosity



Notes: Red line fits a linear relationship between luminosity in 1992 and luminosity in 2013. All variables are in logarithm. Data source: V4 DMSP-OLS (NOAA's National Geophysical Data Center 2016) Nighttime Lights Time Series.

Figure A 2.5: Relationship between luminosity and math test scores in sample countries



Notes: The graphs show the correlation of luminosity with math test scores by sample country. We show only countries with at least five regions (i.e., Malawi, Swaziland, and Uganda are not shown). Variables are in logarithm. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007).

Table A 2.1: Key characteristics of national education systems in sample countries

Country	Children not in school*	Duration of Primary School	Duration of Secondary School	Duration of Compulsory Education	Entrance age to primary education	Entrance age to lower secondary education	Pupil-teacher ratio in primary school	Government expenditure for education as a share of GDP	Share of repeaters in primary education of total enrollment	Share of trained primary teachers
Botswana	15.03	7	5	a	6	13	25.22	8.85	4.89	97.39
Kenya	12.91	6	6	8	6	12	41.95	7.03	5.83	98.70
Lesotho	24.84	7	5	n.a.	6	13	37.20	11.27	20.99	91.03
Malawi	6.83	6	6	8	6	12	78.12	3.54	20.68	88.38
Mauritius	3.50	6	7	11	5	11	21.51	3.18	3.72	100
Mozambique	19.13	7	5	n.a.	6	13	64.80	4.28	5.93	63.19
Namibia	12.02	7	5	7	7	14	29.95	6.04	18.11	94.81
Seychelles	4.49	6	7	10	6	12	12.47	4.77	n.a.	2.43
Swaziland	24.22	7	5	n.a.	6	13	32.44	6.71	18.04	94.04
Tanzania	4.13	7	6	7	7	14	53.09	4.04	4.22	100.00
Uganda	6.44	7	6	n.a.	6	13	49.56	2.39	14.79	93.36
South Africa	11.70	7	5	9	7	14	33.19	4.97	7.96	77.92
Zambia	8.84	7	5	7	7	14	55.35	1.24	6.50	92.66
Zimbabwe	10.89	7	6	7	6	13	35.78	1.97	2.12	88.15

Notes: All statistics refer to 2007 (out-of-school children refer to 2005 in the Seychelles, to 2009 in Swaziland and Uganda, to 2008 in Tanzania, to 2006 in South Africa, and to 2011 Zimbabwe; pupil-teacher ratio refers to 2010 in Zimbabwe; government expenditure for education as a share of GDP refers to 2006 in for Kenya, Lesotho, Mozambique, Namibia, and Swaziland, to 2006 in the Seychelles, and to 2010 in Malawi, Uganda, and Zimbabwe; share of repeaters refers to 2006 in Kenya, to 2008 in Namibia, to 2004 in South Africa, and to 2012 in Zimbabwe; share of trained teachers refers to 2013 in Lesotho, to 2011 in the Seychelles, to 2002 in South Africa, and to 2012 in Zambia and Zimbabwe). In years or percent. Lesotho, Mozambique, Swaziland, and Uganda do not have compulsory schooling. *As percentage of people in primary school age. Tanzania includes Zanzibar. Data Sources: World Development Indicators (World Bank 2018a) (for primary school duration, pupil-teacher ratio, government expenditure, share of repeaters, tertiary

enrollment); UIS (2018)(for out-of-school children, secondary school duration, compulsory school, entrance age, percentage of trained primary teachers).

Table A 2.2: Sectoral composition in sample countries

Sector	Agriculture	Industry	Services
Botswana	1.70	29.20	69.10
Kenya	35.3	17.2	47.9
Lesotho	5.3	34.6	60.1
Malawi	28.1	15.8	56.1
Mauritius	22.50	37.80	39.70
Mozambique	22.3	23	54.7
Namibia	6.6	25.8	67.6
Seychelles	2.50	13.80	83.70
South Africa	2.8	29.7	67.5
Swaziland	6.5	45	48.6
Tanzania	23.4	28.6	47.6
Uganda	25.8	23.2	51
Zambia	5.4	35.6	59
Zimbabwe	12.5	26.9	60.6
Average	14.3	27.6	58.1

Notes: This table shows the share of agriculture, industry, and services, respectively, in the GDP. Tanzania includes Zanzibar. Values are in percent. Data source: CIA World Factbook (2017).

Table A 2.3: Variable definitions

Variable	Description	Dataset	Year	Format	Source
Outcomes					
Luminosity	Average nightlight intensity in a region.	V4 DMSP-OLS	2007	Raster	https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html
Luminosity per capita	Average nightlight intensity in a region per ten-thousand inhabitants.		2007		
GDP per capita	Gross domestic product (per capita) in constant 2005 purchasing power parity (PPP) dollars.	Gennaioli et al. (2013)	we use the year closest to 2007	Excel with geocodes	https://scholar.harvard.edu/shleifer/publications/human-capital-and-regional-development
Educational achievement					
Math test scores	Test scores in SACMEQ III math assessment. Measured on a scale with mean of 500 and standard deviation of 100.	SACMEQ III	2007	Stata	http://www.sacmeq.org/
Reading test scores	Test scores in SACMEQ III reading assessment. Measured on a scale with mean of 500 and standard deviation of 100.				

(Continued on next page.)

Variable	Description	Dataset	Year	Format	Source
Educational attainment					
Years of schooling	The average years of schooling from primary school onward for the population aged 15 years or older.	Gennaioli et al. (2013)	we use the year closest to 2007	Excel with geocodes	https://scholar.harvard.edu/shleifer/publications/human-capital-and-regional-development
Primary gross enrollment	Ratio (in %) of the total number of children attending primary school relative to the number of children within the official age range for primary school. Due to children attending primary school outside of the official age range (due to grade repetition), this ratio may exceed 100.	Sub-national African Education and Infrastructure Access Data	we use the year closest to 2007	Stata	https://www.strausscenter.org/ccaps-content/climate-vulnerability-model.html
Primary net enrollment	Ratio (in %) of the number of children of official primary school age attending primary school relative to the total of number of primary school age children.				
Secondary gross enrollment	Ratio (in %) of the total number of children attending secondary school relative to the number of children within the official age range for secondary school. Due to children attending secondary school outside of the official age range, this ratio may be above 100.				
Secondary net enrollment	Ratio (in %) of the number of children of official secondary school age attending secondary school relative to the total of number of secondary school age children.				
Population share with college degree	Share of the population aged 15 years or older whose highest educational level is ISCED 5 or 6. Asked from the household head.	Gennaioli et al. (2013)	we use the year closest to 2007	Excel with geocodes	https://scholar.harvard.edu/shleifer/publications/human-capital-and-regional-development
Literacy rate	Share of region's adult population (aged 15–49 years) that is able to read (we use the year closest to 2007).	Gershman and Rivera (2018)	2007-2013	Stata	http://www.borisgershman.com/research.html

(Continued on next page.)

Variable	Description	Dataset	Year	Format	Source
Population					
Population density	Inhabitants per arc degrees.	Gridded Population density of the World (GPW), v4	2005	Raster	http://sedac.ciesin.columbia.edu/data/set/gpw-v4-Population density-count-rev10
Population	Number of inhabitants in ten-thousands.				
Area	Regions mean surface area in arc degrees.	From ArcGIS			
Geography					
Longitude	Region's mean distance from the Prime Meridian.	From ArcGIS			
Latitude	Region's mean absolute distance to the equator.	From ArcGIS			
Landlocked	Dummy variable equal to one if the region is landlocked (zero otherwise).	Gershman and Rivera (2018)	2007-2013	Stata	http://www.borisgershman.com/research.html
Capital	Region which contains the capital city.	Self-generated			
Nature					
Ruggedness	Terrain Ruggedness Index, in millimeters.	GMTED by Nunn & Puga (2012)	2010	Ascii grid	https://diegopuga.org/data/rugged/
Mineral facilities	Number of mineral facilities per region.	Eros & Candelario-Quintana (2006)	2006	Shapefile and Excel with geocodes	https://pubs.usgs.gov/of/2006/1135/

(Continued on next page.)

Variable	Description	Dataset	Year	Format	Source
Protection against climatic hazards	Index of protection against floods, rainfall anomalies, chronic water scarcity, coastal elevation, tropical cyclones, and wildfires (anomalies between 1970 and 2009).	CCAPS Climate Security Vulnerability Model 3.0	2014	Tiff	https://www.strausscenter.org/ccaps-content/climate-vulnerability-model.html
Land suitability for agriculture	Index of land suitability for rain-fed agriculture (base data from 2000).	Gershman and Rivera (2018)	2007-2013	Stata	http://www.borisgershman.com/research.html
Temperature	Average temperature in a region during the period 1950 to 2000 in degrees Celsius.	Gennaioli et al. (2013)	average during the period 1950-2000	Excel with geocodes	https://scholar.harvard.edu/shleifer/publications/human-capital-and-regional-development
Health					
Infant underweight	Weight-for-age z-scores which are more than two standard deviations below the median of the international reference population (in 2005).	Gennaioli et al. (2013)	we use the year closest to 2007	Excel with geocodes	https://scholar.harvard.edu/shleifer/publications/human-capital-and-regional-development
Share of homebirths	The share of births delivered at home (data available closest to 2007).	Gershman and Rivera (2018)	2007-2013	Stata	http://www.borisgershman.com/research.html

(Continued on next page.)

Variable	Description	Dataset	Year	Format	Source
Fractionalization					
Tribes	Variety of tribes per area (Murdoch's atlas in 1959).	Murdock's ethnic atlas by Nunn & Wantchekon (2011)	1959	Shapefile	http://scholar.harvard.edu/files/nunn/files/murdock_shapefile.zip
Fatalities	Number of reported deaths from a conflict event.	Armed Conflict Location & Event Data Project (ACLED) by Clionadh et al. (2010)	2007	Excel with geocodes	https://www.acleddata.com/download/2909/
Ethno-linguistic fractionalization	Index of ethnolinguistic fractionalization (ELF) adjusted for group similarities.	Gershman and Rivera (2018)	2007-2013	Stata	http://www.borisgershman.com/research.html
Ethno-linguistic polarization	Index of ethnolinguistic polarization (ELP) capturing antagonism between groups by the distance of a society from a perfectly polarized population (assuming that the existence of a sizable ethnic minority alongside the dominant group substantially increases the likelihood of ethnic conflict).				
Religious fractionalization	ELF based on a four-way classification for religion into Christianity, Islam, "traditional" religion, and none.				
Religious polarization	ELP based on a four-way classification for religion into Christianity, Islam, "traditional" religion, and none.				

(Continued on next page.)

Variable	Description	Dataset	Year	Format	Source
Preferences					
Trust	Self-assessment of people only having the best intentions.	Falk et al. (2018) within the Gallup World Poll	2012	Stata	https://www.briq-institute.org/global-preferences/downloads
Patience	Intertemporal choice sequence using staircase method and self-assessed willingness to wait.				
Risk	Lottery choice sequence using staircase method and self-assessed willingness to take risks in general.				
Altruism	Donation decision and self-assessed willingness to give to good causes.				

Table A 2.4: Summary statistics

	N	Mean	SD	Min	Max
Outcomes					
Luminosity	112	2.7	6.4	0.0	33.4
Luminosity per capita	112	19.9	24.7	0.2	125.1
GDP per capita	90	2,543	2,998	5.8	14,634
Educational achievement					
Math test score	112	505.3	48.3	419.2	652.7
Reading test score	112	509.3	53.6	414.0	622.7
Educational attainment					
Years of schooling	90	4.0	2.1	0.5	9.4
Primary gross enrollment	90	107.7	14.9	59.3	145.6
Primary net enrollment	90	80.4	11.0	42.3	92.8
Secondary gross enrollment	90	50.8	25.7	10.1	114.1
Secondary net enrollment	90	34.2	20.7	1.8	78.8
Population share with college degree	90	0.0	0.0	0.0	0.2
Literacy rate	90	0.8	0.2	0.2	1.0
Population					
Population density	112	10,613	44,545	0.0	316,771
Population (in ten-thousands)	112	158.8	411.9	0.3	3,327
Geography					
Area	112	5.0	5.4	0.0	34.7
Latitude	112	-16.1	9.6	-32.7	2.8
Longitude	112	32.1	9.8	15.2	57.9
Landlocked	112	0.7	0.5	0.0	1.0
Nature					
Ruggedness	112	0.2	0.2	0.0	1.0
Mineral facilities	112	3.9	17.8	0.0	167.5
Protection against climatic hazards	112	223.8	26.1	151.6	255.0
Land suitability for agriculture	90	4.9	1.6	2.0	8.0
Temperature	90	20.0	3.8	8.3	27.3
Health					
Infant underweight	90	0.1	0.1	0.1	0.3
Share of home births	85	0.3	0.2	0.0	0.8
Fractionalization					
Tribes	112	30.2	117.6	0.2	833.3
Ethno-linguistic fractionalization	90	0.5	0.3	0.0	0.9
Ethno-linguistic polarization	90	0.6	0.2	0.0	0.9
Religious fractionalization	90	0.2	0.2	0.0	0.6
Religious polarization	90	0.4	0.3	0.0	1.0
Fatalities	112	0.4	1.0	0.0	4.7

(Continued on next page.)

	N	Mean	Std. Dev.	Min	Max
Preferences					
Trust	51	-0.3	0.4	-1.2	0.7
Patience	51	-0.1	0.2	-0.6	0.7
Risk	51	0.6	0.4	-0.4	1.3
Altruism	51	-0.2	0.3	-1.1	0.7

Notes: Non-logarithmized values are reported.

Table A 2.5: Region observations by variable

Variable/Country	BOT	KEN	LES	MAL	MAU	MOZ	NAM	SEY	SOU	SWA	TAN	UGA	ZAM	ZAN	ZIM	Total
Climate hazards	6	8	10	3	5	10	13	6	9	4	10	4	9	5	10	112
Fatalities	6	8	10	3	5	10	13	6	9	4	10	4	9	5	10	112
Gennaioli et al. (2013) (GDP, years of schooling, temperature)	-	8	10	3	-	10	13	-	9	4	10	4	9	-	10	90
Gershman & Rivera (2018) (capital, landlocked, literacy rate, home births, underweight, urbanization, land suitability, ethno-linguistic and religious polarization and fractionalization)	5	8	-	3	-	10	13	-	9	4	10	4	9	5	10	90
Mineral facilities	6	8	10	3	5	10	13	6	9	4	10	4	9	5	10	112
Preferences	5	8	-	3	-	-	-	-	9	-	9	4	-	3	10	51
Primary and secondary enrollment (net & gross)	-	8	10	3	5	10	13	6	9	4	10	4	9	-	10	101
Ruggedness	6	8	10	3	5	10	13	6	9	4	10	4	9	5	10	112
Tribes	6	8	10	3	5	10	13	6	9	4	10	4	9	5	10	112
Number of regions in country	*6*	*8*	*10*	*3*	*5*	*10*	*13*	*6*	*9*	*4*	*10*	*4*	*9*	*5*	*10*	*112*

Notes: Table shows the number of region-level observations by country and variable.

Table A 2.6: Student observations

	Number of students
Per region	1294.629 on average 669.79 SD [354 to 2,972]
Per school	48.1 on average 23.39 SD [4 to 310]
Per class	28.8 on average 14.75 SD [2 to 72]
Total	121,370 students in 112 regions

Notes: Table shows the number of students per region, school, and class, respectively. Reported are average, standard deviation (SD), and range between minimum and maximum (in square brackets).

Table A 2.7: Correlations between human capital variables

	Math test score	Years of schooling	Primary net enrollment	Secondary net enrollment	Population share with college degree	Literacy rate
Math test score	1					
Years of schooling	0.283	1				
	(0.007)					
Primary net enrollment	0.092	0.338	1			
	(0.390)	*(0.001)*				
Secondary net enrollment	0.068	0.845	0.444	1		
	(0.523)	*(0.000)*	*(0.000)*			
Population share with college degree	0.350	0.660	0.187	0.347	1	
	(0.001)	*(0.000)*	*0.077*	*(0.001)*		
Literacy rate	0.276	0.766	0.743	0.750	0.491	1
	(0.009)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	

Notes: p-values in parentheses. The correlation between math test scores and reading test scores (at regional level) is 0.91.

Table A 2.8: Using luminosity in 2013 as outcome

	(1)	(2)	(3)	(4)	(5)
Math test score	11.994***	4.264**	6.841***	4.436*	5.099**
	(3.662)	(1.949)	(2.097)	(2.362)	(2.037)
Population		1.015***		0.879***	
		(0.069)		(0.157)	
Area			-1.044***	-0.174	
			(0.068)	(0.138)	
Population density					0.549***
					(0.026)
Observations	112	112	112	112	112
Adj. R-squared	0.0193	0.778	0.649	0.779	0.762

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity (measured in 2013). All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

Table A 2.9: Clustering standard errors at country level

	(1)	(2)	(3)	(4)	(5)
Math test score	12.948***	5.539**	7.875***	5.822***	6.259***
	(3.775)	(2.333)	(2.058)	(2.260)	(2.128)
Population		0.972***		0.750***	
		(0.072)		(0.145)	
Area			-1.028***	-0.286**	
			(0.054)	(0.129)	
Population density					0.533***
					(0.029)
Observations	112	112	112	112	112
Adj. R-squared	0.058	0.778	0.689	0.787	0.780

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Standard errors are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

Table A 2.10: Alternative functional forms

Functional form:	Log (luminosity) - lin	Log (luminosity pc) - lin	Log (GDP pc) - lin	Log (luminosity) - lin (squared)	Lin (luminosity) - lin
	(1)	(2)	(3)	(4)	(5)
Math test score	0.616***	0.526***	0.465***	0.739***	1.890**
	(0.224)	(0.199)	(0.154)	(0.220)	(0.916)
Math test score^2				-0.191	
				(0.157)	
Population	0.666***			0.715**	4.238**
	(0.178)			(0.303)	(1.832)
Area	-0.818***	-0.035	-0.075	-0.805***	-0.787*
	(0.237)	(0.089)	(0.083)	(0.209)	(0.459)
Observations	112	112	90	112	112
Adj. R-squared	0.588	-0.013	0.157	0.600	0.699

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity in Columns 1 and 4, log luminosity per capita in Column 2, log GDP per capita in Column 3, and luminosity in Column 5. Test score, population, and area are standardized to mean 0 and standard deviation 1 across countries. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

Table A 2.11: Sample composition

Panel A: Excluding regions

Sample excluding	Capital regions	Two most densely populated regions in each country	Least densely populated region in each country	Countries where SACMEQ regions do not match administrative regions
	(1)	(2)	(3)	(4)
Math test score	5.354**	5.655**	5.743**	6.072**
	(2.479)	(2.361)	(2.727)	(2.811)
Population	0.635***	0.678***	-0.291	0.669***
	(0.167)	(0.150)	(0.239)	(0.246)
Area	-0.363**	-0.287**	0.756***	-0.380
	(0.152)	(0.130)	(0.274)	(0.291)
Observations	96	82	98	78
Adj. R-squared	0.680	0.733	0.759	0.789

*** p<0.01, ** p<0.05, * p<0.10

Notes: In Column 1, region which contains a country's capital city are excluded. In Column 2 (Column 3), the two most densely populated regions (the least densely populated region) in each country are excluded. In Column 4, sample consists of countries where education regions in SACMEQ matched administrative regions (i.e., Kenya, Lesotho, Mauritius, Mozambique, Namibia, Swaziland, South Africa, Zambia, and Zimbabwe). Dependent variable: log luminosity. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

(Continued on next page.)

Panel B: Excluding one country at a time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Math test score	5.608**	5.169**	5.903***	5.671**	5.941**	4.998**	3.518**	5.632***	6.902***	5.613**	5.932**	5.733**	5.557***	5.907**	6.099*
	(2.463)	(2.612)	(2.174)	(2.504)	(2.638)	(2.330)	(1.755)	(2.156)	(2.599)	(2.295)	(2.346)	(2.537)	(2.059)	(2.472)	(3.270)
Population	0.721***	0.732***	0.644***	0.712***	0.739***	0.750***	0.875***	0.731***	0.696***	0.718***	0.715***	0.707***	0.699***	0.737***	0.723***
	(0.153)	(0.189)	(0.123)	(0.145)	(0.165)	(0.137)	(0.160)	(0.165)	(0.182)	(0.155)	(0.178)	(0.175)	(0.145)	(0.136)	(0.196)
Area	-0.333**	-0.320**	-0.392***	-0.340**	-0.319**	-0.299**	-0.233*	-0.334***	-0.365**	-0.335***	-0.318*	-0.345**	-0.340**	-0.293**	-0.336*
	(0.139)	(0.156)	(0.110)	(0.141)	(0.135)	(0.120)	(0.136)	(0.120)	(0.155)	(0.119)	(0.171)	(0.140)	(0.140)	(0.146)	(0.200)
Excluding	BOT	KEN	LES	MAL	MAU	MOZ	NAM	SEY	SOU	SWA	TAN	UGA	ZAM	ZAN	ZIM
Observations	106	104	102	109	107	102	99	106	103	108	97	108	103	107	102
Adj. R-squared	0.765	0.772	0.807	0.789	0.798	0.793	0.799	0.785	0.787	0.790	0.795	0.798	0.795	0.791	0.740

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Country abbreviations: KEN=Kenya, LES = Lesotho, MAL= Malawi, MOZ =Mozambique, NAM= Namibia, SOU= South Africa, SWA=Swaziland, TAN= Tanzania, UGA=Uganda, ZAM= Zambia, ZAN= Zanzibar, ZIM= Zimbabwe. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

Table A 2.12: Preferences

	(1)	(2)	(3)	(4)	(5)	(6)
Math test score	3.106*	3.597*	3.196	3.107	3.082	3.848
	(1.793)	(2.118)	(2.447)	(2.118)	(1.897)	(2.958)
Trust		0.721				0.976*
		(0.464)				(0.543)
Patience			0.731			0.893
			(1.229)			(1.581)
Risk aversion				-0.126		-0.348
				(0.688)		(1.323)
Altruism					-0.703	-1.048*
					(0.439)	(0.605)
Population	0.792***	0.823***	0.792***	0.790***	0.760***	0.780***
	(0.115)	(0.110)	(0.128)	(0.116)	(0.152)	(0.158)
Area	-0.300**	-0.282**	-0.296**	-0.300*	-0.343**	-0.334**
	(0.122)	(0.113)	(0.136)	(0.162)	(0.146)	(0.143)
Observations	51	51	51	51	51	51
Adj. R-squared	0.851	0.855	0.850	0.848	0.853	0.857

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Center for International Earth Science Information Network (CIESIN) (2017), Falk et al. (2018).

Table A 2.13: Results from principal component analysis

	Share of variation explained by first principal component	Loadings on first principal component
Geography	0.376	
Longitude		-0.716
Latitude		0.4700
Landlock		0.5125
Capital		0.0547
Nature	0.476	
Ruggedness		0.6002
Climatic hazards		-0.2314
Land suitability		0.496
Temperature		-0.5833
Poor health care	0.636	
Underweight children		0.7071
Home births		0.7071
Fractionalization	0.379	
Tribes		-0.1296
Fatalities		0.1849
Ethno-linguistic fractionalization		0.4517
Ethno-linguistic polarization		0.4542
Religious fractionalization		0.5182
Religious polarization		0.5197

Notes: This table refers to Table 2.7.

Table A 2.14: Reading Skills

	(1)	(2)	(3)	(4)	(5)
Reading test score	11.621***	5.087**	6.711***	5.191**	5.480***
	(3.294)	(1.990)	(2.065)	(2.103)	(1.619)
Population		0.949***		0.746***	
		(0.072)		(0.144)	
Area			-0.997***	-0.263**	
			(0.083)	(0.131)	
Population density					0.520***
					(0.026)
Observations	112	112	112	112	112
Adj. R-squared	0.127	0.790	0.701	0.798	0.790

*** p<0.01, ** p<0.05, * p<0.10

Notes: Dependent variable: log luminosity. All variables are in logarithm. See Table A 2.3 for definition of variables. All regressions include country fixed effects. Adj. R-squared refers to within-country R-squared (i.e., country fixed effects are partialed out). Bootstrapped standard errors in parentheses are clustered at the country level. Data sources: V4 DMSP-OLS Nighttime Lights Time Series, SACMEQ (2007), Gennaioli et al. (2013), Center for International Earth Science Information Network (CIESIN) (2017).

3 Testing [66]

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

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

While there have been previous evaluations of the impact of accountability systems, largely within the United States (Figlio & Loeb 2011), it is unclear how to generalize from

[66] This chapter is joint work with Eric A. Hanushek of Hoover Institution, Stanford University, CESifo, IZA, and NBER, and Ludger Woessmann of University of Munich, ifo Institute, CESifo, and IZA.

these.[67] These policies operate within a specific institutional environment of national school systems; as such, the evaluations necessarily neglect overall features that are common across a nation. Moreover, testing policies are often set at the national level, making it difficult to construct an adequate comparison group for evaluation of policy outcomes. By moving to international comparisons, it is possible to consider how overall institutional structures interact with the specifics of student assessments and school accountability systems. This cross-country approach allows us to investigate which aspects of student assessment systems generalize to larger settings and which do not. Of course, this advantage comes at a cost, because identifying the impact of various schooling policies across nations offers its own challenges.

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

Because this is a period of rapid change in student assessment policies across countries, we can link policies to outcomes in fixed-effects panel models. Our identification relies on changes in student assessment regimes within countries over time. While using the individual student data for estimation at the micro level, we measure our treatment variables as country aggregates at each point in time to avoid bias from within-country selection of students into schools. Conditioning on country and year fixed effects allows us to account for unobserved time-invariant country characteristics as well as common time-specific shocks.[68]

[67] There is broader geographic representation of studies of student exit exams, a type of assessment that we analyze here (see Woessmann 2018 for a review). More detailed references to the existing literature on different forms of testing will follow in Section 3.1.1 below, where they can be discussed within the setting of our conceptual framework.

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

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

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

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

The next section develops a conceptual framework that highlights the achievement effects of different dimensions of student assessments. Section 3.2 introduces the data and Section 3.3 the empirical model. Section 3.4 presents our results including analyses of heterogeneous effects. Section 3.5 reports a placebo test and other specification tests, and Section 3.6 shows a series of robustness analyses. Section 3.7 concludes.

3.1 An incentive framework of different dimensions of assessments

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

3.1.1 Conceptual framework: Principal-agent relationships

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

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

$$\text{Parents: } \max \mathrm{V_p} = \mathrm{f_p}[\mathrm{A_s}, \mathrm{R_s}, \mathrm{E_s}] \tag{3.1}$$

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

Parents, however, cannot directly choose the elements of this value function but must work indirectly to achieve their ends. In particular, they may offer short-term rewards for

learning *R* to their child and try as best as possible to observe and control child effort *E*. Similarly, achievement *A* is only partially controlled by parents but as a general rule relies heavily upon purchasing the services of schools. This is natural because of economies of scale in producing knowledge, of the limited ability of parents to provide the full array of school services, and of the benefits of specialization.

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

$$A_s = A_s(I, E_t, E_s) \tag{3.2}$$

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

As effort levels of teachers and children cannot be perfectly observed or controlled by parents, this setup gives rise to a tree of standard principal-agent relationships (Laffont & Martimort ca. 2003).[69] In particular, parents act as principals that contract the teaching of their children to schools and teachers as agents. In the process of classroom instruction, teachers also act as principals themselves who cannot fully observe the learning effort of their students as agents. Teaching in the classroom and studying at your desk may be viewed as classical examples of asymmetric information where the respective principal cannot fully monitor the behavior of the respective agent. Parents, teachers, and students each have specific objective functions that combine with the asymmetric information of the actors. Therefore, one cannot simply assume that the actions of children and teachers will lead to the optimal result for parents.

Let us assume that teachers maximize the following value function:

$$\text{Teachers: } \max V_t = f_t\left[A_s\left(I, \underbrace{E_t}_{(+)}, E_s\right), R_t, \underbrace{E_t}_{(-)}\right] \tag{3.3a}$$

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

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

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

$$\text{Students: } \max V_s = f_s\left[A_s\left(I, E_t, \underbrace{E_s}_{(+)}\right), R_s, \underbrace{E_s}_{(-)}\right] \tag{3.4a}$$

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

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

3.1.2 The technology of student assessment

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

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

$$O_s = \tau(A_s) \tag{3.5}$$

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

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

In a general way, we can think of providing rewards *R* to both teachers and students based on the outcome levels *O* observed by the student assessments:[70]

$$\text{Teachers: } \max V_t = f_t\left[A_s\left(I, \underbrace{E_t}_{(+)}, E_s\right), R_t\left(\underbrace{O_s}_{(+)}\right), \underbrace{E_t}_{(-)}\right] \tag{3.3b}$$

$$\text{Students: } \max V_s = f_s\left[A_s\left(I, E_t, \underbrace{E_s}_{(+)}\right), R_s\left(\underbrace{O_s}_{(+)}\right), \underbrace{E_s}_{(-)}\right] \tag{3.4b}$$

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

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

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

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

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

The first issue is a classical identification problem. We want to know when we can infer effort levels of teachers and students from information on outcomes. If student efforts were constant over time, we could directly relate changes in achievement in a given classroom to the teacher and from that infer teacher effort levels. Alternatively, if we thought teacher effort was constant, we could attribute different performance of students to their own effort. The first is roughly the idea behind value-added modelling (Koedel, Mihaly & Rockoff 2015; Chetty, Friedman & Rockoff 2017). The second is closer to providing consequential exit exams for student achievement (Bishop 1997). Of course, in neither case is it realistic to assume constant effort by the other actor, but the policy choices implicitly assume that one form of effort is much more important than the other. These issues will be discussed more completely in Section 3.5.2 below.

The second issue recognizes the fact that no assessment technology τ today provides complete measurement of the relevant achievement for long-run well-being. Prior discussions of accountability systems have considered various dimensions of this problem (Figlio & Loeb 2011). Perhaps the best-known conceptual discussion is the classic Holmstrom & Milgrom (1991) paper that considers how imperfect measurement of outcomes distorts incentives (see also Dixit 2002). In particular, if there are multiple objectives and only a subset is measured, effort could be distorted to the observed outcomes to the detriment of unobserved outcomes. But there is also more general discussion of such topics as teaching to the test (Koretz 2017),[71] gaming of tests (e.g.,

[71] There are two aspects of teaching to the test. On the one hand, teaching may unduly focus on the form and character of the test itself, which is not in the interest of parents. Creative and flexible designs of tests are required to prevent such activity. On the other hand, if the tests accurately sample from the domains of achievement that parents desire, focusing teaching towards the contents of the test is in fact part of the mechanism of aligning teaching with the parental value function.

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

3.1.3 Assessment dimension 1: Different strengths of incentives

Testing is a ubiquitous component of schooling, but not all tests have the same use or impact in helping to solve the underlying principal-agent problems. By far the most common type of testing is teacher-developed tests that are used both to guide instruction and to provide feedback to students and parents. While the generated student-specific information may be valued by parents, the key feature of teacher-developed tests is that it is generally difficult if not impossible to compare results across teachers. Therefore, while these tests may be useful in providing incentives to students and related information to parents (O_s enters positively in R_s in equation (3.4b)), they do not solve the principal-agent problem between parents and teachers (O_s effectively does not enter R_t in equation (3.3b)). One would not expect the results of these tests to affect teacher effort levels. There is a blurry line between teacher-developed tests and periodic content testing that generally goes under the heading of formative assessments which may also be provided by external producers. In both cases, the information provided by the tests is just used internally by the teacher without parents being able to compare outcomes externally.

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

Of course, the strength of any incentives relating to these various tests will depend upon how they enter into rewards for teachers and students in equations (3.3b) and (3.4b). On the one hand, results of student assessments may just provide information to some or all

actors in the system.[72] On the other hand, performance on any test may also be linked directly to consequences – rewards and punishments to students (including retention and promotion) and teachers.[73] As a general principle, we would naturally expect attaching consequences to results to produce stronger incentives and larger behavioral changes.

3.1.4 Assessment dimension 2: Different addressees of incentives

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

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

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 have strong incentives for student effort (equation (3.4b)), but limited impact on

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

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

teacher effort (equation (3.3b)).[74] On the other hand, testing that is directly linked to consequences for schools such as the NCLB legislation in the US may have limited relevance for students and their efforts.[75] Similarly, school inspectorates and inspections of teacher lessons may be more relevant for school and teacher effort than for student effort. However, even in these cases, strategic complementarity or substitutability of the actors might produce some ambiguity in responses.[76]

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

3.1.5 Assessment dimension 3: Dependence on school environments

The prior conceptual discussion is framed in terms of a series of individual two-way interactions. Understanding the implications of various testing schemes and their usage necessarily involves looking at performance across schools and, in our case, across countries. When we think in these larger terms, it is difficult to believe that behavior is uniform across systems even when confronted with the same incentive structure.[77]

[74] By affecting chances to enter specific institutions and fields of higher education as well as the hiring decisions of potential employers, central exit exams usually have real consequences for students; see Bishop 1997, Wossmann 2003, Jürges, Schneider & Büchel 2005, Woessmann et al. 2009, Luedemann 2011, Schwerdt & Woessmann 2017, and Woessmann 2018 for further analysis of the effects of central exit exams.

[75] For analyses of the effects of NCLB and predecessor reforms, see Hanushek & Raymond 2005, Jacob 2005, Neal & Schanzenbach 2010, Rockoff & Turner 2010, Dee & Jacob 2011, Rouse et al. 2013, Reback, Rockoff & Schwartz 2014 and Deming et al. 2016; see Figlio & Loeb 2011 for a survey.

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

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

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

3.2 International panel data

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

3.2.1 Six waves of PISA student achievement tests

In 2000, the Organisation for Economic Co-operation and Development (OECD) conducted the first wave of the international achievement test called Programme for International Student Assessment (PISA). Since then, PISA has tested the math, science, and reading achievement of representative samples of 15-year-old students in all OECD countries and in an increasing number of non-OECD countries on a three-year cycle (OECD (2016)).[78] 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.[79] PISA does not follow individual

[78] 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.

[79] 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.

students over time. But the repeated testing of representative samples of students creates a panel structure of countries observed every three years.

In our analyses, we consider all countries that have participated in at least three of the six PISA waves between 2000 and 2015.[80] This yields a sample of 59 countries observed in 303 country-by-wave observations. We perform our analysis at the individual student level, encompassing a total sample of 2,187,415 students in reading and slightly less in math and science. The sample, listed in Table 3.1, includes 35 OECD and 24 non-OECD countries that encompass a wide range of levels of economic development and student achievement.

PISA uses a broad set of tasks of varying difficulty to create a comprehensive indicator of the continuum of students' competencies in each of the three subjects. Overall testing lasts for up 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.[81] Until 2012, PISA employed paper and pencil tests. In 2015, the testing mode was changed to computer-based testing, a topic we will come back to in our robustness analysis below.

Figure 3.1 depicts the evolution of math achievement of each country over the 15-year period. While average achievement across all countries was quite stable between 2000 and 2015, achievement has moved significantly up in some countries and significantly down in others. In 14 countries, achievement improved by at least 20 percent of a standard deviation compared to their initial achievement (in decreasing order, Peru, Qatar, Brazil, Luxembourg, Chile, Portugal, Israel, Poland, Italy, Mexico, Indonesia, Colombia, Latvia, and Germany). On the other hand, achievement decreased by at least 20 percent of a standard deviation in eleven countries (United States, Korea, Slovak

[80] 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.

[81] 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.

Republic, Japan, France, Netherlands, Finland, Iceland, United Kingdom, Australia, and New Zealand).

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

Our vector of control variables allows us to condition on a rich set of observed characteristics of students, schools, and countries. The student-level controls include student gender, age, first- and second-generation immigration status, language spoken at home, parental education (measured in six categories), parental occupation (four categories), and books at home (four categories). The school-level controls include school size (number of students), community location (five categories), share of fully certified teachers, principals' assessments of the extent to which learning in their school is hindered by teacher absenteeism (four categories), shortage of math teachers, private operation, and share of government funding. At the country level, we include GDP per capita and, considering the results in Hanushek, Link & Woessmann (2013), the share of schools with academic-content autonomy and its interaction with initial GDP per capita. To avoid sample selection bias from non-response in the survey data, we impute missing values in the student and school background variables by using the respective country-by-wave mean.[82] To ensure that imputed data are not driving our results, all our regressions include a set of dummy variables – one for each variable with missing data – that are set to one for imputed values and zero otherwise.

[82] The share of missing values is generally very low for the covariates, see Appendix Table A 3.1.

3.2.2 Categories of assessment usage

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

Here we summarize the categories of testing that we construct, while the details of questions and sources can be found in the Data Appendix. From a combination of the surveys for principals that accompany the PISA assessments, of the regular publications and data collection of other parts of the OECD, and from data compiled under the auspices of the European Commission, we have 13 separate indicators of the use and purpose of testing, each measured at the country-by-wave level.[83] We combine these into four separate categories that represent quite different aspects of testing in the schools. They differ by the degree of standardization of the assessment data and the specific actors – administrators, teachers, and students – most affected.

Standardized external comparisons. The first category draws on four separate data sources that identify standardized assessments constructed outside of schools and used explicitly to allow comparisons of student outcomes across schools and students. This category includes the proportion of schools where (according to the principals of schools participating in PISA) performance of 15-year-olds is regularly compared through external examinations to students across the district or the nation (which we term "school-based external comparisons"). It also includes indicators of whether central

[83] Appendix Table Table A 3.2 provides an overview of the different underlying assessment indicators. Appendix Table Table A 3.3 indicates the number of country observations by wave for each indicator.

examinations affect student placement at lower secondary level (two sources) and whether central exit exams determine student outcomes at the end of secondary school (which, together, we term "student-based external comparisons").[84] This overall category of exams has strong incentives through the rewards to students but also affects rewards to administrators and teachers by making external information available to parents and policy makers. While not fully explicit from the surveys, the items in this category are roughly ones where consequential outcomes are related to student scores, making for stronger incentives.[85]

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

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

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

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

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

uncertain but generally small impacts on teacher rewards because of the lack of comparability across settings.

Aggregation of separate indicators. We combine the original 13 separate indicators of assessment practices into four main categories as the simple average of the observed indicators in each category.[86] 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 the 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 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.[87]

Some descriptive statistics. Table 3.2 provides descriptive statistics both for the individual indicators of student assessment and for the four combined assessment categories. The measures derived from the PISA background questionnaires are shares bounded between 0 and 1, whereas the other assessments measures are dummy variables.[88] As is evident, some assessment practices are more common than others. For

[86] The variables in each category are calculated as proportionate usage in terms of the specific indicators for each country and wave. Note also that indicator data entirely missing for specific PISA waves are imputed by country-specific linear interpolation of assessment usages, a procedure that retains the entire country-by-wave information but that does not influence the estimated impact of the test category because of the inclusion of imputation dummies in the panel estimates (see Data Appendix for details). 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 3.4).

[87] 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.

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

example, 89 percent of schools in our country-by-wave observations use some form of assessment to inform parents, but only 29 percent have national standardized exams in lower secondary school. Table 3.1 provides country-by-country statistics of the initial and final value of the four separate indicators of standardized external comparison. Of particular relevance, there is a tendency for increased prevalence of the measures of standardized external comparison over time.

For our estimation, the amount of variation over time within individual countries in the different types of test usage is key. To understand the overall patterns of change in our data, Figure 3.2 shows histograms of the 15-year change in the combined measures of the four assessment 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 supportive of our estimation strategy based on a country-level panel approach.[89] Importantly, there is also wide variation in the change of 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 standardized external comparison than for the other three categories, leading us to expect higher precision (lower standard errors) of the coefficient estimates for this category. To provide a more fine-grained picture of the wave-to-wave variation, Figure 3.3 depicts the evolution of using standardized assessments for school-based external comparison from 2000 to 2015 for each country. The increasing use of such external assessments in many countries is quite evident. For example, in five countries, the share of schools that are externally compared with student assessments increased by more than 50 percentage points (Luxembourg, Denmark, Italy, Portugal, and Poland), and in another 18 countries, the share increased by more than 20 percentage points. In three countries, by contrast, the share decreased by more than 20 percentage points (Tunisia, Costa Rica, and Croatia).

[89] The exception in this depiction is internal testing. However, the reduction in this aggregate measure is fully accounted for by a change in the wording of the questionnaire item on the use of assessments to inform parents, where the word “assessments” was replaced by the word “standardized tests” in the 2015 questionnaire (see Appendix Table A 3.2). 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 testing 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 testing are unaffected by dropping the year 2015 from the analysis.

While perhaps desirable, it is beyond the scope of this chapter to provide detailed anecdotal narratives of specific policy reforms that underlie the changes in student assessment measures documented by the PISA school background questionnaires. No data source provides consistent external documentation of the time pattern of different testing policies, forcing us to rely upon the actual school implementation pattern identified by the principals at the time of each testing wave. However, on a number of occasions, it is straightforward to link major policy reforms directly to the overall pattern of expanded accountability measures. For example, the strong increase in school-based assessments used for external comparison in Italy in 2009, clearly visible in Figure 3.3, coincides with the introduction of the *Invalsi* national test.[90] Similarly, the increased external assessment in Denmark in 2006 reflects the 2006 *Folkeskole* Act that introduced a stronger focus on evaluation, assessment, and accountability including national tests (Shewbridge et al. 2011). And the strong increase in external assessments in Luxembourg shows the introduction of standardized national assessments that monitor student outcomes in French, German, and mathematics (Shewbridge et al. 2012). 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.

3.3 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

[90] See https://it.wikipedia.org/wiki/Test_INVALSI.

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

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

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + \beta X_{ct} + \mu_c + \mu_t + \varepsilon_{ict} \quad (3.6)$$

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

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

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

differences across countries.[91] This implies that countries that do not change their assessment practices over the observation period will not enter into the estimation of β.

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

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

3.4 Results

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

[91] Some recent investigations of scores on international assessments have focused on differential effort levels of students across countries (see, for example, Borghans & Schils 2012; Zamarro, Hitt & Mendez 2016; Gneezy et al. 2017; Balart, Oosterveen & Webbink 2018. 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 & Demmrich 2001).

3.4.1 Strength of incentives across usage categories

We start our discussion of results with the average effects of the different categories of student assessment in our country sample. Table 3.3 presents the results for the combined measures of the four assessment categories, first entered separately (Columns 1-4) and then jointly (Columns 5-7). All models are estimated as panel models with country and year fixed effects, conditioning on the rich set of control variables at the student, school, and country level indicated above.[92] Regressions are weighted by students' sampling probabilities within countries, giving equal weight to each country-by-wave cell across countries and waves. Standard errors are clustered at the country level throughout.

Overall, the basic impact results displayed in Table 3.3 suggest that different forms and dimensions of student assessments have very different effects on student achievement. Among the four assessment categories, only standardized testing that is used for external comparisons has a strong and statistically significant positive effect on student outcomes. The coefficients on standardized monitoring and internal testing are insignificant and close to zero, whereas there is quite a sizeable negative coefficient on internal teacher monitoring.[93] These different impacts are consistent with the predictions on differing strengths of incentives from the conceptual discussion.

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

[92] Appendix Table A 3.1 shows the coefficients on all control variables for the specification of the first column in Table 3.6.

[93] Note that, consistent with the larger within-country variation of standardized external comparisons over time documented in Section 3.2.2, the standard error associated with this coefficient estimate is smaller. Still, even with the smaller standard error of this variable, the coefficient estimates on standardized monitoring and internal testing would be far from statistical significance.

monitoring even turns significant in math. With that nuanced exception, results for science and reading achievement are very similar to those for math (Columns 6 and 7).

To establish that our aggregation is not suppressing important heterogeneity within our four categories, we present individual results for each of the 13 underlying country-level indicators of student assessment going into our test usage categories. Table 3.4 displays the estimates for the separate indicators that underlie our aggregates, where each cell represents a separate regression.[94] Of particular interest, each of the four elements of the external comparison composite, with one exception, has a significantly positive impact on student performance in the three subjects. The exception is the use of central exit examinations, which could simply reflect that student performance measured by PISA at age 15 is not very responsive to rewards that only occur at the end of secondary school (when students are usually aged around 18 or 19). While the point estimates are positive in all three subjects, they do not reach statistical significance.[95] The estimated coefficients for the other three indicators taken separately are smaller than the combined measure. As noted, this probably reflects a reduction in measurement error for the correlated indicators and the fact that the different incentives are not perfect substitutes, implying that the combined impact across categories is greater than that for any individual component.[96] We return below to a consideration of separate components of external comparisons as related to schools and to students.

At the individual indicator level in Table 3.4, there is also some evidence of positive effects of standardized testing in the relevant grade for PISA, and some indication of impact from the use of assessment to inform parents. None of the other indicators of standardized monitoring without external comparison, of internal testing, and of internal teacher

[94] The separate regressions of Table 3.4 do not employ any imputation of the separate treatment variables. Thus, the number of countries and waves included in each estimation varies and is determined by the availability of the specific assessment indicator. The fact that these results confirm the previous results of the four combined categories shows that the latter cannot be driven by the interpolated imputations required for the aggregation of the separate indicators.

[95] Consistent with the weaker evidence on central exit exams, constructing the combined measure of standardized external comparison without the central exit exam measure (i.e., based on the other three underlying indicators) yields a slightly larger coefficient estimate of 30.926 in the specification of column 5 of Table 3.3.

[96] A third possibility is that the estimation samples for the separate indicators are varied and smaller than for the combined indicator. However, we reject this explanation because estimating the combined model in column 5 of Table 3.3 just for the smallest sample of countries in the separate indicator models yields a virtually identical coefficient for external comparisons.

monitoring is significantly related to student achievement on average. The individual estimates suggest that the potential negative impact of the internal monitoring of teachers is driven by the two subjective components – monitoring by the principal and by external inspectorates. The aggregate categorical variable is larger than these two subcomponents, potentially again reflecting a reduction in measurement error and possible additivity.

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

3.4.2 School-based versus student-based external comparisons

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

Table 3.5 disaggregates the standardized external comparisons into school-based and student-based external comparisons (each of which is based on standardized exams that have meaning across schools).[97] This table presents simultaneous estimates that include

[97] The measure of student-based external comparison is the simple average of the three underlying indicators of standardized external comparison except for the one on school-based external comparison. Note that the estimates of Table 3.5 are based on smaller student samples from fewer countries, because data on student-based external comparison are available for few countries beyond OECD and European Union countries.

the other three categories. Both school and student incentives are strongly positive and statistically significant, with estimates for the school-based incentives being somewhat larger than for the individual student incentives. At the same time, none of the estimates for the remaining categories are qualitatively affected. The results suggest that focusing incentives on different actors yields different responses and leads to separate effects on outcomes.

3.4.3 Environmental differences in usage impact

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

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

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

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + \beta_1 X_{ct} + \beta_2 (X_{ct} \times A_{c0}) + \mu_c + \mu_t + \varepsilon_{ict} \quad (3.7)$$

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

Table 3.6 presents estimates of the interacted model for the three subjects. The left three columns provide results for the aggregate category of standardized external comparisons, while the right three columns divide the external comparisons into school-based and student-based comparisons. The initial score is centered on 400 PISA points (one standard deviation below the OECD mean). The precise patterns of estimated effects by initial achievement with confidence intervals are displayed in Figure 3.4 for math performance.

In broad generalities, the picture of how the overall achievement environment interacts with the incentives from different test usage can be summarized as follows. First, the impact of standardized external comparisons is stronger in lower achieving countries and goes to zero for the highest achieving countries. In particular, at an initial country level of 400 PISA points the introduction of standardized external comparison leads to an increase in student achievement of 37.3 percent of a standard deviation in math. With each 100 initial PISA points, this effect is reduced by 24.6 percent of a standard deviation. At an initial level of 500 PISA points (the OECD mean), the effect of standardized external comparison is still statistically significantly positive at around 13 percent of a standard deviation in all three subjects. Second, standardized monitoring similarly creates significant incentives in initially low-achieving countries, with effects disappearing for higher-achieving countries (i.e., those with initial scores of roughly above 490 in all subjects). Third, the estimate of internal testing is insignificant throughout the initial-achievement support. Fourth, the estimates for internal teacher monitoring are insignificant for most of the initial-achievement distribution and turn negative only at high levels of initial achievement in math (perhaps reflecting the purely linear interaction). Fifth, when external comparisons are disaggregated into school-based and student-based components, school-based comparisons follow essentially the same heterogeneous pattern as overall standardized external comparisons but go to zero for a somewhat larger set of initially high-achieving countries. By contrast, the impact of student-based external comparisons does not vary significantly with initial achievement levels.

The disaggregated underlying individual indicators of standardized external comparison consistently show the pattern of significantly stronger effects in initially poorly performing countries (Appendix Table A 3.4).[98] Interestingly, the introduction of central exit exams – which did not show a significant effect on average – also shows the pattern of decreasing effects with higher initial achievement, in particular in science. Similarly, all three underlying indicators of standardized monitoring also show the same pattern of significant positive effects at low levels of achievement and significantly decreasing effects with initial achievement. Thus, the positive effect of standardized testing in low-

[98] 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.

achieving countries appears to be quite independent of whether the standardized tests are used for external comparison or just for monitoring. This finding supports the World Bank attention to testing for low achieving countries (World Bank 2018d).[99]

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

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

3.5 Specification Tests

In this section, we come back to a discussion of the identifying assumptions of our specification and a series of tests of their validity. We start with a placebo test and then get to a number of additional analyses.

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

3.5.1 A placebo test with leads of the assessment variables

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

A leading remaining concern of the fixed-effects model is that reforms may be endogenous, in the sense that reforming countries may already be on a different trajectory than non-reforming countries for other reasons, thus violating the usual common-trend assumption of the fixed-effects model. Here the largest concern is that countries that are on a downward trend turn to expanded testing and accountability to reform the system. Note that, if generally true, this would tend to bias our estimated effects downward.

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

Table 3.7 reports the results of this placebo test. As is evident, none of the lead variables of the four assessment categories is significantly related to student achievement (i.e., in the wave before reform implementation). At the same time, the results of the contemporaneous assessment measures are fully robust to conditioning on the lead variables: The use of standardized external comparison has a significant positive effect on the math, science, and reading achievement of students *in the year in which it is implemented*, but not in the wave in which it is not implemented yet. Moreover, the

estimated coefficients for the usage categories are qualitatively similar to those in Table 3.3.[100]

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

Estimating the full interacted model with all four assessment categories and their leads interacted with initial achievement is overly demanding to the data. Nevertheless, focusing just on the main results of Section 3.4, an interacted model that includes just standardized external comparison, its lead, and their interactions with initial achievement gives confirmatory results: standardized external comparison is significantly positive, its interaction with initial achievement is significantly negative, and both the lead variable and its interaction with initial achievement are statistically insignificant (not shown).

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

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

3.5.2 Additional discussion and analysis

Another important possible remaining 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

[100] 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 3.10 below, results are very similar in that specification.

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.

Therefore, all our estimations include information on the time pattern of autonomy reforms for each country. Consistent with prior work (Hanushek, Link & Woessmann 2013), our results confirm that the effect of school autonomy on student achievement is negative in developing countries but positive in developed countries in this extended setting.[101] Importantly, the results on assessment effects are not confounded by the potentially coincidental introduction of policies that alter school decision making and autonomy.

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 categories of testing usage 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 as opposed to public school management, and school 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. To the extent that these are the subject of other contemporaneous policy changes, they would be controlled for.

In fact, some of these school-level variables – in particular, those capturing the composition of teachers – could potentially be endogenous to the testing reforms. However, as the first column of Table 3.8 indicates, qualitative results are unaffected by

[101] 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.

leaving the teacher controls out of the specification for math achievement. The same is true for achievement in science and reading (not shown).

Another approach to gauge the potential relevance 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. As shown in column 2, dropping all covariates from the model does not change the results. 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 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 & Taber 2005).

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 advantage of incorporating several changes per country. 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 (as suggested by the model with testing lags alluded to above). By restricting identification to overall changes, we can both reduce the potential influence of measurement error and gauge the long-run relevance of the policy reforms. Column 3 provides estimates from a model in long differences that considers just the total 15-year change from the first to the last PISA wave. Our main findings are robust in this long-difference specification. Consistent with larger measurement error in shorter-frequency change data, the estimate of the positive effect of standardized external comparison is larger when considering only long-run changes. The estimates of effects of the other three assessment categories remain insignificant.

The long-difference analysis provides a convenient way to illustrate the main results about how changes in standardized comparisons translate into achievement gains. Figure 3.5 displays the added-variable plot for the impact of introducing standardized assessments used for external comparison. It clearly shows that countries that expanded the use of standardized testing for external comparison from 2000 to 2015 saw the achievement of their students improve.

Relatedly, 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, whereas the former may carry particular interest from

a policy perspective. As discussed in Section 3.2.2, the implementation path 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 implementation 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 and Eurydice measures of national standardized testing represent coding by country specialists of the changes in assessment policies – just the kinds of well-identified policy changes that would enter into micro policy evaluations.

While we prefer the combined assessment measures in our baseline specification, it is important to note that the two dummy measures of standardized external comparisons are separately significant in their impact on overall student performance (see second and third lines in Table 3.4). Thus, the more gradual measure of usage of external comparison in schools and the discontinuous reform indicators of formal national policies yield very similar results, indicating that our results do not depend on adopting one of the specific perspectives.

As indicated in Table 3.9, also the results of the interacted specification are unaffected by dropping the teacher controls or all controls (Columns 1 and 2). Similarly, 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 standardized external comparisons (column 4).[102]

To check that the negative effects of standardized monitoring without external comparison and internal teacher monitoring at high levels of initial achievement (indicated in Figure 3.4) are not an artefact of the imposed linearity of the interaction model, Columns 5-8 of Table 3.9 report results of a specification that interacts each of the four assessment categories with four dummies reflecting the four quartiles of initial

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

country achievement. There is no indication of strong nonlinearity.[103] 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. This result may suggest that introducing standardized monitoring without external comparison and internal teacher monitoring in systems that are already performing at a high level may in fact detract teacher attention from more productive forms of instruction.

3.6 Robustness analyses

Our results prove robust to a number of potentially contaminating factors. In particular, we consider possible peculiarities of our country sample, possible effects of student and school exclusions from PISA testing, and possible interactions with changes in PISA testing. For ease of exposition, we present robustness results without heterogeneity by country achievement level in the text (Table 3.10) and the heterogeneity results, which yield similar conclusions, in Appendix Table A 3.5.

To ensure that our results are not driven by the peculiarity of any specific country, we re-ran all our main models (the simultaneous regressions of Columns 5-7 in Table 3.3 and Columns 1-3 in Table 3.6) excluding one country at a time. The qualitative results are insensitive to this, with all significant coefficients remaining significant in all regressions (not shown).

To test whether results differ between developed and less developed countries, we split the sample into OECD and non-OECD countries. As the first two columns of Table 3.10 show, qualitative results are similar in the two subgroups of countries, although the positive effect of standardized external comparison is larger in OECD countries. Patterns of heterogeneity by achievement level are less precisely identified within the two more homogeneous subgroups (Appendix Table A 3.5). In the group of OECD countries, the significant effect of standardized external comparison does not vary significantly with initial achievement, but the demands of the fully interacted model make estimation

[103] The pattern for internal teacher monitoring also has a rather steady pattern when entered without the other three assessment categories (92.3, -3.7, -36.6, and -102.5), suggesting that the joint specification with four interactions of four assessment measures may be rather demanding to depict precise patterns. The separately estimated patterns for the other three measures also indicate rather linear relationships.

difficult with just the 35-country sample. When we drop the insignificant interactions (column 2), the point estimate of the use of standardized scores for comparisons is significant. The heterogeneous effect of standardized monitoring is somewhat more pronounced in OECD countries. But overall, the patterns do not differ substantively between the two country groups.

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 2016c). The average total exclusion rate is below 3 percent, but it varies from 0 percent to 9.7 percent across countries and waves. To test whether this variation affects our analysis, the next column in Table 3.10 (and Appendix Table A 3.5) controls for the country-by-wave exclusion rates reported in each PISA wave. As is evident, results are hardly affected.

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

First, the simplest test of whether our analysis is affected by the 2015 changes in testing methodology is to drop the 2015 wave from our regressions. As is evident from column 4 in Table 3.10 (and column 5 in Appendix Table A 3.5), qualitative results do not change when estimating the model just on the PISA waves from 2000 to 2012, indicating that our results cannot be driven by the indicated changes in testing mode.

Second, to address the changes in the psychometric scaling procedure, PISA recalculated countries' mean scores in the three subjects for all PISA waves since 2006 using the new 2015 scaling approach. In the final column of Table 3.10 (and Appendix Table A 3.5), we run our model with these rescaled country mean scores instead of the original individual scores as the dependent variable for the PISA waves 2006 to 2015. Again, qualitative

results do not change, indicating that the changes in scaling approach do not substantively affect our analysis.

Third, while no similar analysis is possible for the change in testing mode, we analyzed whether countries' change in PISA achievement from paper-based testing in 2012 to computer-based testing in 2015 is correlated with a series of indicators of the computer familiarity of students and schools that we derive from the PISA school and student background questionnaires in 2012. As indicated by Appendix Table A 3.6, 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.

3.7 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 – meaning assessments that enter into reward and incentive systems for some individuals – are inappropriate (Koretz 2017). Others argue that increased use of testing and accountability systems are essential for the improvement of educational outcomes (World Bank 2018d) and, by extension, of economic outcomes (Hanushek et al. 2015; Hanushek & Woessmann 2015).

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

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

The empirical analysis employs the increasingly plentiful international student assessment data that now move toward providing identification of causal implications of national testing.[104] Specifically, the six waves of the PISA assessments between 2000 and 2015 permit country-level panel estimation that relies on within-country over-time analysis of country changes in assessment practices. We combine data across 59 countries to estimate how varying testing situations and applications affect student outcomes.

Focusing on international comparisons has both advantages and costs. A variety of 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 possible influences on student outcomes. Some issues of measurement error, imprecise wording of questionnaire responses, and other possible influences on student outcomes 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.

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

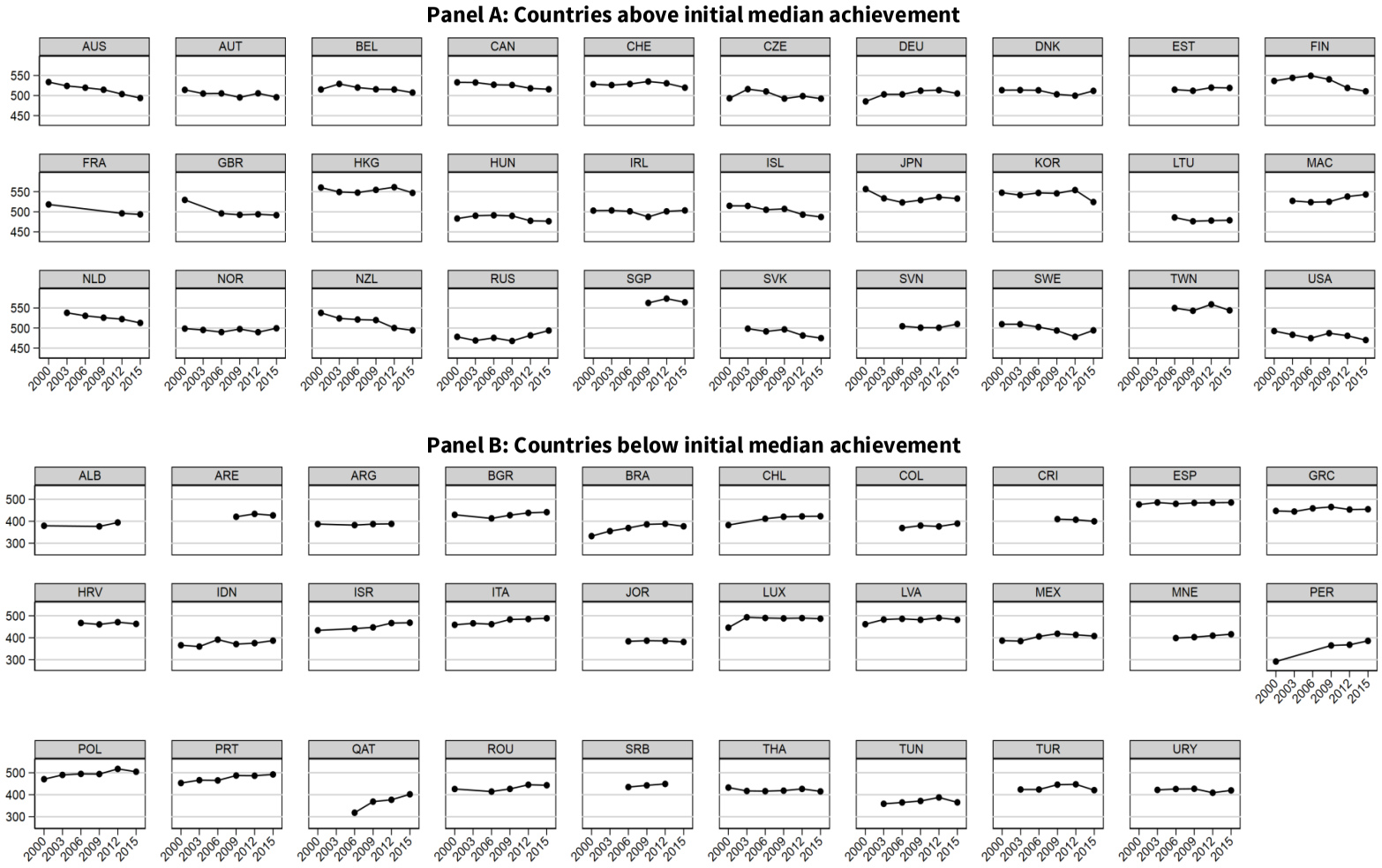
[104] Interestingly, even the international testing – conducted on a voluntary basis in a low-stakes situation – has come under attack for potentially harming the educational programs of countries. Recent analysis, however, rejects this potential problem (Ramirez, Schofer & Meyer 2018).

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

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

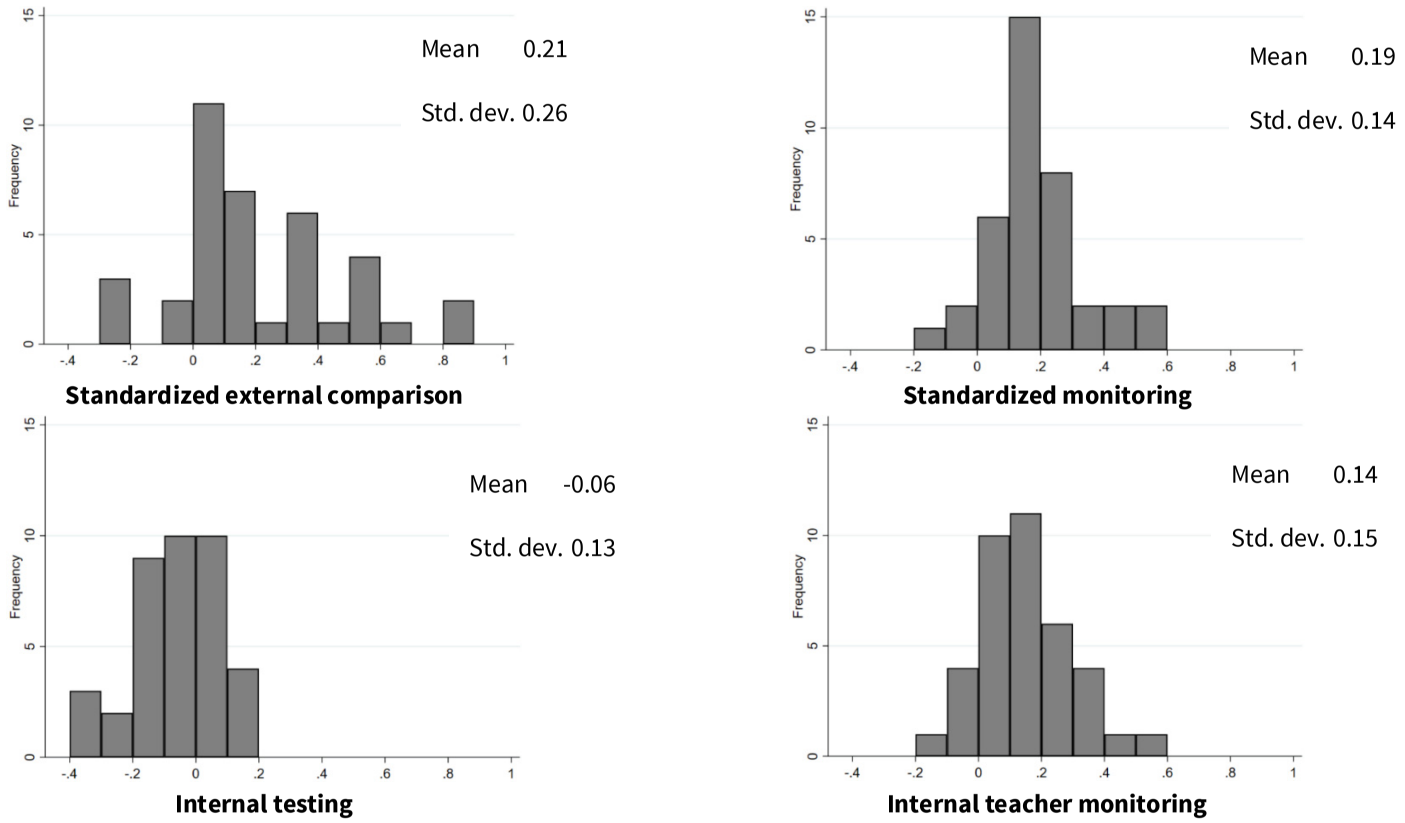
Overall, the results from international comparisons of performance suggest that school systems gain from measuring how their students and schools are doing and where they stand in a comparative way. Comparative testing appears to create incentives for better performance and allows rewarding those who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

Figure 3.1: PISA math achievement in 2000-2015



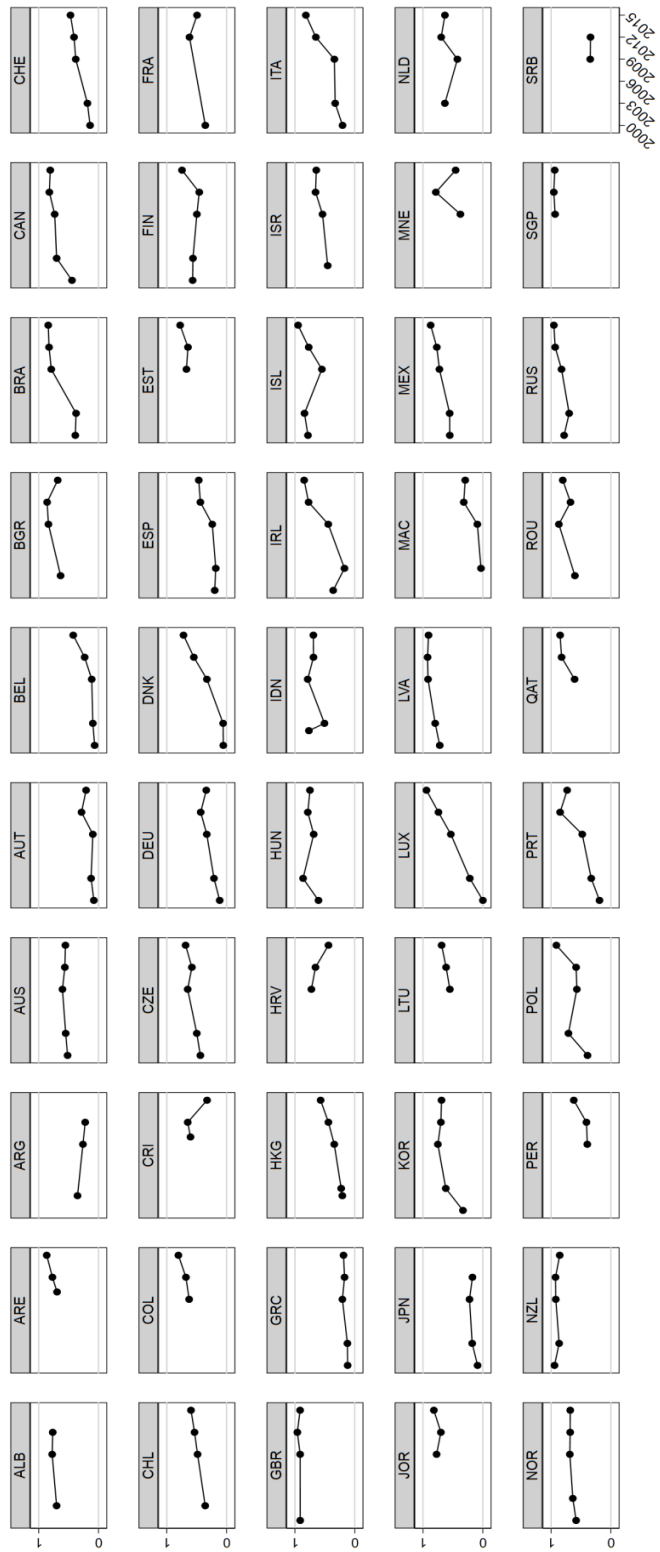
Notes: Country mean achievement in PISA math test. Country sample split at median of initial achievement level for expositional reasons. Country identifiers are listed in Table 3.1. Own depiction based on PISA micro data.

Figure 3.2: Histograms of change in four categories of student assessments, 2000-2015



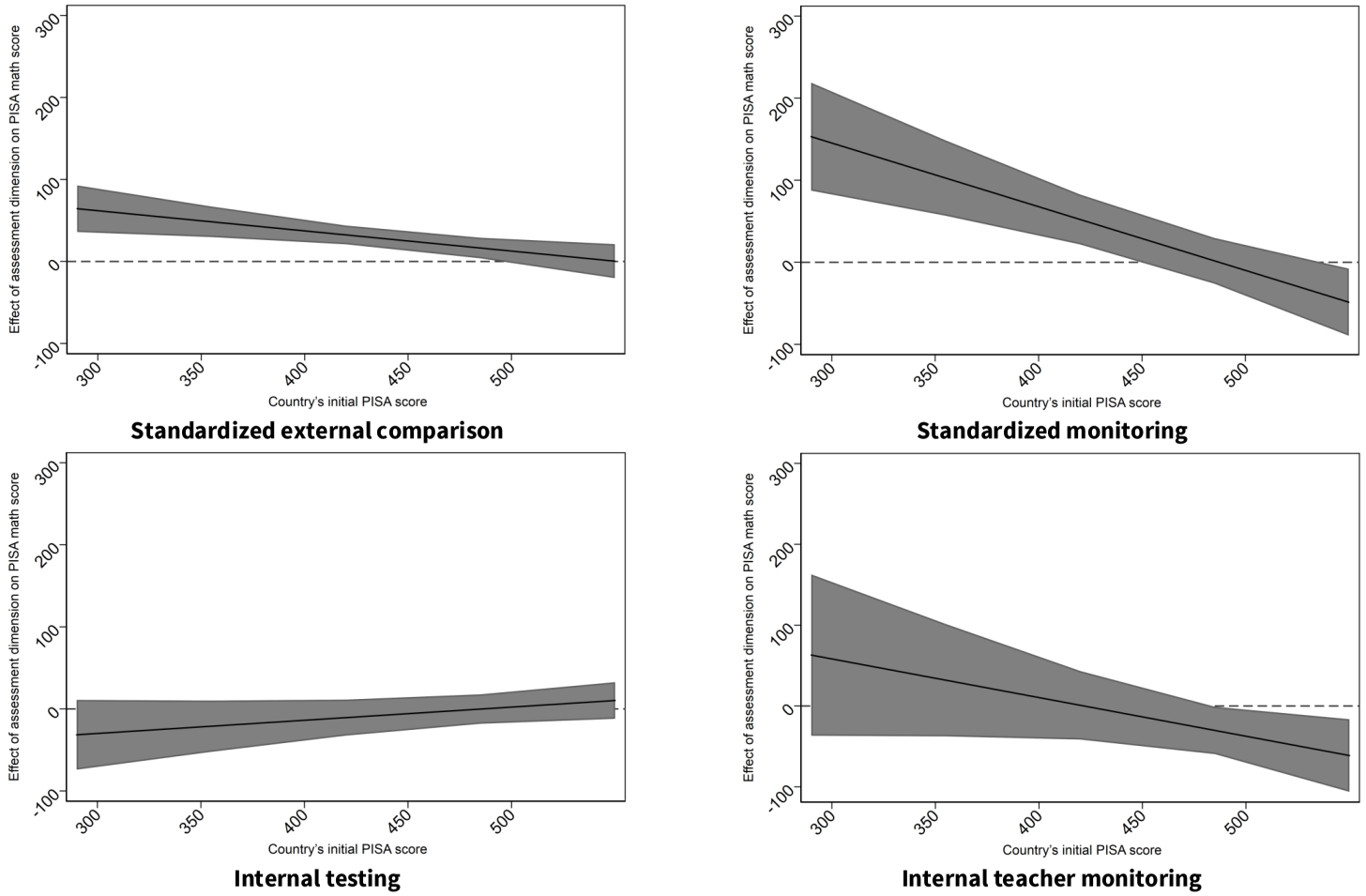
Notes: Histograms of change between 2000 and 2015 in the four combined measures of student assessment for the 38 countries observed both in the first and last PISA waves.

Figure 3.3: School-based external comparison in 2000-2015



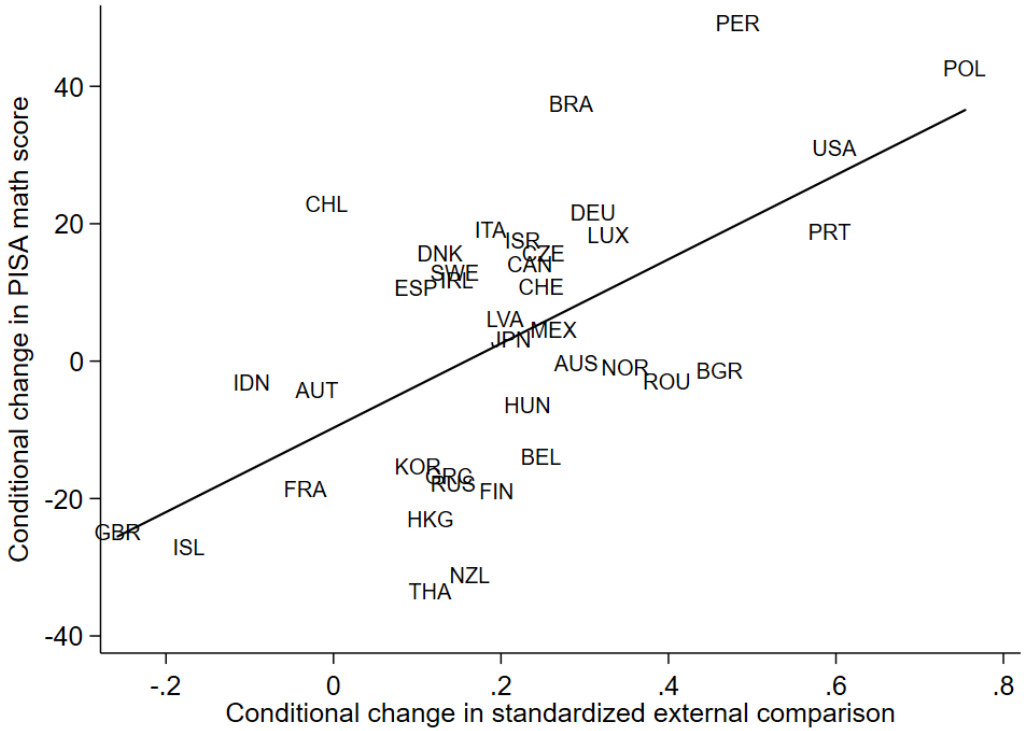
Notes: Country share of schools with use of assessments for external comparison. Country identifiers are listed in Table 3.1. Own depiction based on PISA micro data.

Figure 3.4: Effect of student assessments on math performance by initial achievement levels



Notes: Average marginal effects of student assessments on PISA math score by initial country achievement, with 95 percent confidence intervals. See first column of Table 3.6 for underlying model.

Figure 3.5: Fifteen-year changes in use of 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 use of standardized testing for external comparison, 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 3.8 for underlying model.

Table 3.1: Selected indicators by country

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

(continued on next page)

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

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

Table 3.2: Descriptive statistics of assessment measures

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

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

Table 3.3: The effect of different dimensions of student assessments on student achievement: Fixed-effects panel models

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

Notes: Dependent variable: PISA test score in subject indicated in the header. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. Control variables include: student gender, age, parental occupation, parental education, books at home, immigration status, language spoken at home; school location, school size, share of fully certified teachers at school, teacher absenteeism, shortage of math teachers, private vs. public school management, share of government funding at school; country's GDP per capita, school autonomy, GDP-autonomy interaction; imputation dummies; country fixed effects; year fixed effects. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3.4: Baseline model for separate underlying assessment indicators

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

Notes: Each cell presents results of a separate regression. Dependent variable: PISA test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 3.3 for included control variables. Number of observations and R^2 refer to the math specification. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3.5: Disaggregation of standardized external comparisons into school-based and student-based comparisons

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

Notes: Dependent variable: PISA test score in subject indicated in the header. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 3.3 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3.6: Effects of student assessments by initial achievement level: Fixed-effects panel models

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

Notes: Dependent variable: PISA test score in subject indicated in the header. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Initial score: country's PISA score in the initial year (centered at 400, so that main-effect coefficient shows effect of assessments on test scores in a country with 400 PISA points in 2000). Sample: student-level observations in six PISA waves 2000-2015. See Table 3.3 for included control variables. Complete model of specification in column 1 displayed in Table A 3.1. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3.7: Placebo test with leads of assessment reforms

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

Notes: Dependent variable: PISA test score in subject indicated in the header. Lead indicates values of test usage category from subsequent period, i.e., before its later introduction. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 3.3 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3.8 Specification tests: Base specification

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

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 3.3 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3.9: Specification tests: Interacted specification

	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 external comparison	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 monitoring	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 testing	-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)*					
Internal teacher monitoring	-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 3.3 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3.10: Robustness tests: Base specification

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

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. Rescaled test scale available for waves 2006-2015 only. See Table 3.3 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Appendix

A Data appendix: sources and construction of assessment measures

We derive a series of measures of different categories of the use of student assessments over the period 2000-2015 from the PISA school background questionnaires and other sources. Information on testing usage is classified into four groups with varying strength of generated incentives: standardized external comparison, standardized monitoring, internal testing, and internal teacher monitoring. We aggregate each assessment measure to the country-by-wave level. Below, we also discuss how we combine the different indicators into an aggregate measure for each of the four assessment categories. Details on the precise underlying survey questions and any changes in question wording over time are found in Appendix Table A 3.2.

A.1 Standardized external comparison

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

First, from the PISA school background questionnaires, we measure the share of schools in each participating country that is subject to assessments used for external comparison. In particular, school principals respond to the question, “In your school, are assessments of 15-year-old students used to compare the school to district or national performance?” Figure 3.3 in the text provides a depiction of the evolution of this measure from 2000 to 2015 for each country.

Second, in the 2015 version of its Education at a Glance (EAG) publication, the OECD (2015) published an indicator of the existence of national/central examinations at the lower secondary level together with the year that is was first established. The data were collected by experts and institutions working within the framework of the OECD Indicators of Education Systems (INES) program in a 2014 OECD-INES Survey on Evaluation and Assessment. National examinations are defined as “standardized student tests that have a formal consequence for students, such as an impact on a student’s eligibility to progress to a higher level of education or to complete an officially-recognized

degree" (OECD (2015), p. 483). According to this measure, five of the 37 countries with available data have introduced national standardized exams in lower secondary school between 2000 and 2015.[105]

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

Fourth, Leschnig, Schwerdt & Zigova (2017) compile a dataset of the existence of central exit examinations at the end of secondary school over time for the 31 countries participating in the Programme for the International Assessment of Adult Competencies (PIAAC). They define central exit exams as "a written test at the end of secondary school, administered by a central authority, providing centrally developed and curriculum based test questions and covering core subjects." Following Bishop (1997), they do not include commercially prepared tests or university entrance exams that do not have direct consequences for students passing them. Central exit exams "can be organized either on a national level or on a regional level and must be mandatory for all or at least the majority of a cohort of upper secondary school." We extend their time period, which usually ends in 2012, to 2015. Five of the 30 countries in our sample introduced central exit exams over our 15-year period, whereas two countries abandoned them.

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

A.2 Standardized monitoring

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

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

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

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

A.3 Internal testing

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

In particular, school principals also report on the use of assessments of 15-year-old students in their school for purposes other than external comparisons. Our first measure of internal testing captures whether assessments are used “to inform parents about their

child's progress." The second measure covers the use of assessments "to monitor the school's progress from year to year." Each measure is coded as the share of schools in a country using the respective type of internal assessments.

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

A.4 Internal teacher monitoring

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

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

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

A.5 Constructing combined measures for the four assessment categories

Many of the separate assessment indicators are obviously correlated with each other, in particular within each of the four groups of assessment categories. For example, the correlation between the EAG measure of national standardized exams in lower secondary school and the Eurydice measure of national tests used for career decisions is 0.59 in our

pooled dataset (at the country-by-wave level) and 0.54 after taking out country and year fixed effects (which reflects the identifying variation in our model). Similarly, the two internal-testing measures of using assessments to inform parents and using assessments to monitor school progress are correlated at 0.42 in the pooled data and 0.57 after taking out country and year fixed effects (all highly significant).

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

Our construction of the combined measures takes into account that the different indicators are available for different sets of waves and countries, as indicated in Appendix Table A 3.3. Before combining the indicators, we therefore impute missing observations in the aggregate country-by-wave dataset from a linear time prediction within each country. We then construct the combined measures of the four assessment categories as the simple average of the individual imputed indicators in each category. To ensure that the imputation does not affect our results, all our regression analyses include a full set of imputation dummies that equal one for each underlying indicator that was imputed and zero otherwise.

The combined measures of the four assessment categories are also correlated with each other. In the pooled dataset of 303 country-by-wave observations, the correlations range from 0.278 between standardized external comparison and internal teacher monitoring to 0.583 between standardized monitoring and internal testing. After taking out country and year fixed effects, the correlations are lowest between standardized external comparison and all other categories (all below 0.2), moderate between standardized monitoring and the other categories (all below 0.3), and largest between internal testing and internal teacher monitoring (0.485). Because of potential multicollinearity, we first run our analyses for each aggregate assessment category separately and then report a model that considers all four categories simultaneously.

B Appendix tables

Table A 3.1: Descriptive statistics and complete model of basic interacted specification

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

(continued on next page)

	Descriptive statistics			Basic model	
	Mean	Std. dev.	Share imputed	Coeff.	Std. err.
School characteristics					
Number of students	849.0	696.7	0.093	0.012***	*(0.002)*
Privately operated	0.193	0.383	0.071	7.500*	*(4.396)*
Share of government funding	0.802	0.289	0.106	-16.293***	*(4.596)*
Share of fully certified teachers at school	0.822	0.294	0.274	6.662**	*(2.793)*
Shortage of math teachers	0.202	0.394	0.041	-5.488***	*(1.031)*
Teacher absenteeism					
No	0.337	0.427	0.213		
A little	0.484	0.447	0.213	-0.325	*(1.175)*
Some	0.140	0.310	0.213	-6.089***	*(1.556)*
A lot	0.039	0.173	0.213	-7.715***	*(2.413)*
School's community location					
Village or rural area (<3,000)	0.092	0.281	0.056		
Town (3,000-15,000)	0.208	0.397	0.056	5.238***	*(1.768)*
Large town (15,000-100,000)	0.311	0.451	0.056	9.935***	*(2.148)*
City (100,000-1,000,000)	0.251	0.422	0.056	14.209***	*(2.594)*
Large city (>1,000,000)	0.137	0.336	0.056	17.482***	*(3.447)*
Country characteristics					
Academic-content autonomy	0.597	0.248	-	-11.666	*(8.826)*
Academic-content autonomy × Initial GDP p.c.	5.043	7.578	-	1.871***	*(0.475)*
GDP per capita (1,000 $)	27.30	20.80	-	0.009	*(0.123)*
Country fixed effects; year fixed effects				Yes	
Student observations	2,193,026			2,094,856	
Country observations	59			59	
Country-by-wave observations	303			303	
R^2				0.393	

Notes: Descriptive statistics: Mean: international mean (weighted by sampling probabilities). Std. dev.: international standard deviation. Share imputed: share of missing values in the original data, imputed in the analysis. Basic model: Full results of the specification reported in first column of Table 3.6. Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability. Regression includes imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A 3.2: Measures of student assessments: Sources and definitions

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

(continued on next page)

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

Notes: Own depiction based on indicated sources.

Table A 3.3: Country observations by wave

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

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

Table A 3.4: Estimations for separate underlying assessment indicators: Interacted specification

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

Notes: Two neighboring cells present results of one separate regression, with "main effect" reporting the coefficient on the variable indicated in the left column and "× 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 3.4 for numbers of observations, countries, and waves and Table 3.3 for the included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A 3.5: Robustness tests: Interacted specification

	OECD countries		Non-OECD countries	Control for exclusion rates	Without 2015	Rescaled test scale
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized external comparison	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 monitoring	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 testing	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)*
Internal teacher monitoring	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 3.3 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A 3.6: Correlation of computer indicators in 2012 with change in PISA score from 2012 to 2015 at the country level

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

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

4 Teacher specialization and student gender [106]

Classic economic theory assigns productivity gains to the division of labor because specialization on a certain task allows to improve skills and to increase the likelihood for innovation. As a result, a comparative advantage arises (Smith 1776). This understanding derives from production in the 18th century, where processes were divisible and productivity was evaluated in piece per worker. In the recent past, this understanding has been applied to assembly line production, for example in the automotive industry. A key economic production factor is human capital. While human capital production is not simple and learning output not measured in piece rates, knowledge is divisible into subjects of different fields. Specialization on certain (fields of) subjects may generate a comparative advantage. Thus, specialized teachers may increase student achievement. Yet, teacher specialization seems hardly studied in the economics of education.

Germany provides an ideal setting to examine teacher specialization during studies because candidates usually choose two subjects at the beginning of their tertiary studies and later teach exclusively those subjects.[107] We consider teachers specialized if they choose subjects of one field, such as math, informatics, natural sciences, and technology (MINT) or languages. For example, teachers study two out of four languages (English, German, French, and Latin) or they study two of the seven MINT subjects (math, informatics, natural sciences – including physics, chemistry, biology, and geography – and technology).[108] Teachers without specialization combine languages or MINT with any other subject.[109] For example, non-specialized teachers study English and arts or

[106] This chapter is joint work with Raphael Brade of University of Göttingen and University of Erfurt.

[107] Theoretically, teachers can study and teach more (or less) than two subjects. As the average number of subjects is 1.96 in sciences and 1.78 in languages, we refer to two subjects for simplicity, especially when comparing specialized and non-specialized teachers. In empirical analysis, we control for the number of subjects studied. Furthermore, some subject combinations are not allowed in some Bundesländer. Bavaria does not allow to combine biology with German or French, and German not with informatics. In Rhineland Palatine, a specialization on languages is impossible. Saxony-Anhalt allows field specializations, but choice is limited. In Thuringia, a specialization in sciences is impossible and a specialization on languages is unlikely, because it necessitates to combine another language with Latin.

[108] MINT refers to the subjects studied or taught by teachers including math, informatics, physics, chemistry, biology, and geography. Sciences refer to physics, chemistry, and biology skills of students tested in our data.

[109] Other subjects are music, arts, history, politics, religion, sports, and subjects which are neither MINT nor languages.

chemistry and music. Our approach contrasts existing studies, which have examined teacher specialization on one subject based on the highest value-added in student test scores of teachers in-service (Jacob & Lefgren 2008; Condie, Lefgren & Sims 2014; Fryer 2018). The two approaches have different policy implications. We emphasize the long-term consequences of initial teacher training, while existing studies focus on optimizing the use of personnel in-service.

From a theoretical perspective, there may be positive and negative effects of teacher specialization. On the one hand, specialized teachers affect student achievement positively for several reasons. Specialization creates a deeper understanding of the field which results in a broader content to offer to the students (Anderson 1962). Specialized teachers are likely more confident and more respected. As a result, their lessons may be more interesting and encouraging to their students (Jacob & Rockoff 2011; Fryer 2018). Furthermore, well-versed teachers may have more capacity to concentrate on the pedagogical ongoings in the classroom, which may also originate from economies of scale in lesson planning and lower coordination costs when teachers specialize (Fryer 2018). Another reason for greater effectiveness of specialized teachers may originate from teachers specializing during their studies with a long time until classroom teaching in which they can develop superior skills. As a result, one expects specialized teachers to increase student achievement more than non-specialized teachers.

On the other hand, specialized teachers may be too far from their students' reality to successfully convey knowledge. At worst, students may feel intimidated and discouraged in their skills and may even lose interest in the subject because blinkered specialists may be impervious to the ongoings in the classroom beyond subject content and may lack pedagogical skills (Anderson 1962; Fryer 2018).[110] Consequently, specialized teachers may not be able to increase student achievement more than non-specialized teachers. Taking together positive and negative expectations on teacher specialization, it is unclear whether specialized or non-specialized teachers are best to maximize student achievement.

[110] While the lack of pedagogical skills remains unanswered by the literature, it is no concern for this chapter because specialized and non-specialized teachers receive the same pedagogical training during their studies in Germany and we assume no selection into treatment based on initial pedagogical skills.

To study the effects of teacher specialization we exploit within-student-between-subject variation. This approach eliminates most student-teacher sorting as well as biases due to unobserved subject-invariant student characteristics. This approach is commonly used to estimate effects of teacher characteristics on students' academic achievement (Dee 2005; 2007; Schwerdt & Wuppermann 2011; Metzler & Woessmann 2012; Bietenbeck 2014; Bietenbeck, Piopiunik & Wiederhold 2018; Hanushek, Piopiunik & Wiederhold 2018; Falck, Mang & Woessmann 2018). We apply student-fixed effects to two waves of the German National Assessment Study (*Ländervergleich* 2012 and *Bildungstrend* 2015), surveying representative samples of the ninth-grade population (Lenski et al. 2016; Pant et al. 2017; Schipolowski et al. 2018; Stanat et al. 2018). In 2012, student skills are tested in biology, physics, and chemistry. In 2015, skills are tested in German and English. This allows us to identify from between-subject variation of closely related subjects of one field – sciences or languages – in each wave. This is an advantage over previous studies which used variation between math and languages (Metzler & Woessmann 2012; Bietenbeck, Piopiunik & Wiederhold 2018; Hanushek, Piopiunik & Wiederhold 2018) or between math and sciences (Schwerdt & Wuppermann 2011; Bietenbeck 2014; Falck, Mang & Woessmann 2018). As a result, the assumption that teacher characteristics exert similar effects across subjects is more credible. The biggest remaining concern to identification is teacher selection into specialization. We control for a large set of teacher characteristics and consider whether they form before or after specialization, i.e., whether the characteristics are exogenous mechanisms or endogenous to the treatment. We claim to cover the most relevant characteristics and to avoid classic omitted variable bias (OVB).

Our analysis suggests a significant effect of teacher specialization on student achievement for boys by 0.061 standard deviations (SD) in sciences and 0.087 SD in languages but not for girls. The effects are stable across school tracks, student achievement, and socio-economic status. Potential mechanisms are student attitudes towards the subjects, such as self-concept and subject interest, which is affected by teacher specialization. In sciences, teacher specialization significantly increases subject interest of boys and decreases subject interest of girls. In contrast, there seems to be no statistically significant influence of teacher specialization on self-concept. The estimate of teacher specialization on student achievement conditional on student attitudes remains unchanged. In languages, self-concept and subject interest seem to increase with teacher specialization for boys but not for girls. The estimate of teacher

specialization on student achievement seems to decrease conditional on student attitudes. Hence, self-concept and subject interest are well-identified, non-causal mechanisms of transmitting teacher specialization on student achievement. Other potential mechanisms do not yield significant estimates neither as outcome nor as control variables to student achievement, such as teacher-student gender match and further teacher characteristics including self-efficacy, job satisfaction, discipline, teaching style, internal differentiation, and self-assessed teaching skills.

We show in robustness checks that different aggregation methods of self-concept and subject interest from their sub-concepts do not alter our findings. Furthermore, results are constant across the sub-skills in subjects, i.e., insight and knowledge in sciences and reading and listening in languages. Using grades as alternative skills measures indicates that specialized teachers grade girls systematically and statistically significantly worse than boys conditional on their actual skills. This potentially originates from teachers and girls themselves underrating girls' skills (Tiedemann & Faber 1998; Bertrand 2011). Additionally, grades and science skills appear uncorrelated (0.408 at p-value 0.000), hinting at different skills relevant for grading opposed to skills tested in the survey.

We consider an alternative definition of specialization – by teaching instead of studying. A teacher is specialized in several subjects from one field (MINT or languages) based on the subjects she or he teaches opposed to non-specialized teachers who teach subjects from more than one field. We formerly defined specialization by the subjects studied. The estimate of specialization by teaching is positive and significant but smaller than specialization by studying. Conditioning on both specializations renders specialization by teaching insignificant and leaves specialization by studying significant. This suggests that specialization by studying is the more relevant determinant of student achievement in the German setting compared to specialization by teaching.[111]

We mitigate remaining threats to identification from teacher selection into specialization by two tests. First, we use certain teacher characteristics as outcome to teacher specialization. Second, we condition on those teacher characteristics in the main specification. Additionally, the gender-specific treatment effect reassures us that specialized teachers are not simply better teachers because we would then expect to find

[111] We do not know whether specialization by teaching (on one field) is similar to the highest value-added (in one subject) of teachers.

a treatment effect on girls as well. Lastly, we address the concern that the effect of teacher specialization may be driven by non-specialized teachers who choose easy subjects such as music, arts, history, politics, religion, sports, and other subjects which are neither MINT nor languages. We construct two new comparison groups of non-specialized teachers. Non-specialized teachers who combine subjects from MINT or languages with easy subjects and non-specialized teachers who combine subjects from MINT with languages and vice versa. In sciences, we find that the effect of teacher specialization on boys is only significant when compared with non-specialized teachers who choose easy subjects. In languages, we find that the effect of teacher specialization is larger when comparing to non-specialized teachers who choose MINT subjects.

This chapter relates to two strands in the literature: Teacher specialization and the gender gap in student achievement. First, the paper linked closest to ours regarding teacher specialization is a field experiment in primary school by Fryer (2018). He assigned teachers to one subject based on their highest value-added score and finds a negative effect of -0.11 SD over one year on a combined test score of math and reading. His main explanation for the deterioration through specialization is inefficient pedagogy due to less interaction of teachers with each student, outweighing the comparative advantage in knowledge. Fryer (2018)'s finding underlines Bishop (2006)'s research on the importance of pedagogy and personal relationship for academic achievement. Similar to Fryer (2018)'s tests on instruction methods and job satisfaction, we condition on teacher's self-efficacy, their opinion on the profession, classroom discipline, teaching style, internal differentiation, and self-assessed teaching skills. These controls may be endogenous because specialized teachers may have more capacity for internal differentiation. Thus, specialized teachers yield higher student achievement because of their free capacity and not because of their specialization. This may be problematic for our approach if more efficient teacher candidates select into specialization. Most of our estimates of teacher specialization on teacher characteristics and of teacher characteristics as controls to teacher specialization are insignificant. This suggests that those teacher characteristics are not a mechanism that transmits teacher specialization into student achievement. In contrast, Fryer (2018) finds a negative effect of teacher specialization on job satisfaction and on above-median job performance. The discrepancy of Fryer (2018)'s and our findings may be due to the different definitions of specialization. Fryer (2018) assigns specialization to one subject based on the highest value-added, while we assign specialization to one field of several subjects based on

subject choice during studies. We do not know whether specialized teachers create higher value-added. From the different definitions follow different mechanisms: Fryer (2018) argues that the pedagogical component of teaching is attenuated due to less teacher-student interaction, while our specialized and non- specialized teachers do not vary in this dimension. This time component is also an issue in other primary school studies (Anderson 1962; Gerretson, Bosnick & Schofield 2008). Usually, non-specialized teachers in primary school teach their class all subjects throughout the whole day. As a result, teachers can follow students' development closely. Implementing a study where teachers only teach subjects with their highest value-added necessitates that a class is taught by several teachers and not by one teacher. As a result, teachers with the highest value-added in math teach math for an hour per day to several classes instead of non-specialized teachers teaching math, languages, and arts to one class for at least three hours per day. Hence, this existing approach does not isolate the treatment effect of specialization from time spent with the class. In contrast, our approach allows to regard teacher specialization independently of time spent together because specialization is determined during studies and time spent together occurs in-service. In our analysis, teachers already spend only one hour per day with their class and teach other classes. Thus, our treatment does not necessitate to redesign school schedules. Instead, we compare teacher background – the subject combination studied – independently of the time spent in class.

Further evidence on teacher specialization based on the highest value-added finds 0.05 SD in math and 0.03 SD in reading (Jacob & Lefgren 2008; Condie, Lefgren & Sims 2014). Our estimates for boys are between 0.061 SD in sciences and 0.087 SD in languages. This is surprising given that our definition of specialization during studies on a field with several subjects differs from their definition of specialization in-service to one subject based on teacher value-added. Additionally, teacher training differs between Germany and the United States providing differing subject and pedagogical skills.[112] In practice, our definition refers to teacher training generating resources to the teaching process, while earlier interventions refer to school organization of the available resources. In summary, our approach highlights the importance of initial teacher training.

[112] We discuss differences in teacher training between Germany and the United States in Section 4.1.

A second strand of the literature examines the gender gap, i.e., the differences in academic achievement between girls and boys. For example, the performance gap in math between girls and boys widened between 2000 and 2015 in Germany (from -15 to -18 points), while it decreased in the rest of the world (from -12 to -8 points). In contrast, the performance gap in reading between girls and boys shrank between 2000 and 2015 worldwide (from 32 to 17 points in Germany and from 32 to 26 points internationnally), compare Table A 4.1). This gap seems to arise over the course of school (Fryer & Levitt 2010) through societal factors rather than through biological factors, such as innate ability (Guiso et al. 2008).[113] Hence, the gender gap seems to arise from nurture and not from nature. The gender gap in academic achievement operates at the institutional and at the individual level. In reality, more gender-equal countries appear to close the gap (Guiso et al. 2008; Nollenberger, Rodríguez-Planas & Sevilla 2016; Lippmann & Senik 2018) and boys seem to enjoy competition more than girls (Niederle & Vesterlund 2010; Bertrand 2011; Reuben, Wiswall & Zafar 2017).[114] Competitiveness may transfer to student achievement by self-esteem, which seems to be less pronounced among girls (and more pronounced among boys) leading to lower (higher) educational attainment (Araujo & Lagos 2013).[115] Even with the same objective performance, girls seem to underestimate

[113] Examining the origins of gender differences, culture comes into play. Societal factors of the gender gap were investigated early by Gneezy, Niederle & Rustichini (2003) in laboratory experiments suggesting that men are more competitive compared to women in competitive environments. Subsequently, Gneezy, Leonard & List (2009) found that women may be equally competitive as men, when part of the dominant group of a culture, such as the matrilineal Khasi in India opposed to the patrilineal Maasai in Tanzania. These two environment-specific studies show that the societal environment drives the gender gap.

[114] Investigations on the influence of gender on student achievement are often set in single-sex schools opposed to coeducational settings. Studies suggest positive effects for girls' achievement from single-sex schooling due to less disruptions (Lavy & Schlosser 2011; Link 2012). Single-sex schools may even increase interest, self-efficacy, expectations, and actual college major choice of MINT for boys, but not for girls (Park, Behrman & Choi 2018). In contrast, studies identifying from reorganizing single-sex to coeducational school seems to lower academic achievement for girls and boys (Dustmann, Ku & Kwak 2018). In summary, existing evidence disagrees on the direction of effects of single-sex as opposed to coeducational settings.

[115] Self-esteem relates to the broader field of non-cognitive skill contributing to educational achievement (Heckman & Rubinstein 2001; Cunha & Heckman 2007). Non-cognitive skills comprise motivation, tenacity, trustworthiness, perseverance and so on. Intrinsic motivation is difficult to separate from ability as it is often self-reported or derived from response times, item response rates, or performance over the course of a test (Eklöf 2010; DeMars & Wise 2010; Zamarro, Hitt & Mendez 2016). When effort, i.e., motivation, can be incentivized exogenously, it seems to raise performance in low-stakes tests holding ability constant (Gneezy et al. 2017).

their skills and expect lower results. This may be due to two reasons. First, girls seem to attribute failure to their lacking skills and less to lacking effort (Tiedemann & Faber 1998; Bertrand 2011). Second, teachers seem to perceive girls as less competent, i.e., less able, than boys. If a self-fulfilling prophecy comes to effect, girls yield lower performance than boys (Tiedemann 2000). This chapter captures a student's self-concept in several sub-categories.[116] We find that teacher specialization insignificantly decreases girls' self-concept by 0.11 SD in sciences, which confirms the literature's reasoning.

The main contribution of this chapter is threefold. First, we are the first to examine the influence of teacher specialization on student outcomes defined by initial teacher training independent of time spent with students. This implies that policy makers and researchers may revisit initial teacher training rather than managing in-service teachers. Second, we investigate gender-specific effects of teacher specialization in secondary school when a gender gap in achievement has developed opposed to previous studies in primary school. Third, we investigate the German setting while most studies focus on the United States where teacher training differs as described in the following section.

This chapter continues in Section 4.1 with background information on teacher training. Section 4.2 presents the empirical strategy applied to the data introduced in Section 4.3. Section 4.4 reports estimation results and Section 4.5 concludes.

4.1 Background on teacher training

Germany's teacher training is distinct from teacher training in the United States, which most studies focus on. Germany employs the concurrent model of teacher studies (*Lehramt*) where candidates usually study academic content in two subjects and

[116] In 2012, self-concept is predefined as of four sub-categories (self-assessed performance, satisfying grades, fast learning, rating relative to other subjects) and in 2015 of seven items (mostly understand the subject, fast learning, effort, talent, no hopeless case, less difficulties than peers, needing less time than peers).

pedagogical knowledge together with all teacher candidates.[117] An advantage of the concurrent model is that learning is more integrated, linking academic content to pedagogical knowledge. The disadvantage is that the program is less flexible than consecutive training, since candidates need to decide in the beginning of their university career whether they want to become a teacher or not. This excludes many candidates who feel the calling to become a teacher at a later stage (Musset 2010; Hoffmann & Richter 2016).

Becoming a teacher later in life is one of the key differences between the German and the American teacher training system. While the United States offer alternative routes into teaching for candidates without traditional teacher training, Germany is very restrictive in qualifying candidates who did not initially opt for teacher studies.[118] This limitation for late teacher candidates in Germany is one reason for our approach – once teachers have studied two subjects, they usually teach them and new candidates hardly ever become teachers. Hence, teacher specialization has a long time to manifest and evolve.

This difference in teacher training may be a reason why the U.S. studies find no effect of initial teacher training on student achievement. This means that there seems to be no significant link between teacher candidates' grades from undergraduate studies or college exams and student achievement (Harris & Sass 2011). This may be caused by the diverse qualification and employment backgrounds of teacher in the United States. While it seems more common to leave the teaching profession in the United States, 72 percent of German teachers are civil servants and have little incentive to leave their profession (Statista 2012). Hence, with less variation in the German teacher population than in the American, the link between teachers' university and high school grades and student

[117] Theoretically, teachers can study and teach more (or less) than two subjects. As the average number of subjects is 1.96 in sciences and 1.78 in languages, we refer to two subjects for simplicity, especially when comparing specialized and non-specialized teachers. In empirical analysis, we control for the number of subjects studied. Furthermore, some *Bundesländer* do not allow all subject combinations. Bavaria does not allow to combine biology with German or French, and doesn't allow German to be combined with informatics. In Rhineland Palatine, a specialization on languages is impossible. Saxony-Anhalt allows field specializations, but choice is limited. In Thuringia, a specialization in sciences is impossible and a specialization in languages is unlikely because it necessitates to combine another language with Latin.

[118] Some *Bundesländer*, such as Berlin or North-Rhine Westphalia, have recently been experiencing severe shortages in their teaching staff and increasingly employ teachers without subject qualification (*fachfremd*) or candidates outside of the teaching profession (*Quereinsteiger*).

achievement may be stronger. Yet, German empirical evidence is limited as linking teachers with achievement information to students with achievement information is restricted.[119] For example, Enzi (2017) applies a value-added model to NEPS and investigates the influence of teacher candidates' grades in teacher state exams and high school exams on student achievement. He finds that teachers at the top performance quartile link to significantly higher student achievement and more efficient classroom management. His result suggests that teachers who are more effective in increasing student achievement may be identified at university entry and exit. Thus, one may expect a significant link of teacher grades and student achievement in Germany opposed to the zero-effect in the United States. Overall, this highlights the importance of initial teacher training in Germany.

German teacher training at university ends after four to six years with a state exam testing theoretical knowledge in the two subjects and in pedagogy. The mandatory practical teacher training (*Referendariat*) of 18 to 24 months follows, where candidates teach classes, prove their teaching skills in demonstration lessons, and take oral and written exams. Teacher training is completed with a second state examination (Terhart 2007; Musset 2010).

In summary, teacher training in Germany is more inflexible than in the United States in two ways that define our empirical strategy. First, German teachers fix their subjects at the beginning of their studies and they usually only teach those subjects throughout their career. Second, teacher candidates without traditional teacher training are scarce. In contrast, American teacher candidates have more alternative paths to enter the profession at a later stage and have more options to switch the subjects they teach.

4.2 Empirical strategy

Estimating the effect of teacher specialization on student achievement with ordinary least squares (OLS) poses two problems for causal interpretation. First, specialized teachers may not be randomly distributed across students because students may sort

[119] Some few data sets allow linking teacher achievement to student achievement, such as the German National Educational Panel Study (NEPS), which may be accessed by application.

into schools and classrooms based on their (or their parents') preferences for teachers. For example, high ability students may prefer specialized teachers because more apt students value deeper understanding of contents by specialized teachers. Thus, any model that does not account for sorting will likely provide biased estimates. Previous studies on the effect of teacher characteristics have addressed this issue by using student-fixed effects. The approach identifies from within-student-between-subject variation, eliminating issues of sorting into schools and classrooms from subject-invariant student characteristics, such as ability. An advantage of our variable of interest is that teachers cannot adjust their specialization acquired by studying, i.e., before the time of the treatment, or adjust their teaching specialization once assigned for the school year in reaction to the students assigned.

The second problem for analysis with OLS arises if teachers select into specialization based on unobserved teacher characteristics that correlate with student achievement. This poses a problem for identification with OLS and student-fixed effects. We aim at mitigating this problem by controlling for a rich set of teacher characteristics and investigating the influence of teacher specialization on those characteristics.

To employ student-fixed effects models, one needs several observations per student – either over time or across subjects.[120] We rely on multiple observations per student across different subjects at the same point in time. One assumes that the differences in achievement between subjects originate from differences in teacher specialization between the subjects. We observe each student in two language subjects and in up to three science subjects at a time. We follow the literature on student fixed effects (see Dee 2005; 2007; Schwerdt & Wuppermann 2011; Metzler & Woessmann 2012; Bietenbeck 2014; Bietenbeck, Piopiunik & Wiederhold 2018; Hanushek, Piopiunik & Wiederhold 2018; Falck, Mang & Woessmann 2018) by estimating the effect of teacher specialization on student achievement with the following equation:

$$Y_{ijs} = \alpha + \beta\, TS_{ijs} + \gamma\, X_{js} + \lambda_i + \varepsilon_{ijs} \qquad (4.1)$$

[120] Longitudinal data with several observations per student over time allows for a value-added model that uses student achievement as the outcome variable on the left-hand side of the equation and lagged achievement of the same student as the control variable on the right-hand side to account for the history of educational production up until the period of the lag. Yet, longitudinal data of individual students is scarce in Germany.

Where Y_{ijs} denotes student i's test score in subject s taught by teacher j which is determined by teacher specialization TS_{ijs} and by a vector of other teacher and class characteristics, X_{js}. β is our coefficient of interest. λ_i is the student-fixed effect controlling for subject-invariant determinants of test score. ε_{ijs} is a student-by-subject-specific error term.

When estimating the effects of teacher specialization, we rely on four assumptions. First, we assume that the effect from teacher specialization on student achievement does not vary across subjects. This seems credible in our case, as we use very similar subjects, such as chemistry, physics, and biology or German and English.

Second, subject-invariant covariates have the same relation to student achievement in the different subjects. For example, student gender remains constant for German and for English.

Third, and most importantly, we assume that conditional on the observed background characteristics, teacher specialization is distributed randomly across teachers within students. This means that if there is an unobserved characteristic that determines both selection into treatment and student achievement, estimates will be biased downwards or upwards. For a downward bias, specialized teachers may be more content-related thinkers and may give less attention to the emotional relationship with their students. If those unemotional teacher candidates rather select into specialization than emotional teachers and if students need a warm relationship to their teachers to achieve high test scores, the true effect of teacher specialization on student achievement will be underestimated. Hence, this teacher characteristic decreases student achievement through another mechanism than teacher specialization and estimates will be downward biased. As downward bias works against finding an effect, this bias is not as worrisome as the following bias. For an upward bias, teacher specialization is non-randomly distributed across teachers with a characteristic that promotes student achievement. For example, specialized teachers may advocate standardized student assessments because they are more conscientious, which makes them adhere to the tested curriculum. At the same time, more conscientious teachers may rather select into teacher specialization. As a result, students of specialized teachers achieve higher student achievement due to conscientious teachers but not due to teacher specialization. In summary, non-random distribution of teacher specialization with characteristics promoting higher student

achievement conjures an effect where there is none. In comparison to the first bias of underestimation, the latter is more detrimental to identification. All studies estimating the effect of teacher characteristics on student outcomes are threatened by this omitted variable bias. We address this issue by controlling for a rich set of teacher and class characteristics and test whether teacher specialization alters those characteristics. It is reassuring that we find effects of teacher specialization on student achievement only for boys but not for girls. This indicates that specialized teachers are not *per se* different than non-specialized teachers.

The fourth identifying assumption implies that learning accumulates over the whole school time. This is threatened by the excellent teacher in the year before we observe students created an advantage in student performance and persists, which the current average skilled teacher cannot replicate. Student-fixed effects cannot capture previous learning as the method uses observations from one period only. Therefore, student-fixed effects models necessitate one of the following assumptions. First, either full decay of past inputs, i.e., there is no aggregate learning process. This is most improbable. Second, full consideration through subject-invariant-fixed effects with exogenous covariates, which still suffers from potential omitted variable bias. Third, no correlation with current teachers' characteristics relevant for the treatment. This assumption is rather credible and may be approached by controlling for teacher characteristics.

4.3 Data

Achievement data and student and school background information originate from the German National Assessment Study in 2012 (*Ländervergleich*) and 2015 (*Bildungstrend*) (Lenski et al. 2016; Pant et al. 2017; Schipolowski et al. 2018; Stanat et al. 2018). The survey targets a representative set of ninth graders drawn in two stages. First, in each *Bundesland* schools are drawn randomly. Second, within schools, one class (Gymnasium

track) or two classes (other tracks) are drawn randomly and all students of a class participate.[121]

We use the following subject-specific student outcomes from the student questionnaire: test scores, self-concept, and subject interest. Student performance is psychometrically scaled with a mean of 500 test-score points and a standard deviation of 100 test-score points, which we standardize to mean zero and standard deviation one. The 2012 assessment tests skills in biology, physics, and chemistry.[122] The 2015 assessment tests skills in German and English. In reality, ninth-grade students in the highest track and most lower tracks are obliged to take biology, chemistry, physics, German, and English.[123]

Both waves report skills in sub-dimensions of plausible values (PV): subject knowledge (*Fachwissen*) and insight (*Erkenntnisgewinn*) in sciences; and reading and listening in languages. We use PV1 and our results are robust to using the other PVs. We construct a global test score for each subject by standardizing the mean of the standardized sub-

[121] In 2012, two classes per school participated. In 2015, one class per school participated. Hence, we receive a different number of classes and schools in 2012, while the number of classes and schools are the same in 2015.

[122] Math skills were also tested in 2012 which we do not use for our analysis for three reasons. First and foremost, only some of the students that were tested in math, were also tested in sciences (4,921 out of 24,709). As we explain in more detail at the end of this section, in 2012 only around 30 percent of the students can be linked to a teacher. We therefore would yield about 1,500 additional observations if we include math. Second, we aim at comparing very similar subjects to adhere to the identifying assumption of the student-fixed effects model demanding for teacher characteristics to be randomly distributed across subjects. In our view, biology, chemistry, and physics are closer related to each other than to math. Third, math is measured in different sub-skills (numbers, measuring, space and form, functional relationship, data and coincidence) compared to sciences (insight and knowledge).

[123] 9th graders are obliged to take all three science subjects (biology, chemistry, and physics) and both languages (German and English) in the highest track (*Gymnasium*). In the lower tracks (*Realschule*, *Mittelschule*, *Hauptschule*, or the joint-tracks of *Gesamtschule*), natural sciences are sometimes combined in one subject which is tested in the same manner as for the separate subjects (Jansen et al. 2014). This is the case for *Hauptschule* in Baden-Wurttemberg, *Mittelschule* in Bavaria, in Lower Saxony in the *Oberschule*, in North-Rine Westphalia in the *Sekundarschule* and the *Hauptschule*, in Rhineland-Pallatine chemistry and physics are joint in *Hauptschule*, and natural sciences in the *Gesamtschule*, in Saxony-Anhalt and Schleswig-Holstein and Thuringia all three as natural sciences in the *Gemeinschaftsschule*. In Hesse, there is no physics in *Realschule* and no biology in *Hauptschule*. The fact that not all lower tracks in all Bundesländer teach all subjects poses no computational problem, as we keep only students in the sample who are taught in all subjects. Yet, this subject limitation across tracks may increase sample selection.

dimensions in each subject. We also use the standardized sub-scores in robustness tests to check whether it is one skill domain driving the overall effect.

The subject-specific self-concept is a pre-defined aggregate of several items. In 2012, the items are “not good in the subject”, “receive good grades in the subject”, “fast learning in the subject”, and “one of the best subjects” answerable with the options “fully applies”, “rather applies”, “does rather not apply”, “does not apply at all”. We also construct an index of self-concept by taking the simple average of the standardized sub-concepts that we standardize. In 2015, the items of students’ self-concept are “I mostly understand the subject matter”, “I have more difficulties in this subject than others in my class”, “no matter how hard I try, I cannot cope with this subject”, “I do not have a talent for this subject”, “I need more time than others to solve tasks”, “I am no hopeless case”, and “I learn fast”. The answer options are the same as above. This wave also provides a pre-defined aggregate index with and without imputations and we construct a standardized average of the standardized sub-concepts.[124]

The subject-specific interest in 2012 and in 2015 derives from the pre-defined aggregate index of the items “personally important”, “enjoy”, “interest”, and “one of my favorite activities” answerable by the options “fully applies”, “rather applies”, “does rather not apply”, “does not apply at all”. We also construct a standardized average of the standardized sub-concepts. 2015 again provides a pre-defined aggregate index with imputations.

Another potential student outcome are grades. Yet, grades are a mixture of performance and social relations, which complicates their interpretation. Testing mechanisms and robustness of our findings, we use grades standardized to mean zero and standard deviation of one.[125]

[124] Self-concept is missing for 51.5 percent for the observations in 2012 and for subject interest 72.3 percent. In 2015 the data provides several iterations of imputed values: 36.1 percent for self-concept and 36.3 percent for subject interest. We use the first iteration in our estimations but the results are robust to using the other iterations. To eliminate concerns about the construction of the overall index and imputed data, we construct our own index from the sub-concepts and show that our findings are robust.

[125] In Germany, the highest grade is 1 and the lowest grade is 6. We take the inverse of the grade scale to facilitate interpretation (higher values of the variable indicate higher performance).

To construct our treatment variabl, we use two items asked of the teachers: the subjects teachers obtained a teaching qualification for and which subjects they teach in the current school year. We define teacher specialization by studying only subjects from one field – either MINT or languages. In 2012, the science field consists of the MINT subjects math, informatics, biology, chemistry, physics, and geography.[126] We observe students skills in biology, chemistry, physics, and we have information on teacher studies in all MINT subjects. Overall, 81 percent of physics teachers specialized in MINT subjects during their studies, 79 percent of chemistry teachers and 43 percent of biology teachers. In 2015, the language field consists of German, English, French, and Latin. We observe students skills in German and English and have information on teacher studies in all four languages. On average, 26 percent of German teachers specialized in languages during their studies and 42 percent of English teachers did.

Similar to teacher specialization by studying, we observe teacher specialization by teaching, which is the current subject combination taught. 79 percent of teachers taught physics and another MINT subject, 73 percent taught chemistry and another subject, and 44 percent taught biology and another subject. 25 percent of teachers taught German and another language and 40 percent taught English and another language.

We do not explicitly include the intensity of specialization, i.e., the number of subjects studied in the field, because the variance is low: On average, teachers studied 1.96 subjects with a standard deviation of 0.65.[127] Instead, we condition on the number of subjects a teacher studied, to avoid putting too much weight on teachers who studied only one subject and would be labelled specialist without any other subject studied. The second key teacher control is whether the teacher studied the subject she or he teaches.

Covariates of the student-fixed effect model from the teacher questionnaire are female, age, migration background, years as a teacher, years at this school, no standard teacher training (i.e., no *Lehramt*), the employment volume in hours per week, the employment contract, the type of teaching license, the type of institution of teacher training,

[126] There is no „technology“ subject.

[127] Specifically, 63 percent of teachers studied two MINT subjects, 23 percent studied one science subject, and 7 percent studied three MINT subjects. In contrast, 65 percent studied one language subject and 23 percent studied two language subjects.

participation in professional development in the last school year, and the total hours of professional development in the last school year. More endogenous teacher characteristics are the attitudes towards general student assessments and attitudes towards VERA.[128] Those attitudes may be correlated with the treatment because specialized teachers are rigorous and effective in increasing student achievement and therefore may support standardized assessments more than non-specialized teachers.

In a robustness test, we consider further teacher characteristics in the categories self-efficacy, job satisfaction, discipline, teaching style, internal differentiation, and self-assessed teaching skills. We generate those six aggregates from 55 items sorted in the categories by the survey questionnaire (details in Table A 4.6). We take the standardized mean of each item and standardize the index.[129]

The OLS controls at the student level contain: female, age, migration background, and grade repetition; at family level: number of books at home in six categories, mother or father born abroad, highest educational degree in the family (ISCED 1, 2, 3B and 3C, 3A and 4, 5B, and 6), highest EGP level in the family in five categories[130] (higher occupation (*obere Dienstklasse*), lower occupation (*untere Dienstklasse*), routine services in trade and administration, self-employed, skilled worker and worker with personnel responsibility), and the highest ISEI,[131] and at school level: city size in four categories (town with less than 3,000 to 15,000 inhabitants, large town with 15,000 to 100,000 inhabitants, city with 100,000 to 1,000,00 inhabitants, and large city with more than 1 million inhabitants), private operation, school size,[132] share of students with German mother tongue; and at classroom level: the number of students in a class. Table A 4.2 displays summary statistics.

[128] VERA (*Vergleichsarbeiten*) are German standardized student assessments in third and eighth grade.

[129] We have considered more elaborate methods of aggregating items into categories, such as principal component analysis (PCA) or factor analysis, but regression results are qualitatively the same.

[130] EGP signifies Erikson, Goldthorpe & Portocarero (1979) to categorize the socio-economic status of parental occupation.

[131] ISEI stands for International Socio-Economic Index, following Ganzeboom et al. (1992).

[132] School size is larger in 2012 than in 2015 because the largest category of 2015 includes all schools with more than 400 students (top-coded), while 2012 further differentiates large schools.

For heterogeneity analysis, we distinguish school track into the highest track, Gymnasium, and all other tracks.[133] Furthermore, we create achievement quartiles based on the average student performance across all subjects in sciences or languages. Lastly, we use the surveys' index of socio-economic status and create quartiles of socio-economic status.

To avoid losing observations from missing values in control variables, we replace missing values at the individual level with a constant and include dummies for the missing values for each control variable in our regressions.

The final sample is composed as follows. 24,796 students were tested in natural sciences in the 2012 wave of the German National Assessment.[134] After we drop all students that attend special education (*Förderschule*), we can link 8,583 to their biology teacher, 7,546 to their physics teacher, and 8,387 to their chemistry teachers. The low quota of student-teacher matches of under 50 percent is due to teachers not answering questionnaires or the part of the questionnaire on class information identifiers.[135] The low answer ratio among teachers may be because it was not mandatory for teachers in each Bundesland to answer the questionnaire. Further, we lose 6,790 students because we only observe them in one subject, and we need at least two observations per student for our fixed-effects approach. After these exclusions, we are left with a sample of 7,641 students who we observe in 2.3 subjects on average. The students are in 640 classes in 529 schools and are taught by 1,368 teachers.

[133] In the 16 German *Bundesländer* with independent education policies, tracking is mostly implemented after fourth grade based on previous student performance into high (*Gymnasium*), middle (*Realschule* or *Mittelschule*), and low track (*Hauptschule*). Some schools, the *Gesamtschule*, host several tracks within one institution, but classes are still tracked. Special education is located in the *Förderschule*, which we exclude completely.

[134] Some of those students were also tested in math. We focus on sciences subjects because the number of students that were tested in math and in science is rather low (4,921) and the way subject skills are tested differs between math (numbers, measuring, space and form, functional relationship, data and coincidence) and science (knowledge and insight). Further details on our reasoning are given in Footnote 122.

[135] The low share of student-teacher matches is also documented in the survey report, compare IQB (2013) Chapter 12.6 page 381. Additionally, the ideal teacher population is only vaguely defined, as teachers in biology, physics, chemistry, or math of tested classes. Yet, some of the teachers may have taught several of the four subjects to the tested class and other teachers only one subject. As a result, the number of teachers per school depends on subject combination per class.

In the 2015 German National Assessment, 34,982 (excluding special education) students were tested in English and German. Of those we can link 28,113 students to a teacher. We likely lose observations in this step because it was not mandatory to fill out the teacher questionnaire in all *Bundesländer* (IQB 2016). We exclude another 3,142 students for whom we do not observe test scores in German or English.[136] After dropping all students for which we do not observe test scores in both subjects, we arrive at the final sample of 19,223 students who are observed in both English and German from 868 classes in 868 schools (one class per school was sampled) and are taught by 1,736 teachers.

4.4 Results

A first take on estimating the link between teacher specialization and student achievement is to use ordinary least squares (OLS) with a range of controls, see Table 4.1. The OLS estimates can be compared with the estimates from our main specification, providing an intuition about the direction and strength of the bias from OLS. The table is structured by the two unconnected waves of tested skills in sciences and languages. The point estimate of teacher specialization with student, family, teacher, and school controls is sizeable but insignificant in sciences and small and insignificant in languages. Conditional on class-fixed effects, this pattern turns around. Because we have eliminated variation between classes, this change in coefficient hints at potential selection of student and teachers into schools and classes and supports the use of an empirical method eliminating selection into schools and classes, such as the student-fixed effects approach.

Interacting teacher specialization with student gender uncovers substantial heterogeneity in the effects between boys and girls. We find that teacher specialization increases achievement of boys by 0.036 SD in sciences and 0.072 SD in languages, when controlling for class-fixed effects. The negative interaction between teacher specialization and student gender leads to very small and insignificant effects for girls in both fields.

[136] Students of one class may distribute over several courses in English. Hence the target population of the study is unclear (refer to the technical report (IQB 2016)).

4.4.1 Main results

To eliminate further unobserved heterogeneity between individuals, we identify from within-student across-subject variation. This student-fixed effects approach holds constant subject invariant student characteristics, such as innate ability or motivation.

Table 4.2 presents the results of our main specification using student-fixed effects structured by the two unconnected waves of tested skills in sciences and languages with stepwise inclusion of teacher controls. As shown in Columns 1 and 7, teacher specialization has little to no overall effect on student skills. However, this effect masks heterogeneity by student gender. Columns 2 and 8 show the estimates without controls and with student-fixed effects. For boys, we find a significant advantage in skills from teacher specialization of 0.055 SD in science and 0.082 SD in languages. The interaction with student gender is significant and indicates that the effect of teacher specialization for girls is lower by 0.09 SD in science and 0.095 SD in languages. This leads to a significant negative effect of 0.35 SD for girls in sciences and to an insignificant effect for girls in languages that is close to zero. Including controls yields marginal changes in the coefficients. Columns 5 and 11 report the results of the specification we use for the following analyses including all previous controls and professional development. Coefficients are very similar to the preceding specifications. Columns 6 and 12 condition on teachers' attitudes towards assessments which may be endogenous, i.e., bad controls, because they are affected by the treatment. This is if specialized teachers tend to be proponents of assessments rather than non-specialized teachers. This may originate from specialized teachers may adhere more to the curriculum that is tested in assessments. Estimates in languages are more affected than in sciences. As a result, we do not condition on teachers' attitude towards assessments in the following.

Overall, compared to the OLS estimates with class-fixed effects, student-fixed effects yield larger estimates in absolute terms but qualitatively similar results. While OLS produces a good approximation of the link between teacher specialization and student achievement, it underestimates the positive effect of teacher specialization on boys' achievement.

Compared to other dimensions of teacher quality, this gender-specific effect from teacher specialization is not small. Teacher value-added from within-school variation yields an advantage of 0.13 SD in reading and 0.7 SD in math (Hanushek & Rivkin 2012), teaching

experience in elementary and middle school yields an advantage in student achievement of 0.02 SD to 0.06 (Harris & Sass 2011), or higher teacher numeracy skills yield an advantage in student achievement of 0.1 SD to 0.15 SD (Hanushek, Piopiunik & Wiederhold 2018). In summary, our estimate of teacher specialization of around 0.06 to 0.09 SD equals 17 to 26 percent of a year of learning.

Surprisingly, the effect of teacher specialization is similar across fields. One could have expected the estimates to differ between the fields for the following three reasons. First, comparative advantages may differ, e.g., complementarities in languages could be higher than in science or vice versa. Second, specialized teachers may differ in their characteristics, motivation, or ability between fields. Third, non-specialized teachers, i.e., our reference group, in sciences and languages may differ in their subject selection. We investigate the third reason in Section 4.4.5, where we explore whether non-specialized teachers choose easy or difficult subjects.

Overall, the results suggest positive effects from teacher specialization for boys but not for girls. This may hint at a different human capital production function for girls compared to boys.

4.4.2 Heterogeneity

This section investigates potential differences in the effect of teacher specialization on student achievement across school tracks, achievement quartiles, and socio-economic quartiles. Table 4.3 displays the main specification in the full sample, in lower tracks, and in the highest track. The estimate of teacher specialization on student achievement appears unaffected across the samples. Figure 4.1 depicts the coefficient plot of Table 4.3 to illustrate the various estimates. Table 4.3 and Figure 4.1 indicate no significant difference of the estimate of teacher specialization on student achievement across tracks. Still, our estimates tentatively suggest that girls in the highest track may be negatively affected by teacher specialization. This may be due to differences in average achievement and socio-economic status, which we explore next. Figure 4.2 reports results for achievement quartiles.[137] Together with Table 4.4, the analysis suggests that the estimates of teacher specialization on boys' achievement are stable across quartiles except for the lowest quartile in languages. The interaction of student gender and teacher

[137] We divide students in achievement quartiles based on their average achievement per field.

specialization seems less constant in magnitude but appears stable in the estimates' direction across both fields. Interestingly, teacher specialization seems to decrease achievement of high-achieving girls in sciences and languages.

Figure 4.3 and Table 4.5 present results for decomposing the effect of teacher specialization on student achievement by socio-economic status. In sciences, the effect of teacher specialization and the interaction with gender are robust across socio-economic quartiles. This pattern is similar in languages except for the highest quartiles where students seem largely unaffected by teacher specialization. This may originate from extensive home support in families of higher socio-economic status, which creates a ceiling effect. In summary, there seems to be little heterogeneity in the effect of teacher specialization on student achievement.

4.4.3 Mechanisms

This section aims at uncovering mechanisms of teacher specialization affecting girls and boys in a different way. One potential explanation for the heterogeneous effect across gender may be that teacher specialization affects academic attitude of girls and boys differently, which we measure by students' subject-specific self-concept and their subject-specific interest.

Table 4.6 reports estimates of teacher specialization on subject-specific self-concept and subject interest. We find that teacher specialization significantly increases self-concept and subject interest in languages for boys. In sciences, we find that teacher specialization increases the subject interest of boys, while it reduces the subject interest of girls. We find no significant effect on self-concept in sciences but the negative effect for girls is still sizeable.

We deconstruct the two indices into their sub-concepts, as observable in Table 4.7 Panel A shows the single items of self-concept in each field. All items were rescaled for easier interpretation in the same direction. The number of observations in Table 4.7 differs from Table 4.6 as the sub-concepts do not contain imputed values. In sciences, self-concept covers very different dimensions of self-concept. For example, asking for receiving good grades involves the feedback at school while fast learning is rather self-related. Consequentially, the effects vary substantially between the different sub-concepts in sciences, and we do not put much weight on these results. The two last sub-concepts

relate the students' self-assessed self-concept relative to their classmates and estimates suggest that students of specialized teachers report to have significantly less difficulties than their classmates and report to need insignificantly less time than their classmate. The net effect for girls is insignificant. Overall, the effects across the different sub-concepts in languages are far more consistent with each other than in sciences, which gives credibility to using the combined score. Panel B of Table 4.7 seems to yield more consistent estimates across the sub-concepts of subject interest within each field and across fields. The interaction of teacher specialization and female students yields highly significant similarly sized coefficients in sciences and insignificant similarly sized coefficients in languages. Overall, subject interest seems to be surveyed more consistently within and across fields compared to self-concept. In summary, teacher specialization seems to affect subject interest in sciences and languages but self-concept only in languages, which may be due to the survey questions that were used for self-concept in sciences. To examine whether those student attitudes are mechanisms of transmitting teacher specialization to student achievement, we condition on them in Table 4.8. This delivers explorative, non-causal evidence in a bad control manner because subject interest and self-concept in languages are themselves subject to the treatment, i.e., *well-identified*, while self-concept in sciences seems unaffected by the treatment, i.e., it is not well-identified.

Table 4.8 shows the effect of teacher specialization on student achievement conditional on student attitudes. The estimates suggest that there is no significant change in coefficients conditional on self-concept, subject interest or both, except for the language skills of boys, which decreases by about 22 percent. Hence, subject interest and self-concept are potential mechanisms of transmitting teacher specialization on boys' language skills.

Table 4.9 examines the effect of teacher specialization by teaching, which is defined as teaching only subjects from one field, MINT or languages. Specialization by teaching contrasts the previous specialization by studying. The difference between the two teacher specializations arises from the point in time where it is defined: in-service or during studies. Specialization by teaching relates closer to the existing literature, which defines teacher specialization on one subject based on the highest value-added of in-service personnel. We regress student achievement on teacher specialization by teaching and condition on teacher specialization by studying. We expect larger coefficients of

specialization by studying as it had more time to manifest. Estimation results suggest significant estimates of gender-specific teacher specialization by teaching, which turn insignificant conditional on teacher specialization by studying. As a result, teacher specialization by studying appears dominant relative to teacher specialization by teaching and underlines the importance of investigating initial teacher specialization compared to reshuffling in-service personnel.

We explore additional mechanisms, such as teacher-student gender match and further teacher characteristics. First, we test them as outcome variables to teacher specialization. Second, we employ them as (well-identified) controls to the main specification. For the first step, Table A 4.8 reports the estimates of regressing teacher gender, teacher-student gender match, and further teaching characteristics on teacher specialization. Most coefficients are insignificant except for two. In languages, there seem to be significantly more women among specialized teachers than among non-specialized teachers. In sciences, specialized teachers seem to be significantly more satisfied with their job than non-specialized teachers. Yet, it is encouraging that the gender match coefficient is insignificant, suggesting that there is no selection of students to teachers (and vice versa) based on gender. Yet, conditioning on job satisfaction does not alter the estimate of teacher specialization on student achievement, as Table A 4.3 reports. The results indicate that specialized teachers are not systematically different from non-specialized teachers except for higher job satisfaction in sciences and more women in languages. Hence, the identifying assumption of random assignment of students to teachers seems to hold.

In the second step, we employ teacher-student gender match and further teacher characteristics as controls to the main specification. In Table A 4.9 we show that the effect of teacher specialization on student achievement is robust to including teacher-student gender match. We find a main effect of gender match of 0.033 SD in sciences for girls and boys. This is consistent with literature suggesting positive effects on student achievement from being taught by a teacher of the same gender as the student (Dee 2005; 2007; Clotfelter, Ladd & Vigdor 2010). While the interaction of teacher specialization and gender match is always sizable, it is never significant. This tentatively suggests that the effect of teacher specialization may depend on matching genders between teacher and student but we cannot draw definite conclusions.

Conditional on other teaching characteristics, such as teacher-reported self-efficacy and job satisfaction, perceived discipline in the class, teaching style, internal differentiation, and self-assessed teaching skills, Table A 4.3 reports no change in the effect of teacher specialization on student achievement.[138]

In summary, one potential mechanism of transmitting the effect of teacher specialization on student achievement are boys' subject interest and self-concept in languages. Teacher-student gender match and further teaching characteristics seem to be no relevant mechanisms.

4.4.4 Robustness checks

We show that our results are robust to creating our own indices of self-concept and subject interest as well as to using sub-skills in each field. Furthermore, we report an alternative outcome instead of test scores.

The survey provides indices of self-concept and subject interest. To reduce concerns on results being driven by the way the pre-defined indices are constructed, we create our own indices from the sub-concepts following Kling, Liebman & Katz (2007). In 2015, the pre-defined indices also included imputed values. For our own indices, we do not impute values to demonstrate that the effects on student attitudes are not driven by the imputed observations. reports that coefficients and standard errors in sciences are very similar in size and significance to the survey indices. In languages, coefficient signs go in the same direction as for the survey indices.

As a second robustness check, we regress use the sub-skills surveyed in each subject instead of the global, i.e., average, score from those sub-skills in preceding analyses. In sciences, insight and knowledge were tested. In languages, reading and listening were tested. In Table A 4.5, we regress the sub-skills on teacher specialization. The coefficients of sub-skills are very similar to the average score. Hence, no sub-skill seems to dominate the global effect.

[138] We have aggregated items following Kling, Liebman & Katz (2007). For the single items and their categorization see Table A 4.6. Results are robust to including items separately and to alternative aggregation methods, such as principal component analysis (PCA) or factor analysis.

We consider grades as alternative outcome to student achievement rather than test scores.[139] Table 4.10 reports estimates from regressing teacher specialization on grades, which yields insignificant estimates of the main effect. Yet, results suggest that girls receive lower grades from specialized teachers across both fields, even conditional on subject-specific skills. This finding is in line with the literature hinting at teachers perceiving girls as less competent (Tiedemann 2000).

In Table A 4.7, we investigate the interplay of teacher specialization, student attitude and achievement, and grades. Panel A reports the effects of teacher specialization on student attitudes when conditioning and not conditioning on grades. The estimates of grades on attitudes are positive and highly significant in both fields (0.415 to 0.560 SD).[140] Panel B shows the link between grades and attitudes on achievement. In sciences, there seems to be no statistically significant correlation of grades and achievement (0.008 SD). In languages, grades yield a significant and positive estimate on achievement (0.173 SD). This may be due to the survey testing different science skills than relevant for grading in school. Yet, the result is surprising, as science skills are expected to be evaluated more objectively than language skills. Conditional on attitude, the coefficient of grades remains insignificant in sciences and significant in languages. Self-concept and subject interest yield a positive and significant estimate on achievement.

4.4.5 Discussion

Below, we discuss two remaining threats to identification: Non-random selection to specialization and subject combinations of non-specialized teachers. First, the selection of teachers to their specialization may not be random. We have included teacher characteristics in Table 4.2 and have shown that the treatment coefficient remains unaffected. The standard teacher characteristics, initial teacher training, and professional development are most likely exogenous to the treatment. Therefore, we apply those controls to all models. Teachers' attitudes towards assessments may be endogenous because specialized teachers may be in favor of assessments due to their closer adherence to the curriculum, which is tested in assessments. As they slightly

[139] We are skeptical against grades as a skill measure because grades are composed of performance and social relations between teachers and students. Furthermore, the timing of grades and test scores is lagged by about 3 months. Hence, it is unclear how to interpret grades.

[140] We cannot exclude reverse causality, i.e., that more interested students receive better grades.

influence the coefficient of teacher specialization in Table 4.2, we do not include those controls in our models. We cannot capture all potential omitted variables with the available items but we are confident to have covered the most relevant teacher characteristics. Additionally, it is reassuring that we find a positive effect of teacher specialization for boys but not for girls. This indicates that specialized teachers are not different in their characteristics than non-specialized teachers but that students absorb knowledge differently across gender.

Furthermore, we have tested whether certain teacher characteristics are affected by the treatment, such as teacher gender and self-efficacy, job satisfaction, classroom discipline, teaching style, internal differentiation, and self-assessed teaching skills (compare Table A 4.8). Two of the characteristics showed a statistically significant estimate. First, specialized teachers in sciences tend to be more satisfied with their job. Conditional on job satisfaction, the coefficient of teacher specialization changes only marginally. Hence, job satisfaction and other teacher characteristics including self-efficacy, classroom discipline, teaching style, internal differentiation, and self-assessed teaching skills are no mechanism of transmitting teacher specialization to student achievement (see Table A 4.9). Second, more female teachers seem to specialize in languages. This is not surprising, as more women teach languages than sciences (39 percent of physics teachers are female, 64 percent of chemistry teachers, and 73 percent of biology teachers opposed to 75 percent of German teachers and 77 of English teachers). Moreover, the teacher-student gender match seems statistically unrelated to teacher specialization, which supports that there is no selection of girls to their female teachers. Another check whether the effect of teacher specialization is driven by fewer female teachers in certain subjects, e.g., physics and chemistry, is to compare two instead of three subjects. Table A 4.10 shows estimation results for identifying student-fixed effects across physics versus biology, biology versus chemistry, and physics versus chemistry. We find the coefficients to be very similar when using physics and biology as well as chemistry and biology. The similar share of female teachers in both subjects indicates that different shares of female teachers do not drive our effects. However, we find no significant effect of teacher specialization when using physics and chemistry. This may originate from the high share of specialized teachers in physics and chemistry (81 percent in physics and 79 percent in chemistry opposed to 43 percent in biology). In summary, the advantage of teacher specialization on boys' achievement is not driven by teacher gender.

The second concern refers to the subjects non-specialized teachers choose. Potentially, sciences and languages are more academically demanding fields than other subjects, such as religious studies or arts. During their studies teachers may therefore choose subjects based on unobserved characteristics. For example, they may combine math with sports because they value a work-life balance and expect this combination to be less time consuming than math and German. Other unobserved characteristics determining subject combinations may be ability, motivation, or effort. Hence, the effect of teacher specialization on student achievement may depend on the subject combinations of non-specialized teachers. We explore this by dividing the non-specialized teachers into two groups. The first group are non-specialized teachers that studied languages (in 2012) or MINT subjects (in 2015). The second group are non-specialized teachers that studied music, arts, history, religion, politics, or sports. The latter may be regarded as easier subjects. We estimate the effect of teacher specialization compared to the two new control groups of non-specialized teachers in separate samples. Table A 4.11 reports additional subjects studied by specialized and non-specialized teachers. For example, 28.2 percent of chemistry teachers and 44.2 percent of biology teachers choose sports as one of their additional subjects. In languages, 31.4 percent of German teachers and 22.3 percent of English teachers choose history as a second subject.

Table 4.11 reports estimation results of these two new comparison groups, which result in smaller samples than the original mixed comparison group. In sciences, teacher specialization has no effect on boys compared to non-specialized teachers who studied languages. The interaction of student gender and teacher specialization remains significant and negative. For non-specialized teachers who studied easy subjects, we find a significant positive estimate of teacher specialization on boys' skills. The estimate for girls is insignificant and close to zero. In languages, the coefficient of teacher specialization is larger when comparing to non-specialized teachers who studied MINT. The gender interaction is not significant but around the same magnitude as in the main results. For non-specialized teachers who studied easy subjects, the estimate of teacher specialization is smaller.

We can only speculate on the different patterns in sciences and languages as there are no proxies for teacher ability or motivation. In sciences, the comparison with easy subjects may lead to a larger estimate for teacher specialization because many of the non-specialized teachers chose sports. This may have two consequences. First, sports yields

probably few complementarities in teaching with other subjects, such as chemistry. Second, it may signal a lack of effort because teachers expect lower workload from teaching sports. In languages, the estimate of easy subjects may be smaller because subjects as history may produce complementarities with German. Unfortunately, we do not observe teacher ability and motivation, which may drive the decision for the chosen subjects but also the effectiveness of specialization.

In summary, we have eliminated identification concerns regarding the selection of teachers into specialization based on (omitted) variables and we have discussed the effect of the subject combinations of non-specialized teachers. If teachers do not select into specialization based on unobserved characteristics in their studies (such as ability, motivation, content knowledge, effort, family situation, life goals, or preferences for teaching methods) and since all teacher candidates receive the same pedagogical training, our results should provide estimates that are close to the true effect of teacher specialization on student achievement.

4.5 Conclusion

We investigated the effect of teacher specialization on student achievement. We define specialization in several subjects of one field, MINT or languages, as chosen in studies opposed to the literature's definition of specialization on the one subject with the highest value-added in-service. Using the German National Assessment Studies of 2012 and 2015, we apply a within-student across-subjects fixed effects approach. Our findings suggest an advantage in science and language skills from teacher specialization for boys but not for girls. The results are consistent across school tracks, student achievement, and socio-economic status. We propose subject interest and self-concept in languages of boys as potential well-identified mechanisms, while teacher-student gender match or further teaching characteristics do not seem to act as mechanisms. We show that our results are robust to different ways of aggregating student attitude, i.e., subject interest and self-concept. Results appear similar across subject sub-skills, i.e., insight or knowledge in sciences and reading or listening in languages. Using grades as alternative skill measure yields two striking findings. First, girls are systematically graded lower by specialized teachers in sciences conditional on actual skills, which may originate from a mutually

reinforcing self-fulfilling prophecy of underestimation by teachers and by girls themselves. Second, grades and skills appear statistically unrelated in sciences, which may hint at different skills relevant for grading as opposed to the survey's test. Lastly, we discussed remaining potential threats to identification – whether the second subject of non-specialized teachers drives our results. We find the estimate of teacher specialization to be most pronounced when comparing specialized teachers in MINT to non-specialized teachers in easy subjects, and specialized teachers in languages to non-specialized teachers who studied MINT subjects.

Table 4.1: The link between teacher specialization to student achievement using OLS

Dep. var.	Science skills					Language skills				
Teacher specialization	0.056	0.016	0.104***	0.036*	0.040*	0.007	0.040	0.039	0.072**	0.079***
	(0.035)	(0.016)	(0.039)	(0.022)	(0.022)	(0.029)	(0.025)	(0.034)	(0.029)	(0.028)
Female student x teacher specialization			-0.093***	-0.039	-0.040			-0.065**	-0.065**	-0.067***
			(0.032)	(0.028)	(0.028)			(0.028)	(0.025)	(0.025)
Net effect for girls			0.011	-0.003	0.000			-0.025	0.007	0.012
			(0.039)	(0.021)	(0.021)			(0.031)	(0.027)	(0.026)
Controls										
Student, family, teacher	x	x	x	x	x	x	x	x	x	x
Teacher attitude towards assessments	-	-	-	-	x	-	-	-	-	x
School	x	-	x	-	-	x	-	x	-	-
Class fixed effects	-	x	-	x	x	-	x	-	x	x
Observations	17,546	17,546	17,546	17,546	17,546	38,446	38,446	38,446	38,446	38,446
N students	7,641	7,641	7,641	7,641	7,641	19,223	19,223	19,223	19,223	19,223
Clusters	640	640	640	640	640	868	868	868	868	868

Note: The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. We condition on the number of subjects studied and whether the teacher has studied the subject she or he teaches. Fixed effects on class level, in 2012; and at school level (one class per school), in 2015. When indicated by x, the model controls the following factors: at student level for age, born abroad, German not mother tongue, ever repeated a class; at family level for number of books at home in six categories, mother or father born abroad, highest educational degree in family, the highest ISEI; on the teacher level for gender, age, born abroad, no standard teacher training, years working at the school, years of experience as teacher, the employment volume in hours per week, the employment contract, the type of teaching license, the type of institution of teacher training, the total hours of professional development, the attitudes towards general student assessments and attitude towards VERA; at school level for private operation, city size, the number of students at a school, the share of German mother tongue speakers; and at the class level for class size. Additionally, we use imputation dummies. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 4.2: The effect of teacher specialization on student achievement using student-fixed effects

Dep. var.	Science skills						Language skills					
Teacher specialization	0.009	0.055***	0.056***	0.061***	0.061***	0.066***	0.035*	0.082***	0.082***	0.084***	0.087***	0.095***
	(0.015)	(0.018)	(0.018)	(0.020)	(0.019)	(0.019)	(0.021)	(0.027)	(0.029)	(0.031)	(0.031)	(0.029)
Female student x teacher specialization		-0.090***	-0.089***	-0.090***	-0.089***	-0.091***		-0.095***	-0.094***	-0.094***	-0.094***	-0.100***
		(0.022)	(0.021)	(0.021)	(0.021)	(0.020)		(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Net effect for girls		-0.035*	-0.033*	-0.028	-0.028	-0.025		-0.013	-0.012	-0.010	-0.008	-0.005
		(0.019)	(0.018)	(0.019)	(0.019)	(0.019)		(0.021)	(0.023)	(0.025)	(0.025)	(0.025)
Controls												
Standard teacher controls	-	-	x	x	x	x	-	-	x	x	x	x
Initial teacher training	-	-	-	x	x	x	-	-	-	x	x	x
Professional development of teachers	-	-	-	-	x	x	-	-	-	-	x	x
Teacher attitude towards assessments	-	-	-	-	-	x	-	-	-	-	-	x
Student-fixed effects	x	x	x	x	x	x	x	x	x	x	x	x
Observations	17,546	17,546	17,546	17,546	17,546	17,546	38,446	38,446	38,446	38,446	38,446	38,446
N students	7,641	7,641	7,641	7,641	7,641	7,641	19,223	19,223	19,223	19,223	19,223	19,223
Clusters	640	640	640	640	640	640	868	868	868	868	868	868

Note: The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. We condition on the number of subjects studied and whether the teacher has studied the subject she or he teaches. When indicated by x, the model controls the following factors: at the teacher level for standard controls (gender, age, born abroad, no standard teacher training, years working at the school, years of experience as teacher, the employment volume in hours per week, the employment contract), initial teacher training (the type of teaching license, the type of institution of teacher training), the total hours of professional development, and teacher attitude towards assessments (towards general student assessments and towards VERA). Additionally, we use imputation dummies. Fixed effects apply at student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 4.3: Heterogeneity by school track: the effect of teacher specialization on student achievement

Dep. var.	Science skills			Language skills		
Sample	All tracks	Lower tracks	Highest track	All tracks	Lower tracks	Highest track
Teacher specialization	0.061***	0.071***	0.051*	0.087***	0.091**	0.061
	(0.019)	(0.018)	(0.028)	(0.031)	(0.045)	(0.042)
Female student x teacher specialization	-0.089***	-0.075***	-0.096***	-0.094***	-0.083*	-0.100***
	(0.021)	(0.025)	(0.027)	(0.027)	(0.046)	(0.032)
Net effect for girls	-0.028	-0.004	-0.045**	-0.008	0.008	-0.040
	(0.019)	(0.021)	(0.023)	(0.025)	(0.044)	(0.033)
Student-fixed effects	x	x	x	x	x	x
Observations	17,546	8,040	9,506	38,446	17,404	21,042
N students	7,641	3,979	3,979	19,223	8,702	10,521
Clusters	640	335	305	868	423	445

Note: We divide the sample by school track attended by students (either the highest track – Gymnasium – or all other tracks). The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. We condition on the number of subjects studied and whether the teacher has studied the subject she or he teaches. The model controls at the teacher level for gender, age, born abroad, no standard teacher training, years working at the school, years of experience as teacher, the employment volume in hours per week, the employment contract, the type of teaching license, the type of institution of teacher training, and the total hours of professional development. Additionally, we use imputation dummies. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

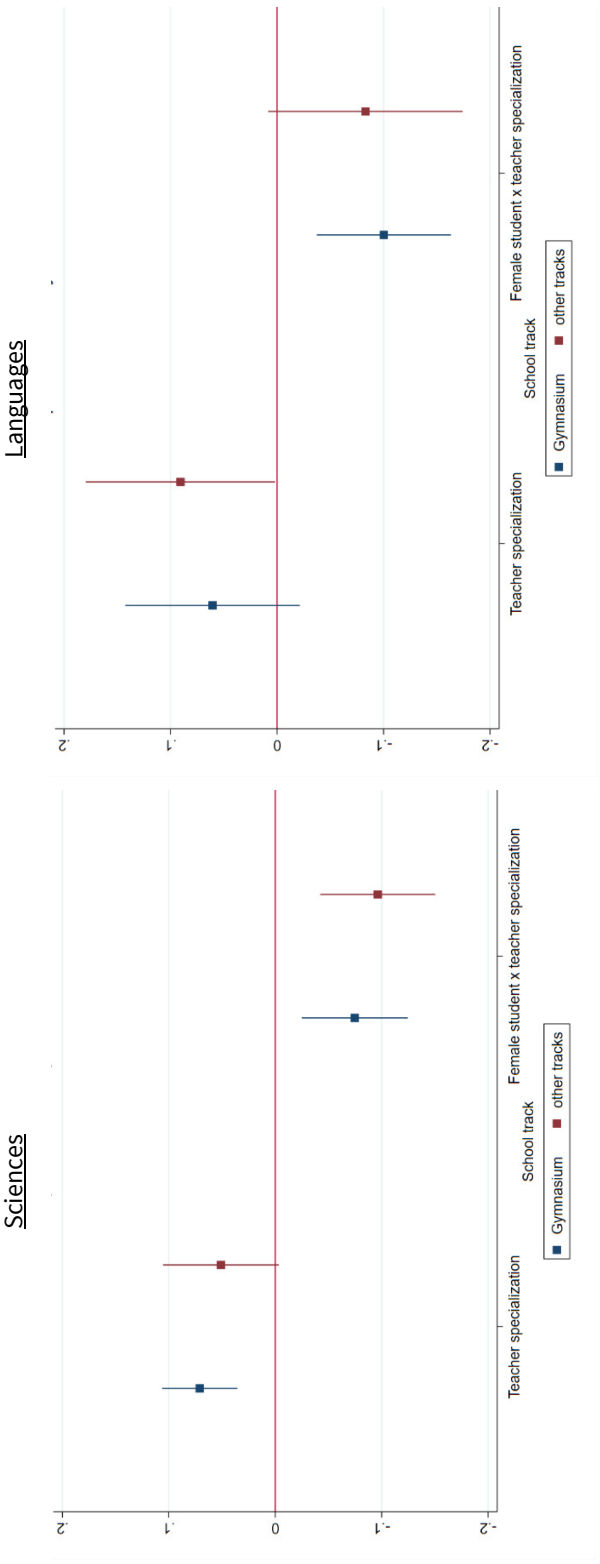


Figure 4.1: Heterogeneity by school tracks: the effect of teacher specialization on student achievement

Note: The coefficients depicted in this graph correspond to those presented in Table 4.3.

Table 4.4: Heterogeneity by achievement quartile: the effect of teacher specialization on student achievement

Dep. var.	Science skills					Language skills				
Sample	All	Lowest quartile	2nd quartile	3rd quartile	Highest quartile	All	Lowest quartile	2nd quartile	3rd quartile	Highest quartile
Teacher specialization	0.061***	0.043*	0.103***	0.042	0.044	0.087***	0.005	0.121***	0.126**	0.071
	(0.019)	(0.023)	(0.024)	(0.026)	(0.045)	(0.031)	(0.057)	(0.042)	(0.050)	(0.044)
Female student x teacher specialization	-0.089***	-0.014	-0.143***	-0.071*	-0.110***	-0.094***	-0.042	-0.063	-0.088*	-0.161***
	(0.021)	(0.035)	(0.037)	(0.038)	(0.0396)	(0.027)	(0.064)	(0.052)	(0.053)	(0.048)
Net effect for girls	-0.028	0.030	-0.040	-0.029	-0.066**	-0.008	-0.037	0.058	0.037	-0.090**
	(0.019)	(0.028)	(0.027)	(0.033)	(0.031)	(0.025)	(0.054)	(0.046)	(0.041)	(0.037)
Student-fixed effects	x	x	x	x	x	x	x	x	x	x
Observations	17,546	4,388	4386	4386	4,386	38,446	9,612	9,612	9,612	9,610
N students	7,641	1,970	1931	1892	1,848	19,223	4,806	4,806	4,806	4,805
Clusters	640	405	543	519	393	868	625	805	758	605

Note: We divide the sample by average performance quartile. The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

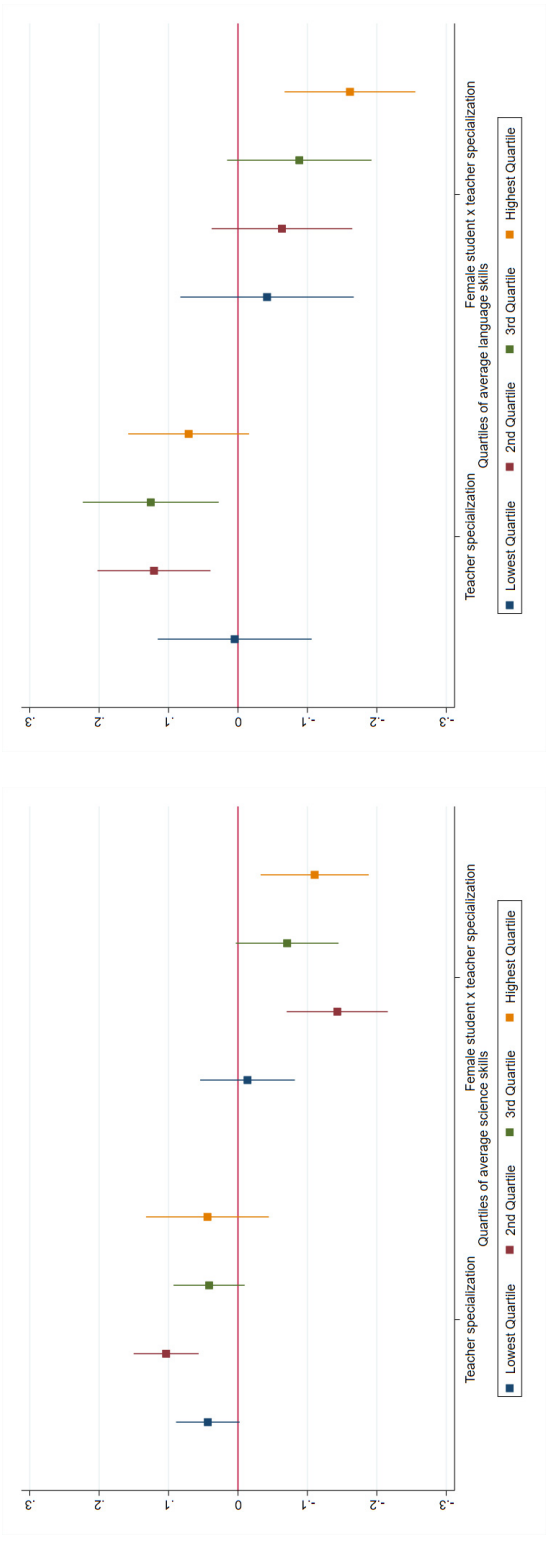


Figure 4.2: Heterogeneity by achievement quartile: the effect of teacher specialization on student achievement

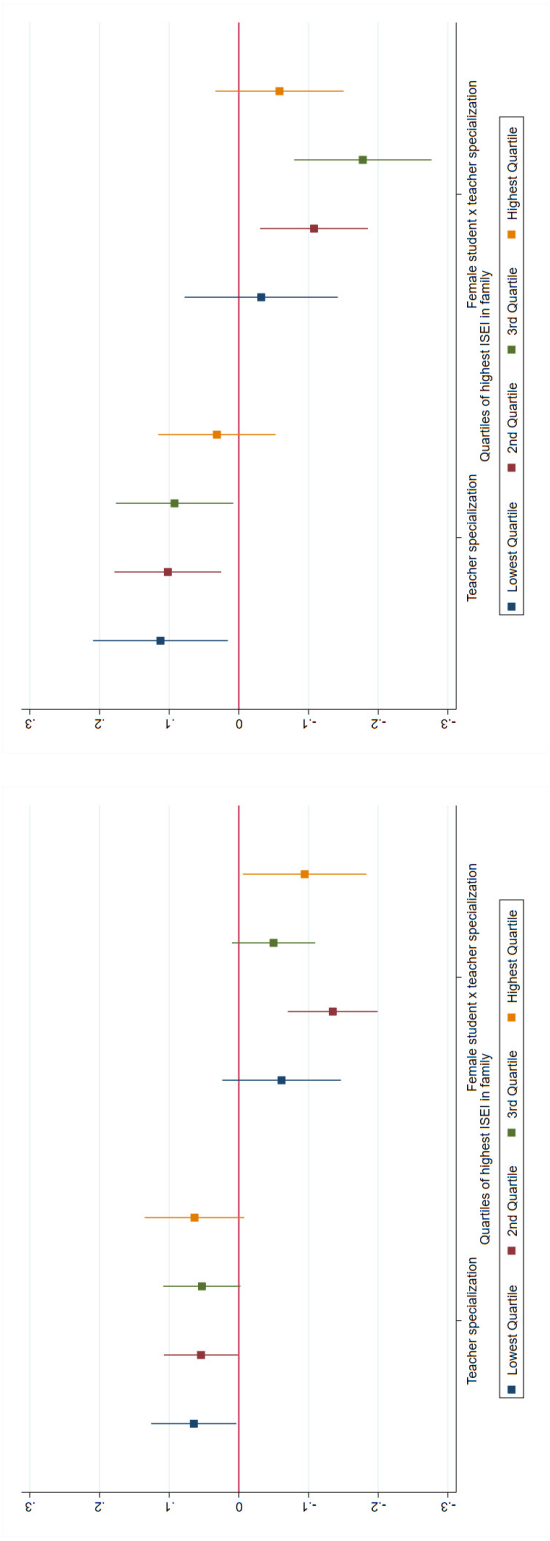
Note: The coefficients depicted in this graph correspond to those presented in Table 4.4.

Table 4.5: Heterogeneity by socio-economic status: the effect of teacher specialization on student achievement

Dep. var.	Science skills					Language skills				
Sample	All	Lowest quartile	2nd quartile	3rd quartile	Highest quartile	All	Lowest quartile	2nd quartile	3rd quartile	Highest quartile
Teacher specialization	0.061***	0.065**	0.054**	0.053*	0.064*	0.087***	0.113**	0.102***	0.092**	0.032
	(0.019)	(0.031)	(0.027)	(0.028)	(0.036)	(0.031)	(0.049)	(0.039)	(0.043)	(0.043)
Female student x teacher specialization	-0.089***	-0.061	-0.135***	-0.050	-0.095**	-0.094***	-0.032	-0.108***	-0.178***	-0.058
	(0.021)	(0.043)	(0.033)	(0.031)	(0.045)	(0.027)	(0.056)	(0.039)	(0.050)	(0.047)
Net effect for girls	-0.028	0.003	-0.080***	0.003	-0.031	-0.008	0.080*	-0.006	-0.086**	-0.027
	(0.019)	(0.033)	(0.027)	(0.026)	(0.034)	(0.025)	(0.045)	(0.039)	(0.039)	(0.043)
Student-fixed effects	x	x	x	x	x	x	x	x	x	x
Observations	17,546	4,405	4,381	4,414	4,346	38,446	9,912	11,568	7,448	9,518
N students	7,641	1,947	1,910	1,917	1,867	19,223	4,956	5,784	3,724	4,759
Clusters	640	550	599	590	480	868	824	861	813	775

Note: We divide the sample into quartiles based on socio-economic status, which is a survey item. The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Figure 4.3: Heterogeneity by socio-economic status: the effect of teacher specialization on student achievement



Note: The coefficients depicted in this graph correspond to those presented in Table 4.5.

Table 4.6: The effect of teacher specialization on student attitude

Sample	Sciences				Languages			
Dep. var.	Self-concept		Subject interest		Self-concept		Subject interest	
Teacher specialization	-0.043	0.022	0.007	0.198**	0.031	0.079*	0.061	0.098*
	(0.047)	(0.058)	(0.058)	(0.094)	(0.037)	(0.047)	(0.047)	(0.058)
Female student x teacher specialization		-0.132		-0.371***		-0.097*		-0.075
		(0.084)		(0.135)		(0.052)		(0.050)
Net effect for girls		-0.110		-0.173**		-0.018		0.023
		(0.068)		(0.085)		(0.043)		(0.047)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	8,451	8,451	4,889	4,889	38,446	38,446	38,446	38,446
N students	3,704	3,704	2,113	2,113	19,223	19,223	19,223	19,223
Clusters	352	352	212	212	868	868	868	868

Note: The dependent variables are students' self-concept and subject interest standardized to mean zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 4.7: The effect of teacher specialization on student attitude's sub-concepts

Panel A: Self-concept

Sample	Sciences				Languages						
Dep. var.	Good	Good grades	Learn fast	Best subject	Mostly understand	Learn fast	Effort does help, I can cope	Talent	No hopeless case	Less difficulties than others	Need less time than others
Teacher specialization	0.042	-0.007	-0.066	0.105*	0.075	0.120**	0.128**	0.083	0.034	0.139***	0.037
	(0.056)	(0.060)	(0.054)	(0.056)	(0.060)	(0.053)	(0.061)	(0.061)	(0.055)	(0.050)	(0.040)
Female student x teacher specialization	-0.131	-0.088	-0.005	-0.222***	-0.028	-0.098	-0.115	-0.094	-0.085	-0.145**	-0.077
	(0.092)	(0.070)	(0.076)	(0.084)	(0.060)	(0.069)	(0.073)	(0.074)	(0.077)	(0.065)	(0.047)
Net effect for girls	-0.089	-0.095	-0.070	-0.117*	0.048	0.021	0.014	-0.012	-0.051	-0.006	-0.040
	(0.067)	(0.069)	(0.060)	(0.061)	(0.042)	(0.048)	(0.053)	(0.052)	(0.052)	(0.047)	(0.037)
Student-fixed effects	x	x	x	x	x	x	x	x	x	x	x
Observations	8,402	8,374	8,381	8,373	24,422	24,280	24,326	24,316	24,342	24,372	24,337
N students	3,696	3,689	3,691	3,687	12,271	12,271	12,239	12,260	12,254	12,254	12,257
Clusters	352	352	352	352	617	0.047	617	617	617	617	617

Note: The dependent variables are sub-concepts of students' self-concept standardized to mean zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Panel B: Subject interest

Sample	Sciences				Languages			
Dep. var.	Personally important	Enjoy	Interest	Favorite activity	Personally important	Enjoy	Interest	Favorite activity
Teacher specialization	0.179**	0.125	0.187**	0.224***	0.121	0.181**	0.162**	0.125**
	(0.084)	(0.100)	(0.083)	(0.085)	(0.074)	(0.073)	(0.066)	(0.057)
Female student x teacher specialization	-0.311**	-0.291**	-0.350**	-0.383***	-0.008	-0.083	-0.044	-0.007
	(0.121)	(0.134)	(0.136)	(0.116)	(0.070)	(0.069)	(0.063)	(0.048)
Net effect for girls	-0.132	-0.166*	-0.163*	-0.159**	0.113**	0.098*	0.117**	0.118**
	(0.082)	(0.087)	(0.089)	(0.065)	(0.052)	(0.056)	(0.049)	(0.046)
Student-fixed effects	X	x	x	X	x	x	x	x
Observations	4,865	4,870	4,864	4,853	24,426	24,379	24,309	24,337
N students	2,110	2,109	2,107	2,104	12,248	12,244	12,241	12,233
Clusters	212	212	212	212	616	616	616	616

Note: The dependent variables are sub-concepts of students' subject interest standardized to mean zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 4.8: The effect of teacher specialization on student achievement conditional on student attitude by student gender

Panel A: Girls

Dep. var.	Science skills				Language skills			
Teacher specialization	0.004	0.007	0.010	0.010	-0.018	-0.013	-0.018	-0.015
	(0.024)	(0.025)	(0.025)	(0.025)	(0.027)	(0.026)	(0.027)	(0.026)
Self-concept		0.036**		0.017		0.158***		0.127***
		(0.017)		(0.022)		(0.010)		(0.010)
Subject interest			0.037***	0.026			0.117***	0.067***
			(0.014)	(0.017)			(0.008)	(0.009)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	2,476	2,476	2,476	2,476	19,498	19,498	19,498	19,498
N students	1,069	1,069	1,069	1,069	9,749	9,749	9,749	9,749
Clusters	202	202	202	202	860	860	860	860

Panel B: Boys

Dep. var.	Science skills				Language skills			
Teacher specialization	0.111***	0.110***	0.106***	0.108***	0.092***	0.078***	0.077***	0.072**
	(0.027)	(0.027)	(0.027)	(0.027)	(0.032)	(0.030)	(0.029)	(0.029)
Self-concept		0.030**		0.023		0.165***		0.134***
		(0.013)		(0.016)		(0.009)		(0.010)
Subject interest			0.026**	0.009			0.129***	0.076***
			(0.012)	(0.015)			(0.010)	(0.010)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	2,402	2,402	2,402	2,402	18,948	18,948	18,948	18,948
N students	1,040	1,040	1,040	1,040	9,474	9,474	9,474	9,474
Clusters	195	195	195	195	856	856	856	856

Note: We divide the sample by student gender. The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. We control for students' self-concept and subject interest standardized to mean zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 4.9: The effect of teacher specialization by teaching on student achievement

Dep. var.	Science skills				Language skills			
Teacher specialization by studying			0.072***	0.062***			0.055	0.078*
			(0.020)	(0.020)			(0.038)	(0.045)
Female student x teacher specialization			-0.088***	-0.067***			-0.095***	-0.142***
			(0.021)	(0.025)			(0.027)	(0.042)
Teacher specialization by teaching	-0.001	0.038**	-0.016	0.002	0.046**	0.071***	0.042	0.012
	(0.015)	(0.018)	(0.016)	(0.019)	(0.021)	(0.027)	(0.027)	(0.039)
Female student x teacher specialization by teaching		-0.075***		-0.037		-0.049*		0.059
		(0.022)		(0.026)		(0.028)		(0.042)
Net effect of specialization by teaching for girls		-0.037*		-0.036*		0.022		0.071*
		(0.019)		(0.021)		(0.023)		(0.029)
Net effect of specialization by studying for girls			-0.016	-0.005			-0.041	-0.064**
			(0.018)	(0.020)			(0.031)	(0.031)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	17,546	17,546	17,546	17,546	38,446	38,446	38,446	38,446
N students	7,641	7,641	7,641	7,641	19,223	19,223	19,223	19,223
Clusters	640	640	640	640	868	868	868	868

Note: The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 4.10: Alternative outcome: the effect of teacher specialization on grades

Dep. var.	Science grade		Language grade	
Teacher specialization	0.028	0.026	0.050	0.028
	(0.047)	(0.047)	(0.041)	(0.040)
Female student x teacher specialization	-0.137***	-0.134***	-0.105***	-0.080**
	(0.048)	(0.048)	(0.039)	(0.038)
Subject-specific skills		0.037		0.258***
		(0.037)		(0.014)
Net effect for female	-0.109**	-0.108**	-0.055	-0.052
	(0.046)	(0.046)	(0.039)	(0.040)
Student-fixed effects	x	x	x	x
Observations	17,160	17,160	36,302	36,302
N students	7,505	7,505	18,170	18,170
Clusters	631	631	819	819

Note: The dependent variables are grades from the student questionnaire concerning the past semester with higher values representing higher achievement. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 4.11: Change of control group: the effect of teacher specialization on student achievement by easy or difficult subject

Dep. var.	Science skills				Language skills			
Teacher specialization vs. MINT or languages	-0.022	0.021			0.129**	0.168**		
	(0.035)	(0.043)			(0.059)	(0.076)		
Female student x teacher specialization vs. MINT or languages		-0.088**				-0.078		
		(0.035)				(0.076)		
Teacher specialization vs. other subjects			0.024*	0.071***			0.049*	0.098***
			(0.014)	(0.018)			(0.028)	(0.035)
Female student x teacher specialization vs. other subjects				-0.087***				-0.099***
				(0.026)				(0.034)
Net effect for girls		-0.067*		-0.017		0.090		-0.001
		(0.037)		(0.021)		(0.063)		(0.031)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	13,182	13,182	14,904	14,904	17,942	17,942	28,279	28,279
N students	7,213	7,213	7,459	7,459	13,710	13,710	17,486	17,486
Clusters	608	608	625	625	786	786	616	616

Note: We define teacher specialization by having studied only subjects from sciences (physics, chemistry, and biology) or from languages (German and English) and divide the control group of non-specialized teachers into two new groups based on the subjects studied: (i) whether the subjects belonged to MINT (math, informatics, physics, chemistry, biology, and geography) or to languages (German, English, French or Latin) (ii) whether they belonged to neither to MINT nor languages (music, arts, history, politics, religion, sports, and other subjects). The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Appendix

Table A 4.1: Evolution of PISA scores by gender

Subject	Math			
Year	2000		2015	
Gender	girls	boys	girls	boys
All PISA countries	493.4	505.0	462.7	470.9
International gender gap in math (girls-boys)	-11.6		-8.2	
Germany	493.5	508.2	498.6	517.1
German gender gap in math (girls-boys)	-14.7		-17.5	
Subject	Reading			
Year	2000		2015	
Gender	girls	boys	girls	boys
All PISA countries	515.5	483.9	483.9	457.9
International gender gap in reading (girls-	31.6		26	
Germany	514.0	482.3	521.1	503.9
German gender gap in reading (girls-boys)	31.7		17.2	

Note: The table reports achievement by gender and by year in PISA points. The sample of all PISA countries contains 30 countries in 2000 (Austria, Australia, Belgium, Brazil, Canada, the Czech Republic, Germany, Denmark, Span, Finland, France, the United Kingdom, Greece, Hungary, Ireland, Iceland, Italy, Japan, Korea, Luxemburg, Latvia, Mexico, Norway, the Netherlands, Poland, Portugal, Russia, Sweden, and the United States) and 56 countries in 2015 (additionally the United Arab Emirates, Bulgaria, Chile, Colombia, Costa Rica, Estonia, Hong Kong, Croatia, Indonesia, Israel, Jordan, Lithuania, Macao, Montenegro, the Netherlands, Peru, Qatar, Romania, Singapore, the Slovak Republic, Slovenia, Thailand, Tunisia, Turkey, Taiwan, and Uruguay). Data Source: OECD (2000-2015).

Table A 4.2: Summary statistics

Panel A: Teacher characteristics

Wave	2012						2015			
Subject	Physics		Chemistry		Biology		German		English	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Teacher Diversity/Generalist										
Field specialist by studies	0.813	*0.391*	0.785	*0.411*	0.430	*0.496*	0.258	*0.438*	0.421	*0.494*
Field specialist by teaching	0.786	*0.411*	0.734	*0.443*	0.436	*0.496*	0.248	*0.432*	0.403	*0.491*
Number of subjects studied	1.970	*0.763*	1.967	*0.634*	1.951	*0.575*	1.798	*0.816*	1.710	*0.864*
Studied what teaches	0.867	*0.340*	0.924	*0.265*	0.924	*0.266*	0.856	*0.351*	0.840	*0.367*
Female	0.388	*0.488*	0.640	*0.481*	0.733	*0.443*	0.752	*0.432*	0.772	*0.420*
Share missing	0.042	*0.201*	0.0266	*0.161*	0.0233	*0.151*	0.024	*0.154*	0.035	*0.183*
Age	47.87	*10.23*	46.82	*10.92*	45.45	*11.13*	45.69	*11.06*	46.02	*11.13*
Share missing	0.057	*0.231*	0.037	*0.189*	0.032	*0.176*	0.037	*0.189*	0.043	*0.202*
Born abroad	0.046	*0.209*	0.025	*0.157*	0.022	*0.146*	0.017	*0.128*	0.035	*0.184*
Share missing	0.030	*0.170*	0.027	*0.161*	0.025	*0.158*	0.036	*0.186*	0.047	*0.212*
No standard teacher training	0.146	*0.353*	0.095	*0.294*	0.052	*0.221*	0.063	*0.243*	0.077	*0.267*
Share missing	0.037	*0.189*	0.031	*0.173*	0.015	*0.121*	0.032	*0.177*	0.030	*0.171*
Years at this school	11.31	*9.559*	11.82	*10.51*	10.79	*9.867*	10.42	*9.043*	11.18	*9.442*
Share missing	0.039	*0.195*	0.045	*0.208*	0.030	*0.170*	0.040	*0.197*	0.047	*0.212*
Years as teacher	19.42	*12.55*	19.10	*12.77*	17.62	*12.71*	17.68	*12.64*	17.96	*12.62*
Share missing	0.044	*0.206*	0.045	*0.208*	0.030	*0.170*	0.037	*0.189*	0.045	*0.207*
Employment volume (hours per week)	22.05	*5.121*	21.57	*5.419*	22.27	*5.037*	22.88	*4.590*	22.90	*4.843*
Share missing	0.022	*0.147*	0.025	*0.155*	0.0233	*0.151*	0.033	*0.180*	0.033	*0.180*
Employment contract										
Civil servant	0.648	*0.478*	0.648	*0.478*	0.618	*0.486*	0.634	*0.482*	0.649	*0.478*
Permanent contract	0.281	*0.450*	0.277	*0.448*	0.302	*0.459*	0.294	*0.456*	0.287	*0.453*
Temporary contract > 1 year	0.026	*0.158*	0.045	*0.207*	0.035	*0.183*	0.024	*0.153*	0.028	*0.164*
Temporary contract < 1 year	0.046	*0.210*	0.030	*0.170*	0.046	*0.209*	0.049	*0.216*	0.036	*0.187*
Share missing	0.035	*0.183*	0.040	*0.194*	0.023	*0.151*	*0.037*	*0.188*	0.045	*0.207*

(Continued on next page.)

Wave			2012				2015			
Subject	Physics		Chemistry		Biology		German		English	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Level of teacher qualification										
Primary	0.005	*0.070*	0.006	*0.078*	0.0107	*0.103*	0.013	*0.112*	0.008	*0.09*
Primary and secondary I	0.055	*0.228*	0.045	*0.208*	0.0792	*0.270*	0.099	*0.299*	0.085	*0.279*
Secondary I	0.358	*0.480*	0.381	*0.486*	0.385	*0.487*	0.357	*0.479*	0.320	*0.467*
Secondary II general or highest track	0.520	*0.500*	0.547	*0.498*	0.571	*0.496*	0.538	*0.499*	0.559	*0.497*
Secondary II vocational	0.000	*0.000*	0.004	*0.064*	0.000	*0.000*	0.001	*0.034*	0.002	*0.048*
Special education	0.000	*0.000*	0.002	*0.045*	0.002	*0.046*	0.002	*0.048*	0.000	*0.000*
Share missing	0.010	*0.100*	0.004	*0.064*	0.011	*0.103*	0.000	*0.000*	0.000	*0.000*
Type of teacher training institution										
Institute for teacher training	0.005	*0.0704*	0.008	*0.091*	0.006	*0.080*	0.013	*0.112*	0.016	*0.126*
Pedagogical university	0.259	*0.438*	0.307	*0.462*	0.229	*0.420*	0.230	*0.421*	0.183	*0.387*
University or applied university	0.607	*0.489*	0.603	*0.490*	0.728	*0.445*	0.682	*0.466*	0.718	*0.450*
Other university	0.000	*0.000*	0.006	*0.078*	0.004	*0.065*	0.009	*0.096*	0.007	*0.083*
Share missing	0.010	*0.100*	0.006	*0.078*	0.008	*0.012*	0.000	*0.000*	0.000	*0.000*
Participation in professional dev.	0.836	*0.371*	0.855	*0.352*	0.841	*0.366*	0.836	*0.371*	0.837	*0.370*
Share missing	0.099	*0.298*	0.0656	*0.248*	0.053	*0.224*	0.116	*0.321*	0.109	*0.312*
Hours of professional dev.	22.20	*44.22*	21.50	*40.43*	19.22	*22.88*	23.52	*39.59*	25.35	*47.19*
Share missing	0.202	*0.402*	0.209	*0.407*	0.191	*0.393*	0.146	*0.354*	0.121	*0.326*
Attitude towards general student ass.										
Should be done regularly	2.589	*0.872*	2.647	*0.788*	2.599	*0.838*	2.581	*0.818*	2.533	*0.834*
Important for school's work	2.406	*0.870*	2.482	*0.804*	2.433	*0.834*	2.405	*0.840*	2.365	*0.829*
Creates trouble	2.591	*0.902*	2.661	*0.824*	2.612	*0.854*	2.278	*0.844*	2.290	*0.850*
Creates problems	2.727	*0.856*	2.774	*0.785*	2.697	*0.822*	2.254	*0.789*	2.253	*0.808*
Schools make a greater effort	2.306	*0.824*	2.345	*0.741*	2.268	*0.780*	2.103	*0.749*	2.020	*0.751*
Basis to evaluate school objectively	2.367	*0.876*	2.451	*0.843*	2.421	*0.821*	2.377	*0.825*	2.382	*0.834*
Basis for discussion among colleagues	2.407	*0.853*	2.505	*0.825*	2.449	*0.817*	2.280	*0.815*	2.254	*0.833*
Share missing	0.017	*0.130*	0.037	*0.189*	0.017	*0.129*	0.035	*0.183*	0.038	*0.191*

(Continued on next page.)

Wave	2012						2015			
Subject	Physics		Chemistry		Biology		German		English	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Attitude towards VERA										
Should be done regularly	2.567	*0.870*	2.708	*0.812*	2.578	*0.821*	2.392	*0.856*	2.407	*0.882*
Important for school's work	2.434	*0.865*	2.576	*0.805*	2.460	*0.820*	2.267	*0.835*	2.232	*0.847*
Creates trouble	2.698	*0.862*	2.751	*0.815*	2.618	*0.839*	2.282	*0.863*	2.291	*0.885*
Creates problems	2.784	*0.809*	2.825	*0.752*	2.744	*0.834*	2.269	*0.847*	2.280	*0.858*
Schools make a greater effort	2.316	*0.777*	2.395	*0.748*	2.318	*0.773*	2.104	*0.743*	1.969	*0.754*
Basis to evaluate school objectively	2.457	*0.848*	2.587	*0.797*	2.484	*0.825*	2.311	*0.824*	2.315	*0.845*
Basis for discussion among colleagues	2.436	*0.814*	2.559	*0.786*	2.451	*0.851*	2.254	*0.817*	2.204	*0.847*
Share missing	0.032	*0.176*	0.066	*0.248*	0.040	*0.197*	0.043	*0.202*	0.044	*0.205*
Further teaching characteristics										
Self-efficacy	2.880	*0.376*	2.907	*0.383*	2.936	*0.422*	-	-	-	-
Job satisfaction	3.190	*0.534*	3.195	*0.557*	3.192	*0.552*	-	-	-	-
Discipline	-	-	-	-	-	-	1.717	*0.569*	1.725	*0.565*
Teaching style	-	-	-	-	-	-	2.604	*0.354*	2.628	*0.396*
Internal differentiation	-	-	-	-	-	-	2.756	*0.466*	2.806	*0.469*
Self-assessed teaching skills	-	-	-	-	-	-	3.065	*0.234*	3.076	*0.250*
N	406		488		472		868		868	

(Continued on next page.)

Panel B: Student characteristics

Wave	2012		2015	
	Mean	*SD*	Mean	*SD*
Student test score (not standardized)				
Biology	533.939	*87.155*	-	-
Chemistry	534.935	*88.063*	-	-
Physics	533.977	*87.192*	-	-
German	-	-	525.811	*86.935*
English	-	-	519.770	*91.145*
Student grade (not standardized, not rescaled)				
Biology	2.882	*0.960*	-	-
Share missing	0.062	*0.242*	-	-
Chemistry	2.943	*1.005*	-	-
Share missing	0.039	*0.195*	-	-
Physics	2.963	*0.992*	-	-
Share missing	0.039	*0.193*	-	-
German	-	-	2.882	*0.882*
Share missing	-	-	0.055	*0.229*
English	-	-	2.912	*0.934*
Share missing	-	-	0.056	*0.230*
Self-concept (not standardized)				
Biology	2.851	*0.657*	-	-
Share missing	0.521	*0.500*	-	-
Chemistry	2.530	*0.799*	-	-
Share missing	0.517	*0.500*	-	-
Physics	2.541	*0.761*	-	-
Share missing	0.516	*0.500*	-	-
German	-	-	3.288	*0.564*
Share missing	-	-		
English	-	-	3.208	*0.710*
Share missing	-	-		
Subject interest (not standardized)				
Biology	2.589	*0.742*	-	-
Share missing	0.725	*0.446*	-	-
Chemistry	2.263	*0.850*	-	-
Share missing	0.725	*0.447*	-	-
Physics	2.249	*0.832*	-	-
Share missing	0.724	*0.447*		
German	-	-	2.442	*0.678*
Share missing	-	-	0.000	*0.000*
English	-	-	2.801	*0.745*
Share missing	-	-	0.000	*0.000*
Female	0.507	0.500	0.507	*0.500*
Share missing	0.000	0.000	0.000	*0.000*
Age	15.52	0.640	15.43	*0.569*
Share missing	0.000	0.011	0.000	*0.000*
Born abroad	0.043	0.203	0.041	*0.199*
Share missing	0.093	0.290	0.050	*0.219*
No German mother tongue	0.092	0.289	0.126	*0.332*
Share missing	0.017	0.129	0.047	*0.211*
Grade repetition	0.147	0.354	0.143	*0.350*
Share missing	0.091	0.288	0.047	*0.211*

(Continued on next page.)

Wave	2012		2015	
	Mean	*SD*	Mean	*SD*
Number of books at home				
0-10	0.089	*0.285*	0.054	*0.227*
11-25	0.106	*0.308*	0.088	*0.283*
26-100	0.262	*0.440*	0.266	*0.442*
101-200	0.207	*0.405*	0.205	*0.404*
201-500	0.195	*0.396*	0.223	*0.417*
More than 500	0.141	*0.348*	0.164	*0.370*
Share missing	0.156	*0.363*	0.067	*0.249*
Mother born abroad	0.144	*0.351*	0.183	*0.387*
Share missing	0.157	*0.364*	0.065	*0.246*
Father born abroad	0.148	*0.355*	0.185	*0.388*
Share missing	0.179	*0.383*	0.070	*0.256*
Highest parental education				
ISCED 1	0.010	*0.098*	0.005	*0.071*
ISCED 2	0.010	*0.100*	0.016	*0.124*
ISCED 3A 3B 3C	0.103	*0.304*	0.065	*0.246*
ISCED 3A 4	0.132	*0.338*	0.262	*0.440*
ISCED 5B	0.246	*0.431*	0.181	*0.385*
ISCED 6	0.147	*0.354*	0.151	*0.358*
Share missing	0.232	*0.422*	0.000	*0.000*
Parental occupation				
Lower occupation	0.299	*0.458*	0.302	*0.459*
Routine services in trade and administration	0.306	*0.461*	0.308	*0.462*
Self-employed	0.202	*0.402*	0.201	*0.401*
Skilled worker and worker with personnel responsibility	0.062	*0.241*	0.055	*0.229*
Unskilled worker, agricultural worker	0.068	*0.252*	0.076	*0.265*
Share missing	0.000	*0.000*	0.000	*0.000*
Highest ISEI	52.33	*20.28*	52.92	*20.26*
Share missing	0.000	*0.000*	0.000	*0.000*
N	7,641		19,223	

(Continued on next page.)

Panel C: School characteristics

Wave	*2012*		*2015*	
	Mean	*SD*	Mean	*SD*
Private operation	0.064	*0.245*	0.064	*0.245*
Share missing	0.117	*0.322*	0.048	*0.215*
City size (inhabitants) of school location				
Large city (>5000,000)	0.150	*0.357*	0.166	*0.373*
City (100,000 - 1,000,00)	0.113	*0.317*	0.104	*0.306*
Large town (15,000-100,000)	0.074	*0.262*	0.085	*0.279*
Town (<3,000-15,000)	0.285	*0.452*	0.274	*0.446*
Small town (<3,000)	0.326	*0.469*	0.308	*0.462*
Share missing	0.130	*0.337*	0.051	*0.219*
School size	697.6	*334.5*	376.0	*112.7*
Share missing	0.123	*0.329*	0.053	*0.224*
Share of German speakers at school				
More than 90 percent	0.729	*0.445*	0.667	*0.472*
76 - 90 percent	0.151	*0.358*	0.167	*0.373*
51 - 75 percent	0.058	*0.234*	0.100	*0.300*
26 - 50 percent	0.041	*0.198*	0.040	*0.197*
25 percent or less	0.022	*0.145*	0.027	*0.162*
Share missing	0.121	*0.326*	0.054	*0.226*
N of schools	*529*		*868*	
Classroom characteristics				
Number of students in class	22.55	*4.603*	24.32	*4.490*
N of classes	640		868	

Table A 4.3: Teacher specialization on student achievement conditional on further teaching characteristics

Dep. var.	Science skills		Language skills			
Teacher specialization	0.060***	0.057***	0.085***	0.087***	0.085***	0.086***
	(0.019)	(0.019)	(0.031)	(0.031)	(0.031)	(0.031)
Female student x teacher specialization	-0.089***	-0.085***	-0.094***	-0.093***	-0.095***	-0.095***
	(0.022)	(0.021)	(0.027)	(0.027)	(0.027)	(0.027)
Self-efficacy	-0.010					
	(0.007)					
Job satisfaction		0.001				
		(0.007)				
Discipline			-0.017			
			(0.014)			
Teaching style				0.029		
				(0.022)		
Internal differentiation					-0.014	
					(0.024)	
Self-assessed teaching skills						0.015
						(0.020)
Net effect for girls	-0.029	-0.028	-0.010	-0.009	-0.008	-0.010
	(0.019)	(0.019)	(0.025)	(0.025)	(0.025)	(0.025)
Student-fixed effects	x	x	x	x	x	x
Observations	17,175	17,075	38,446	38,446	38,446	38,446
N students	7,624	7,608	19,223	19,223	19,223	19,223
Clusters	639	638	868	868	868	868

Note: The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. Further teaching characteristics are indices from sub-categories enlisted in Appendix Table A 4.6. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 4.4: The effect of teacher specialization on student attitude (own index)

Sample	Sciences				Languages			
Dep. var.	Self-concept		Subject interest		Self-concept		Subject interest	
Teacher specialization	-0.044	0.020	0.008	0.198**	0.052	0.109*	0.151***	0.168**
	(0.047)	(0.058)	(0.058)	(0.094)	(0.042)	(0.060)	(0.057)	(0.075)
Female student x teacher specialization		-0.130		-0.370***		-0.118		-0.036
		(0.084)		(0.135)		(0.072)		(0.068)
Net effect for girls		-0.110		-0.172**		-0.008		0.132**
		(0.068)		(0.085)		(0.052)		(0.055)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	8,451	8,451	4,889	4,889	24,521	24,521	24,492	24,492
N students	3,704	3,704	2,113	2,113	12,277	12,277	12,255	12,255
Clusters	352	352	212	212	617	617	616	616

Note: The dependent variables are students' self-concept and subject interest standardized to mean zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 4.5: The effect of teacher specialization on sub-skills

Sample	Sciences				Languages			
Dep. var.	Insight		Knowledge		Reading		Listening	
Teacher specialization	0.010	0.045**	0.021	0.072***	0.027	0.066**	0.049*	0.099***
	(0.017)	(0.023)	(0.021)	(0.023)	(0.025)	(0.032)	(0.029)	(0.034)
Female student x teacher specialization		-0.069**		-0.100***		-0.079***		-0.101***
		(0.029)		(0.020)		(0.029)		(0.028)
Net effect for girls		-0.024		-0.028		-0.013		-0.001
		(0.023)		(0.022)		(0.025)		(0.030)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	17,546	17,546	17,546	17,546	38,446	38,446	38,446	38,446
N students	7,641	7,641	7,641	7,641	19,223	19,223	19,223	19,223
Clusters	640	640	640	640	868	868	868	868

Note: The dependent variables are science sub-skills' insight and knowledge in 2012 and language sub-skills' reading and listening in 2015 standardized to mean of zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 4.6: Elements of further teaching characteristics

Panel A: Sciences
Self-efficacy
I can even teach the difficult students important content relevant for examination
I can stay in touch with students' parents even in difficult situations
I am sure I can be in touch with difficult students when I make an effort
I am sure to address individual problems even better in the future
Even with lesson interruptions, I am sure to keep calm
Even if I do not feel well, I can still pay attention to my students
Even though I try, I cannot change much
I can develop creative lesson ideas to change inconvenient situations
I can excite students for new projects
I can assert innovative ideas opposed skeptical colleagues
Job satisfaction
It is difficult to be happy in my profession
Advantages of the profession dominate
It is a problem to teach in several classes
If I could choose again, I would immediately become a teacher
I have thought about taking up another profession
There is no better profession for me
Sometimes I regret becoming a teacher

Panel B: Languages
Discipline
Students don't listen to the teacher
It is often loud and confused in the classroom
The teacher has to wait long until students are quiet
Students cannot work undisturbed
Students only start working long after the lesson begun
Teaching style
Frontal
Small groups
Single or quiet work
Interdisciplinary
Discussions
Self-organized
Peer tutoring
Week plan
Project-based

(Continued on next page.)

Internal differentiation
Heterogenous groups
Homogenous groups
Low expectations of low performers
Varying tasks
High expectations of high performers
Additional tasks
Tasks for understanding
Move on faster
Different homework
Self-assessed teaching skills
Control disruptive behavior
Follow class rules
Students value learning
Adapt level of teaching
Support to think critically
Calm down disruptive students
Motivate students with little interest
Support weaker students
Encourage to good performance
Use different teaching methods
Give alternative explanations when students are confused
Create rule for orderly lessons
Answer difficult questions
Assess whether students understood
Formulate expectations

Note: The table shows which teacher questionnaire items form the six categories of teaching characteristics.

Table A 4.7: The interplay of teacher specialization, student attitude and achievement, and grades

Panel A: The effect of teacher specialization on student attitude conditional on grades

Sample	Sciences				Languages			
Dep. var.	Self-concept		Subject interest		Self-concept		Subject interest	
Teacher specialization	0.034	0.042	0.186*	0.192**	0.084*	0.056	0.103*	0.079
	(0.084)	(0.070)	(0.097)	(0.084)	(0.048)	(0.038)	(0.059)	(0.051)
Female student x teacher specialization	-0.144	-0.076	-0.353**	-0.300**	-0.100*	-0.041	-0.080	-0.030
	(0.128)	(0.100)	(0.141)	(0.121)	(0.052)	(0.044)	(0.050)	(0.044)
Grades		0.540***		0.415***		0.560***		0.476***
		(0.036)		(0.037)		(0.014)		(0.016)
Net effect for girls	-0.110	-0.034	-0.167**	-0.109	-0.015	0.015	0.023	0.049
	(0.085)	(0.069)	(0.087)	(0.079)	(0.043)	(0.038)	(0.048)	(0.045)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	4,719	4,719	4,719	4,719	36,302	36,302	36,302	36,302
N students	2,045	2,045	2,045	2,045	18,170	18,170	18,170	18,170
Clusters	207	207	207	207	819	819	819	819

(Continued on next page.)

Panel B: The effect of teacher specialization on student achievement conditional on grades and attitudes

Dep. var.	Science skills					Language skills				
Teacher specialization	0.097***	0.097***	0.095***	0.090***	0.092***	0.085***	0.076**	0.069**	0.069**	0.066**
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.031)	(0.031)	(0.030)	(0.029)	(0.029)
Female student x teacher specialization	-0.092***	-0.091***	-0.088***	-0.081***	-0.084***	-0.094***	-0.076***	-0.071***	-0.073***	-0.070***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.028)	(0.027)	(0.027)	(0.026)	(0.026)
Grades		0.008	-0.013	-0.007	-0.013		0.173***	0.100***	0.130***	0.084***
		(0.010)	(0.011)	(0.011)	(0.012)		(0.010)	(0.009)	(0.010)	(0.009)
Self-concept			0.039***		0.025*			0.130***		0.109***
			(0.013)		(0.015)			(0.007)		(0.007)
Subject interest				0.034***	0.019*				0.092***	0.060***
				(0.010)	(0.011)				(0.007)	(0.007)
Net effect for girls	0.005	0.006	0.007	0.009	0.009	-0.009	0.000	-0.002	-0.004	-0.004
	(0.025)	(0.025)	(0.025)	(0.026)	(0.026)	(0.025)	(0.027)	(0.025)	(0.026)	(0.025)
Student-fixed effects	x	x	x	x	x	x	x	x	x	x
Observations	4,719	4,719	4,719	4,719	4,719	36,302	36,302	36,302	36,302	36,302
N students	2,045	2,045	2,045	2,045	2,045	18,170	18,170	18,170	18,170	18,170
Clusters	207	207	207	207	207	819	819	819	819	819

Note: Dependent variable as of the first line. Science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. Students' self-concept is survey constructed from four items. Students' subject interest is survey constructed from four items and it is standardized. Grades are inverted with higher values representing higher achievement. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 4.8: The effect of teacher specialization on teacher characteristics

Sample	Sciences				Languages					
Dep. var.	Female teacher	Gender match	Self-efficacy	Job satisfaction	Female teacher	Gender match	Discipline	Teaching style	Internal differentiation	Teaching skills
Teacher specialization	-0.038	-0.027	-0.109	0.264**	0.115**	-0.020	-0.060	-0.016	-0.025	0.011
	(0.048)	(0.021)	(0.132)	(0.110)	(0.056)	(0.021)	(0.084)	(0.059)	(0.065)	(0.071)
Student-fixed effects	x	x	x	x	x	x	x	x	x	x
Observations	17,059	17,546	17,175	17,075	37,501	38,446	38,446	38,446	38,446	38,446
N students	7,625	7,641	7,624	7,608	19,187	19,223	19,223	19,223	19,223	19,223
Clusters	639	640	639	638	866	868	868	868	868	868

Note: The dependent variables as of the first line. Gender match takes the value one for teachers and students having the same gender. Further teaching characteristics are indices from sub-categories enlisted in Appendix Table A 4.6. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 4.9: The effect of teacher specialization on student achievement conditional on teacher-student gender match

Dep. var.	Science skills				Language skills			
Sample	girls		boys		girls		boys	
Teacher specialization	-0.036*	-0.014	0.069***	0.091***	-0.022	-0.075	0.085***	0.102***
	(0.019)	(0.031)	(0.020)	(0.020)	(0.028)	(0.051)	(0.033)	(0.035)
Gender match	0.033**	0.061*	0.034**	0.069**	0.022	0.009	-0.016	-0.002
	(0.014)	(0.031)	(0.016)	(0.031)	(0.027)	(0.031)	(0.033)	(0.035)
Gender match x teacher specialization		-0.039		-0.052		0.070		-0.068
		(0.038)		(0.035)		(0.054)		(0.064)
Effect for teacher specialization of gender matches	0.217*	-0.053**	-0.053**	0.039	0.002	-0.005	-0.005	0.034
	(0.114)	(0.022)	(0.022)	(0.033)	(0.076)	(0.030)	(0.030)	(0.061)
Student-fixed effects	x	x	x	x	x	x	x	x
Observations	8,688	8,688	8,371	8,371	19,068	19,068	18,433	18,433
N students	3,867	3,867	3,758	3,758	9,736	9,736	9,451	9,451
Clusters	626	626	628	628	858	858	854	854

Note: We divide the sample by student gender. The dependent variables are science skills in 2012 and language skills in 2015 standardized to mean of zero and standard deviation one. Gender match takes the value one for teachers and students having the same gender. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 4.10: The effect of teacher specialization on student achievement identified from two subjects in sciences

Dep. var.	Physics vs. biology		Biology vs. chemistry		Physics vs. chemistry	
Teacher specialization	0.042*	0.087***	-0.004	0.058**	-0.009	-0.000
	(0.022)	(0.028)	(0.023)	(0.028)	(0.025)	(0.028)
Female student x teacher specialization		-0.090***		-0.122***		-0.017
		(0.027)		(0.032)		(0.027)
Net effect for girls		-0.004		-0.063**		-0.017
		(0.024)		(0.028)		(0.028)
Student-fixed effects	x	x	x	x	x	x
Observations	11,152	11,152	12,483	12,483	11,457	11,457
N students	7,641	7,641	7,641	7,641	7,641	7,641
Clusters	640	640	640	640	640	640

Note: The dependent variable derives from the across-subject identification based on tow subjects as of the first line. The test score is standardized to mean of zero and standard deviation one. Teacher controls are those of Table 4.3. Fixed effects apply at the student level. Standard errors are clustered at the class level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 4.11 The Most Frequent Additional Subjects of Teachers

	2012						2015			
	Physics		Chemistry		Biology		German		English	
Specialized Teachers										
1st subject	Math	82.4	Biology	59.0	Chemistry	64.0	English	40.6	German	42.5
2nd subject	Informatics	10.3	Math	30.8	Geography	16.3	French	14.7	French	19.5
3rd subject	Chemistry	9.4	Physics	7.0	Math	13.3	Latin	3.1	Latin	2.2
Non-Specialized Teachers										
1st subject	Math	16.7	Sport	28.2	Sport	44.2	History	31.4	History	22.3
2nd subject	Sport	13.9	English	12.6	German	19.6	Religion	11.8	Sport	16.3
3rd subject	Politics	11.1	Biology	9.7	English	10.2	Sport	10.4	Geography	14.3

Note: This table reports the three most frequent additional subjects studied by teachers reported as share in percent. The numbers can sum up to more than 100 percent, as teachers may obtain a qualification for more than two subjects during their studies.

5 How did EU membership of Eastern Europe affect student achievement?

While we currently experience Euroscepticism with the Brexit and increasing vote shares of nationalistic parties, research agrees on the economic benefits of European Union (EU) membership at the country level (Darvas 2018).[141] Individual-level consequences of EU membership for adults have been investigated (Sinn et al. 2001; Crespo Cuaresma, Ritzberger-Grünwald & Silgoner 2008; Dobson 2009; Baldwin & Wyplosz 2012), but empirical evidence on adolescents is scarce. This chapter examines the consequences from EU membership for student achievement and suggests a positive influence of EU membership on student skills through higher family wealth and increased school efficiency.

Expectations on the consequences for adolescents from the EU accession of Eastern Europe divide into an economic and a political dimension.[142] Economically, EU membership liberates the trade of goods and labor. Classic economic theory predicts rising welfare on the macro- and on the micro-level because a greater variety of goods and services are available at lower prices and employment increases. As a result, national economies and individual households become wealthier, because the higher developed Western EU countries attract lower skilled labor from Eastern Europe. Hence, returns to (further) education increase. Richer families associate with higher performing children (Carneiro & Heckman 2002; Dahl & Lochner 2012; Bulman et al. 2017). Yet, more parental employment comes at cost of family time in which children and adults interact. At its extreme, parents may leave the family (temporarily) to migrate for work. This may decrease student achievement (Banerji, Berry & Shotland 2013; Bergman 2015). In sum, it is unclear how increased parental employment due to EU membership may affect student achievement. Politically, EU membership changes processes in the accessing

[141] For example, the 2004 accessions augmented GDP per capita by 12 percent across all member states and cost-benefit analysis suggest that benefits outweigh the costs of accession (Campos, Coricelli & Moretti 2014).

[142] There is also a psychological dimension, such as an increase in life satisfaction due to EU accession (for Romania and Bulgaria, see Nikolova & Nikolaev (2017) and Zapryanova & Esipova (2016). Yet, my data provides only scarce information on attitudes and aspirations and I therefore do not regard psychological consequences from EU accession. Still, I capture a social dimension of EU membership when analyzing family structure.

countries because European standards aim at raising efficiency and fairness in policy making incentivized by a report system (European Council 2017).

To investigate the effect of EU membership on student achievement, I build a country panel from six waves over 15 years of the Programme for International Student Assessment (PISA). My sample contains 1,073,652 students in 32 countries.[143] The successive entry of eleven Eastern European countries between 2004 and 2013 offers an ideal setting for a difference-in-differences approach.[144] Hence, I regress student achievement on a dummy of EU membership. In the search of mechanisms of transmitting EU membership to student achievement, I consider several inputs to the human capital production function.[145] First, I use these human capital production inputs as alternative outcomes to the treatment of EU membership to verify whether they are well-identified.[146] Second, I use the inputs as covariates to the main specification (of regressing student achievement on EU membership) which provides explorative, non-causal evidence.[147]

The main specification yields a positive and significant estimate of EU membership on student achievement by a decile of a standard deviation (SD). In explorative evidence, I detect well-identified mechanisms transmitting EU membership to student achievement. Verifying results using mediation analysis yields the following three key mediators of the treatment effect. First, the largest shift in the estimate EU membership on student achievement comes from conditioning on school efficiency measured by student tests for external comparisons, teacher absence and shortage, and school location in a city.

[143] Cyprus and Malta participated only two times in PISA which fails my criterium of having participated at least three times to make meaningful comparisons over time.

[144] In 2004, Poland, the Czech Republic, Hungary, the Slovak Republic, Slovenia, Estonia, Lithuania, and Latvia entered the EU. In 2007, Bulgaria and Romania followed and Croatia entered in 2013.

[145] The human capital production function was formalized by Hanushek (1970) and (1979) and more recently by Hanushek & Woessmann (2011) as $Y_i = f(I_i, R_i, F_i, A_i)$. The education outcome Y_i captures skills measured by test scores from (large-scale) assessments at individual level *i*. The input factors are intuitions I_i, school resources R_i, family background F_i, and student ability A_i.

[146] If the alternative outcomes react to the treatment of EU membership, they are well-identified.

[147] Bad controls are potential outcome variables to the treatment themselves and should not be included as control variables where estimates may have a causal interpretation. Bad controls are potential outcome variables because they determine after the treatment. To employ those controls determined after the treatment, one needs explicit assumptions on the timing of control, treatment, and outcome variables. In contrast, relevant variables measured before the treatment is determined are good controls and cannot become an outcome variable of the treatment (Angrist & Pischke 2009).

Second, family wealth influences the effect of EU membership on student achievement measured by lower-status parental occupation. Lastly, single parenting seems to shift the effect of EU membership on student achievement. Hence, mechanisms comprise family, school, and institutional factors.

Difference-in-differences necessitates three key assumptions on the consistency of the treatment and control populations. First, the parallel-trends assumption requires that treated and untreated countries would follow the same trend in the absence of the treatment. This is violated if untreated countries experience a deterioration in student achievement. This may be caused by the comparison group of non-EU Eastern European countries investing less in skills because they expect to replace low-skilled labor in the new Eastern EU members states. For example, the Ukrainian plumber replaces the Polish plumber instead of becoming a financial accountant because the Polish plumber emigrated to work in Germany. Figure 5.1 depicts the parallel trends in students' reading achievement.[148] Additionally, a placebo test where the outcome variable leads the treatment variable by one period yields an insignificant estimate close to zero of EU membership on student achievement. [149] Testing for lagged outcomes yields a small, positive, and significant estimate of EU membership on lagging student achievement. This indicates that EU membership continued to have an effect one period after accession. In summary, the graphical evidence and the placebo test suggest that the parallel trends assumption holds.

The second assumption on constant populations refers to the stable unit treatment variable. This necessitates the absence of spillovers from the treatment to the control group, i.e., if untreated countries react to the EU accession of Eastern Europe. This means that permanent Western EU members decrease in academic performance because they invest less in skills in expectation of hiring highly qualified labor from Eastern EU-Europe. For example, Hungarian medical doctors are hired in Germany, while Germans pass on

[148] Note, that the volatility of the control group of never members (grey markers) originates from its heterogenous composition of Albania, Iceland, Montenegro, Norway, and Serbia, which are observed in PISA in different points in time. The upward kink in 2003 stems from low-performer Albania (378 points on average) missing in 2003 and 2006, while Montenegro (410 points on average) and Serbia (431 points on average) participated in PISA for the first time in 2006. To mitigate concerns on results being driven by this heterogenous comparison group, I subsequently exclude each country in my robustness analysis in Section 5.5.3.

[149] As PISA is surveyed every three years, one period equals three years.

becoming medical doctors. This may be a reality for some doctors, but it seems unlikely on the large scale and across several professions. One reason is that Western EU-Europe's competitive advantage relies on highly qualified human capital.

The third key assumption on steady populations refers to a change in a country's population composition due to selective emigration. Usually, migration is selective towards higher ability and higher status. Yet, the Western European labor markets demand lower skilled labor. If low ability families emigrate from Eastern European entrant countries and leave behind higher performing students, the effect from EU membership on student achievement is upward biased. I compute the migration ratio as number of emigrants relative to their home population and regress student achievement on it. The coefficient is zero, which is potentially due to the small ratio of 0.001. Hence, even though there is explorative evidence of negative selection into emigration[150] and with it those low-ability students disappear from their home country achievement, the treatment of EU membership would be overestimated by the remaining higher achievers. Yet, the number of violators is too small to affect my results. Additionally, I test for sample composition at the country level and ensure the findings are not driven by one single country or wave.

My results relate to two strands of the literature. First, indirect evidence of the EU altering incentives to education comes from the literature on the returns to skills.[151] A few well-identified studies investigate returns to skills when countries transit from communism to EU membership. Increasing returns seem to incentivize raising educational attainment and achievement (Fleisher, Sabirianova & Wang 2005; Farchy 2009; Anniste et al. 2012; Botezat & Pfeiffer 2014; Staneva & Abdel-Latif 2016). While the PISA data has no information on returns to schooling in earnings, my results suggest an increase in family asset wealth and an increase in parental employment due to EU membership, which co-move with higher earnings. Explorative evidence suggests family asset wealth and parental employment are mechanisms that increase student achievement.

A second strand of literature regards the link of family structure and student achievement. Disrupted families, where less than two parents are present, associate with lower student achievement (Wuertz Rasmussen 2009; Francesconi, Jenkins & Siedler

[150] See *Selective Emigration* in Section 5.5.1.

[151] There is no empirical evidence on the effect of EU membership on student outcomes.

2010; Tartari 2015). Yet, selection complicates causal analysis of family structure and student outcomes, as disrupted families tend to be of disadvantaged socio-economics status. For example, single-parent families seem to have lower employment rates, lower earnings, and more instable relationships (Ermisch & Francesconi 2001; Gruber 2004). Addressing the endogeneity issue using family-fixed effects or instrumental variable approaches, yields small or zero estimates of student outcomes due to family disruption (Björklund & Sundström 2006; Björklund, Ginther & Sundström 2007; Sanz-de-Galdeano & Vuri 2007; Francesconi, Jenkins & Siedler 2010). Using a difference-in-differences approach with repeated observations at the country level, the findings suggest an increase in disrupted families due to EU membership. Disrupted families appear to be a mechanism of EU membership decreasing student achievement.

This chapter continues as follows: In Section 5.1, I present background information on the influence of EU membership on education in Eastern Europe. Section 5.2 introduces the empirical strategy, followed by the description of the data in Section 5.3. The results are presented in Section 5.4 – divided into main results and mechanisms as outcomes and as covariates. Section 5.5 reports robustness checks and Section 5.6 concludes.

5.1 The impact of EU membership on education

EU membership affects educational outcomes of adolescents in Eastern Europe through economic and political mechanisms, which are intertwined: the political decision to access the EU demanded institutional prerequisites which in turn produced economic consequences. Before accessing the EU by the *Treaty of Membership*, a *Process of Stabilization and Association* installs the *Copenhagen Criteria*; comprising democracy, rule of law, and human rights (European Council 2017). To implement these criteria, Eastern European institutions have modernized by reducing corruption and realizing more just processes applying human rights, such as freedom of choice realized in travel, work, study, investment, and retirement (Nikolova & Nikolaev 2017). After accessing the EU, a single market integrated formerly planned economies in free trade under

competitive pressure.[152] Some countries even introded the Euro currency (Halász 2015).[153] Capital and labor were legalized to flow freely and employment increased domestically and abroad.[154] For example, employment in Eastern Europe increased from 68 to 73 percent, between 2000 and 2017 (Eurostat 2018b). Overall, economic development has been built on local business, foreign direct investment and trade, employment regulations, policy facilitation, and structural funds (Nikolova & Nikolaev 2017).[155] Hence, EU membership advanced political institutions and economic development in general, which probably augmented family wealth. As a result, one would expect from wealthier families to have academically higher achieving children.

Education policy is also influenced by EU membership, most likely through *soft coordination*.[156] Central instrument are the *Education and Training* frameworks, most recently *ET 2020* that allow member states to cooperate (OECD 2015; 2016b; 2016a). For example, one of the *ET 2020* goals aims at a minimum of 40 percent of people aged 30–34 having completed higher education in each member state. To achieve this higher education goal, preceding education levels need to provide quality education. As a result, higher quality at all levels of education may increase with EU membership. Another EU education policy is the provision of extensive funding, e.g., the *European Social Fund* (ESF) for investments into human capital, such as teacher training or new school curricula

[152] The Soviet Union ended in 1991 – nine years before my period of analysis starts and thirteen years before the first wave of Eastern European countries access the EU. Therefore, I do not expect aftermaths of the fading socialism to disrupt my analysis.

[153] Slovenia, the Slovak Republic, Estonia, Latvia, and Lithuania introduced the Euro currency.

[154] Free movement of labor was regulated by a 2+3+2-transformation model, where the United Kingdom, Ireland, and Sweden allowed labor migration immediately with EU membership. Two years later, Spain, Portugal, Finland, and Greece opened their market. Only seven years after the first round of accessions, in May 2011, Germany and Austria granted free labor migration to the 2004 entrants. For the 2007 entrants, Romania and Bulgaria, Germany allows migration since 2014. One year later, Croatia received the legal right to free labor movement to Germany (bpb 2016).

[155] The benefits of EU accessions of EU accessions of Eastern Europe contrast earlier rounds of EU accessions where economic analysis attributes benefits to investment in physical capital (Baldwin & Sheghezza 1996) and to technological innovation (Rivera-Batiz & Romer 1991).

[156] Soft coordination or the open method of coordination (OMC) ist he EU's instrument which does not produce legaslative binding rules but recommendations evaluatioed by one another (European Union 1998-2019).

(European Commission 2013).[157] As a result, more school resources (efficiently used) may increase academic performance of EU entrants

Overall, EU membership has affected education in Eastern Europe politically through more efficient institutions and economically through increasing funding and family wealth. Hence, one would expect an increase in student achievement from EU membership.

5.2 Empirical strategy

I use a difference-in-differences approach on a country panel over time to identify the effect of EU membership on student achievement. The estimation equation is as follows:

$$A_{i,c,t} = \beta\, EU\ member_{c,t} + \lambda X_{i,t} + \mu_c + \mu_t + \varepsilon_{i,c,t} \qquad (5.1)$$

The dependent variable $A_{i,c,t}$ is student achievement of student *i* in country *c* at time *t*. The variable of interest is $EU\ member_{c,t}$ and takes the value zero for a country *c* in time *t* which is not a member of the EU, and the value one for member states. The matrix $X_{i,t}$ captures student level covariates *i* in time *t*. Country-fixed effects μ_c account for unobserved time-invariant country characteristics, such as higher education funding in one country compared to another country. Time-fixed effects μ_t account for period-specific factors, such as a global trend towards more education. $\varepsilon_{i,c,t}$ is an individual-level error term clustered at the country level which is the treatment level. To shows that results are robust despite the small number of clusters (32), I bootstrap standard errors for the main results following Cameron, Gelbach & Miller (2008).[158]

Equation (5.1) identifies estimates of β from country-level variation over time. The coefficients are unaffected by systematic, time-invariant differences across countries. Hence, countries that do not change their EU membership status in the observation period do not contribute to the estimation of the coefficient β. This difference-in-differences approach builds on four assumptions.

[157] The ESF 2007-2013 supported human capital with € 25.5 billion (European Commission 2013).

[158] Bootstrapping the complete analysis would lead to very long computation times.

First, the common trends assumption necessitates countries to develop parallelly in student achievement in the absence of the treatment. One advantage of the difference-in-differences approach is that EU membership does not need to be random, only the assumption of parallel trends needs to hold. I show parallel trends in Figure 5.2 with decomposed control groups according to permanent EU (black markers) and permanent non-EU members (grey markers), and decomposed treatment groups according to the three accession waves (red, green, and blue markers). The figure suggests parallel trends.[159]

Second, the stable unit treatment variable assumption denies spillovers from treated to untreated countries, i.e., student achievement changes without the change of a country's EU membership status. For example, the comparison group of non-EU Eastern European countries invests less in skills because they expect to succeed low-skilled labor in the new Eastern EU members states. For example, the Ukrainian plumber replaces the Polish plumber instead of the Ukrainian becoming a financial accountant because the Polish plumber emigrated to work in Germany. Another case of untreated countries reacting to the EU accession of Eastern Europe occurs if original Western EU members decrease in academic performance due to expecting to hire highly qualified labor from Eastern EU-Europe. This may be the case for Hungarian medical doctors working in Germany, but is seems unlikely to occur on a large scale because Western EU-Europe's competitive advantage relies on highly qualified human capital.

Third, the population composition remains constant. If individuals migrate between countries and select into or out of treatment, the assignment is not random. For example, if families of low socio-economic background with low student achievement emigrate from their Eastern European countries and the remaining population is of high socio-economic background with high student achievement, my analysis will be upward biased. I will eliminate this concern in Section 5.5.1.

[159] Note, that the volatility of the control group of never members (grey markers) originates from its heterogenous composition of Albania, Iceland, Montenegro, Norway, and Serbia, which are observed in PISA in different points in time. The upward kink in 2003 stems from low-performer Albania (378 points on average) missing in 2003 and 2006, while Montenegro (410 points on average) and Serbia (431 points on average) participated in PISA for the first time in 2006. To mitigate concerns on results being driven by this heterogenous comparison group, I subsequently exclude each country in my robustness analysis in Section 5.5.3.

Fourth, there are no country-specific changes over time in unobservables between treatment and control, such as economic shocks or improving school quality to one group. For example, if the Czech Republic introduced a policy to support school children of low-socio economic status, estimates would be biased. Hence, I condition the analysis on various school and institutional measures; such as school resources, teacher background, school autonomy and accountability, and government funding. Results are reported in Section 5.4.3. Further robustness checks on this assumption are shown in Section 5.5.4.

5.3 Data

I use six waves of the Programme for International Student Assessment (PISA), conducted every three years between 2000 and 2015. The survey tests 15-year-old students independently of the educational institution or grade they attend. Students' competencies in the subjects reading, math, and science are elicited by a two-hour test of tasks varying in difficulty. Using item response theory, achievement in each domain is plotted on a scale with student achievement to a mean of 500 points and a standard deviation of 100 points. Countries employ a two-stage sampling design. First, they draw a random sample of schools in which 15-year-old students are enrolled (with sampling probabilities proportional to a school's number of 15-year-old students). Second, they randomly sample 35 students of the 15-year-old students in each school. The aim is to ensure random sampling of schools and students and to monitor testing conditions in participating countries. I exclude countries that do not meet the standards.[160] PISA does not follow individual students over time, but the repeated testing of representative samples of students creates a panel structure for countries observed every three years. I

[160] The Netherlands in 2000 and the United Kingdom in 2003. I exclude any country-by-wave observation for which the entire data of a background questionnaire is missing; as in France from 2003-2009 (missing school questionnaire) and Albania in 2015 (missing student questionnaire). Liechtenstein was dropped due to its small size.

consider all European countries with and without EU membership.[161] I require countries to participate at least three out of six waves, to deduct meaningful comparisons over time.[162] My final sample contains 1,073,652 students in 32 countries. Summary statistics are displayed in Table A 5.1 and the frequency with which a country participated in PISA is displayed in Table A 5.2 .

In the following, I present the variables which are considered as outcome and control variables. Test score in reading, the main outcome, varies between 2000 and 2015 by Eastern European country, as depicted in Figure 5.2. Especially Bulgaria, Romania, and Hungary experienced large changes. Top-performing Eastern European countries are Estonia and Poland scoring at the level of the Netherlands, while weak-performing Eastern European countries are Bulgaria and Romania scoring between non-EU members Montenegro and Serbia.

Following the education production function, I aim at including control variables at the student, parent, family, school, and country level. At the student level, I examine student gender, age, and migrant background.

At the parent level, I consider parental background as reported in the student questionnaires. I observe whether at least one parent was born abroad and the highest education level of both parents categorized by the International Standard Classification of Education (ISCED) into no education, primary, lower secondary, upper secondary I, upper secondary II, or university. Parental work status could be full time, part time, searching, or other. The item was not asked in 2006 and not in 2015, which I ipolate at the country level to maintain a maximum number of observables.[163] The type of parental occupation is documented in the International Standard Classification of Occupations (ISCO) in nine gradings (manager, professional, technician, clerical, services and sales,

[161] Non-members are Albania, Montenegro, Serbia, Switzerland, Iceland, and Norway. Permanent EU members are Austria, Belgium, Germany, Denmark, Spain, Finland, France, the United Kingdom, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Sweden (for a list of EU membership status by country over time, see Table A 5.4).

[162] Cyprus and Malta participated only two times in PISA which fails my criterium.

[163] Ipolation on student level is impossible, due to resampling of a nationally representative population each wave. Without ipolation, I would lose half the sample; for some variables, I would lose up to three quarters. I report and control for the ipolated share of all variables.

skilled agriculture/ forestry/ fishing, craft and trade, plant and machine operator, and elementary), which was asked in every wave.

At the family level, I use wealth and family composition. Family wealth is represented by four items. First, I exploit the student background questionnaire to capture family wealth by the survey's index of consumer goods constructed from an own room, access to internet, number of phones, TVs, computers, cars, bathrooms, and DVD players. The family wealth index is provided in every survey wave. Second, I use the survey index cultural goods consisting of literature books, poetry, and art work. The cultural goods index was asked every year except in 2009, which I ipolate at the country level. Third, I use the number of books at home in five categories: 0-10, 11-100, 101-200, 201-500, and more than 500 books. The books variable was gathered every year. Lastly, I employ the home educational resources index consisting of a study desk, a quiet study place, a computer for school work, educational software, books for school work, technical reference books, and a dictionary. The home educational resource index was constructed in every survey wave.

To capture family structure, I use the student item "Who lives in your home?" and create binary variables for single mother, single father, or no parents, but living with sisters, brothers, grandparents, or other individuals.[164] It was asked in 2000, 2003, 2009, and 2012. Hence, I ipolate the waves 2006 and 2015 at the country level.

At the school level, I use the following items from the principal questionnaire: city size in five categories in which the school is located (village of less than 3,000 inhabitants; town of 3,000-15,000; large town of 15,000-100,000; city of 100,000-1,000,000; and large city of more than 1 million), private or public operator, number of students, share of government funding, shortage of math teachers, share of fully certified teachers, and degree to which

[164] To eliminate concerns that disrupted families significantly differ from nuclear families, I report socio-economic characteristics of each family composition before and after EU membership in Table A 5.5.

teacher absence is a problem in four categories (not at all, a little, some, a lot). In 2006, teacher absence was not asked and I impute it.[165]

At the country level, I use the share of schools with academic-content autonomy and its interaction with initial GDP per capita as developed by Hanushek, Link & Woessmann (2013) because one expects better information locally instead of centrally dependent on the level of development of a country. To capture school accountability, I employ the share of schools in a country using different forms of student test, such as (i) school-based tests for external comparison, (ii) student-based tests for external comparison, (iii) standardized monitoring, (iv) internal testing, and (v) internal teacher monitoring. These measures were developed by Bergbauer, Hanushek & Woessmann (2018). Lastly, I consider expenditure on education as share of GDP from the World Bank of which I ipolate missing values at the country level.

5.4 Results

This section first reports the main results on the impact of EU membership on student achievement in Sub-section 5.4.1. In the search of mechanisms to the main specification, I advance in two steps. First, sub-section 5.4.2 examines the influence of EU membership on several inputs of the human capital production function at individual, parental, family, school, and country level. If there is a statistically significant link, then those outputs are *well-identified*. Second, using the well-identified factors as controls to the main specification yields non-causal, explorative evidence on mechanisms of transmitting EU membership to student achievement. Hence, sub-section 5.4.3 reports estimates of conditioning on the potential mechanisms.

[165] If a whole country lacks answers in a specific wave, I linearly ipolate it. In Sweden in 2015, the following school characteristics are missing: number of students, private or public operator, share of school budget from the government, city size. The share of fully certified teachers is missing in Denmark in 2006, 2012, and 2015; in Bulgaria in 2012; in Spain in 2009; in Hungary in 2000, 2012, and 2015. The number of students is missing in Albania in 2012; and in Austria in 2012 and 2015. Private or public operator is not reported for Bulgaria in 2006. The share of government budget misses for Austria in 2003, 2006, 2009, and 2015. Iceland does not provide the city size in 2000. In contrast, individual missing variables are not replaced.

5.4.1 Main results on the effect of EU membership on student achievement

Table 5.1 shows the estimation results of the main specification in the country panel model. Regressions are weighted by students' sampling probabilities within countries, giving equal weight to each country-by-wave cell across countries and waves. Standard errors are clustered at the country level throughout. The dependent variable is the test score in one of the three subjects: reading, math, and science. Columns 1, 3, and 5 show the base model. Columns 2, 4, and 6 show the model with time- and country-fixed effects. The coefficient of *EU member* suggests that entering the EU is related to a statistically significant increase in reading achievement at the five percent level. The effect magnitude is a quarter of a standard deviation for the base model and a tenth of a standard deviation when applying country- and time-fixed effects. The point estimate of EU membership on math achievement is of similar magnitude in the base and in the fixed-effects model compared to reading achievement, but it is not statistically significant. In contrast, the EU-membership coefficient on science achievement in the base model is of similar magnitude and significance compared to reading achievement, but the EU-membership coefficient is not statistically significant, small, and negative in the fixed-effects model. This difference across subjects may be due to universally applicable numeracy skills opposed to language- and country-specific literacy skills or due to measurement error. Overall, there is no significant difference between the subjects. In the following, I concentrate on reading achievement.

To cater concerns on the low number of clusters (32), I wild-bootstrap standard errors 1,000 times and report p-values in square brackets below the clustered standard errors. As expected, significance levels decline, but the pattern remains: estimates of EU membership on reading are significant in the fixed-effects model, but not on math or science. Overall, the main specification suggests a positive effect of EU membership on student achievement.

5.4.2 Results on the effect of EU membership on alternative outcomes

This sub-section documents the influence of EU membership on several inputs of the human capital production function. If the inputs did change with EU membership, they are well identified, but bad controls in the sense of Angrist & Pischke (2009).

The three key individual level characteristics – age, gender, and migrant background – are shown in Table 5.2. The point estimates of EU membership seem unrelated to student gender and age. This is plausible, as tested students were begotten sixteen years prior to the study and parental fertility preferences seem unlikely to be influenced in advance to EU membership. In contrast, the EU-membership estimate suggests that entering the EU is related to a significant decrease in first generation migrants by 3.4 percentage points. This reflects that migration from other countries to Eastern Europe has decreased due to EU membership. This finding appears counter-intuitive to the EU's aim of increasing the population's mobility. Yet, relocating the EU's external borders to the Balkan may have complicated settlement for foreigners because the new member states may have controlled their borders more sincerely and had more to lose. Furthermore, together with a significant estimate of EU membership on reading skills, this may hint at the importance of language skills for migration.

Table 5.3 reports the effect of EU membership on parental characteristics. The estimate suggests that entering the EU is related to a significant decrease in parental migration by the same magnitude as of students (3.5 percentage points). This suggests joint migration of children and parents. Columns 2 and 3 of Panel A show that lower levels of parental education were unaffected by EU membership. Parental education should be unaffected by EU membership as it was likely to be completed before their surveyed children experienced the policy shift. In contrast, EU membership seems to significantly decrease higher levels of education by three to seven percentage points. It is plausible to observe a decrease in parents' higher education as EU membership increased returns to schooling. In reality, average annual net earnings of a single person in the eleven Eastern European EU entrants increased from 3,022€ to 7,482€, between 2000 and 2015 (Eurostat 2018a). This is in line with research from Norway, where the unexpected discovery of oil resources increased returns to lower education and decreased educational attainment (Bütikofer, Dalla-Zuanna & Salvanes 2017). Panel B of Table 5.3 reports estimation results for parental work status. The coefficient of EU membership is never significant for mothers. In contrast, the point estimate of EU membership for fathers working full time is significant at the ten percent level, suggesting an increase by 3.7 percentage points. The point estimate of EU membership for fathers looking for work is significant at the five percent level, suggesting a decrease by 2.1 percentage points. Panel C of Table 5.3 reports estimation results for parental occupation. The coefficient of EU membership suggests that parents were significantly less employed as professionals by 2.7 percentage points

and more as clericals by 1.5 percentage points, in services and sales by 1.8 percentage points, and in elementary jobs by 1.8 percentage points. Overall, Table 5.3 suggests a decrease in parental migration and a decrease in parental higher education, while fathers seem to work more and parents work rather in low-status occupation. These results are in line with expectations where Western EU Europe demands low-qualified labor.

Table 5.4 reports outcomes of the impact of EU membership at the family level. In Panel A, family wealth is expected to rise with EU membership due to an expansion of parental labor. The estimate suggests that entering the EU is related to a significant increase in consumer goods by 18 percentage points, while cultural goods decline by a similar magnitude. Additionally, the EU membership coefficient suggests a highly significant decrease in having more than 100 books at home by seven to two percentage points. In contrast, the EU membership coefficient is insignificant for educational resources. This surprising result suggests a shift in consumer preferences towards an expansion of status goods at the cost of intellectual goods. Panel B shows estimates on the family structure. The estimate suggests that entering the EU is related to a significant increase in single parents by 3.1 percentage points for mothers and 1.4 percentage points for fathers. At its most lonely form, EU membership seems to significantly increase the share of children living without any parent by 1.5 percentage points. Hence, EU membership created *Euro orphans.* Overall, EU membership seems to have increased status goods at the cost of intellectual goods and seems to have disrupted families.

Table 5.5 shows results at the school level. In Panel A, the point estimate suggests that entering the EU is related to an increase in schools located in villages by 3.8 percentage points and a decrease in towns by 3.9 percentage points and in large towns by 6.6 percentage points. This shift to the countryside is consistent with international goals of increasing access to schools in remote rural areas, such as envisaged by the Millennium Development Goal No. 2. Panel B reports results on school resources, which were likely increased by EU funding. The point estimate of EU membership suggests a reduction in the number of students per school by 11.5 percent and a decline in math teacher shortages by 5.4 percentage points. This suggests an improvement in school resources. Whether the resources were transformed effectively into student learning is examined in Section 5.4.3, where school resources are employed as covariates to the effect of EU membership on reading achievement. Other school characteristics, such as private or

public operation, the share of government budget, the share of certified teachers, and problems with absents teachers show no significant link to EU membership.

At the country level, the EU's soft coordination may have triggered policy reforms. Estimation results are documented in Table 5.6. The point estimate of EU membership does not significantly link to school autonomy. This may be due to decentralization of the socialist school systems right after the dissolution of the Soviet states and not due to EU accession. School accountability seems to be significantly affected by EU accession in some dimensions. The EU membership coefficient suggests a decrease in school-based tests with external comparisons by 9.4 percentage points and a decrease in internal teacher monitoring by 7 percentage points, while student-based external tests with external comparisons increased by 28.6 percentage points. Standardized monitoring and internal testing seem unaffected by EU membership. These estimates confirm that accountability regimes gained strength in the 2000s by shifting from internal testing to standardized tests with external comparability. Another country-level education determinant is government expenditure on secondary education. The point estimate suggests a small negative, though insignificant, link of EU membership to government expenditure on secondary education. This may be due to a reduction in domestic education funding in response to increased EU funding.

In conclusion, estimates of EU membership suggest ambiguous effects for educational input factors. On the one hand, the share of adolescent and adult migrants decreased, fathers work more, families own more status goods, schools are more rural and have more resources, and accountability shifted from internal to external testing. On the other hand, parental higher education decreased, parents work in lower-status jobs, intellectual goods diminished, and parents left their children. Hence, while material wellbeing seems to have improved, while social wellbeing deteriorated.

5.4.3 Mechanisms to the effect of EU membership on student achievement

In this sub-section, I employ reading score as outcome to EU membership conditional on the inputs to the human capital production function used as outcomes in the previous sub-section. In the preceding section, some of these variables were affected by the treatment. Hence, they are well-identified though bad controls of the main specification and serve as explorative, non-causal evidence. As a robustness check, I show results from mediation analysis at the end of this section.

The following tables expand its predecessors by a new first column to report the main specification in a reduced sample. The reduced sample results from missing observations in variables, which should not be imputed. Hence, samples vary, but most samples count more than 1 million observations, except for the samples including school- and country-level mechanisms. The succeeding columns report results conditional on mechanisms.

Table 5.7 suggests that the point estimate of EU membership on reading scores decreases slightly in magnitude and significance by 2 points conditional on student characteristics. Gender and age show positive and significant estimates on student achievement. Yet, they manifest pre-treatment. Thus, gender and age are good controls and no mechanisms. Migrant status shows a negative and significant estimate on student achievement and was well-identified in Section 5.4.2. Still, the coefficient of EU membership remains almost unchanged. Hence, student migrant status is a minor mechanism of EU membership influencing student achievement.

Parental characteristics as mechanisms are displayed in Table 5.8. In Panel A, the estimate of EU membership shrinks marginally (1 point) when conditioning on parental migrant status. Hence, parental migrant status is a minor mechanism of transmitting EU membership to student achievement. Conditioning on parental education yields positive and significant estimates increasing in size when moving from primary to university level. The coefficient of EU membership increases by 4 points. This may be due to more educated parents tend to have information or skills to derive advantages in student achievement from EU membership. As a result, parental education is a relevant and well-identified (as of Section 5.4.2) mechanism. Panel B presents the estimates of EU membership on student achievement conditional on parental labor. The point estimate of EU membership remains similar to the main specification when adding mechanisms for maternal and paternal work status. The estimates of parental work status are significant and positive – except for mothers looking for work – and significant and negative for fathers – except for fathers with other work status. These findings suggest that working mothers increase student achievement, while working fathers decrease it. This may be linked to the different kinds of work, hours away from home, income, and time spend with the child by mothers and fathers which I do not observe in my data. In contrast, the coefficient of EU membership increases by 3 points conditional on parental occupation. Higher status occupations, such as professionals or technicians, expose a positive estimate on student achievement. Among the negative estimates, elementary

occupations expose the largest coefficient because low-status occupations tend to link to other dimensions of low socio-economic and low ability background resulting in low student achievement. Yet, the interplay of socio-economic background and student achievement seems unrelated to parental involvement, as correlation analysis shows.[166]

Table 5.9 reports estimates for family characteristics. The estimate of EU membership increases markedly in magnitude and significance by 6 points conditional on family wealth. The estimates of consumer and cultural goods, the number of books, and home educational resources are large, positive, and significant. Especially possessing more than 500 books seems to link to an advantage in reading scores. This may be due to high socio-economic status, i.e., many of books or highly educated parents which may incentivize children to read. Similar to the previous section, consumer goods have a larger coefficient than cultural goods. Conditional on family structure, the estimate of EU membership shows not very responsive (1 point), while the coefficients of single mother and single father are significant, large, and negative.

Potential mechanisms at the school level are documented in Table 5.10. The EU-membership estimate is unaffected qualitatively by potential school mechanisms, even for the well-identified characteristics, such as number of students and shortage of math teachers. As previous studies have shown, school resources are no strong predictors of student achievement (Hanushek & Rivkin 2006).

Table 5.11 shows mechanisms at the country level.[167] The EU membership coefficient is slightly affected by a decrease of 2 points. The estimate of school autonomy is of expected size but insignificant. In the setting of EU accession, this is not surprising, as former socialist countries may have decentralized their education system already in the 1990s. Tests for external comparison (school-based and student-based) and internal testing yield positive point estimates, while internal teacher monitoring and standardized monitoring yield negative point estimates. This is consistent with Bergbauer, Hanushek & Woessmann (2018). The positive estimate of internal testing, which informs or monitors

[166] See pairwise correlations of Table A 5.6. Correlation coefficients are small and indicate little connection. I fall back to pairwise correlations, because PISA background questionnaires provide items on parent-child interactions in single waves. I do not run regressions drawing on only one wave, but report pairwise correlations.

[167] The sample shrinks due to fewer observations of national tests used for student career decision from Eurydice (2009).

progress without external comparability and internal teacher monitoring including inspectorates, was originally found for poorly performing countries when entering the PISA study. The findings suggest that more targeted information creates stronger incentives, i.e., that incentives to students with consequences for their school career and with external comparability are more tangible and contribute more to student achievement. In contrast, testing seems to set adverse incentives to teachers. Importantly, the results on EU membership effects are not confounded by the potentially coincidental introduction of policies that alter autonomy and accountability. Surprisingly, expenditure on secondary education yields a negative but small significant estimate on student achievement given the other country-level mechanisms, which hints at an inefficient use of school resources. In summary of the country-level mechanisms, the institutional frame, i.e., accountability, seems to be more decisive for student achievement than the economic conditions of a country.

In a final exercise of conditioning on mechanisms, I include the entire set of mechanisms as reported in Table A 5.3. The coefficient of EU membership shows unaffected (from 14.298 points without mechanisms in the available sample to 15.736 points with all mechanisms). The fact that results are insensitive to the included set of relevant mechanisms reduces concerns that estimates are strongly affected by omitted variable bias from unobserved characteristics (in the sense of Altonji, Elder & Taber (2005)).

As bad controls suffer from endogeneity and selection bias, mediation analysis seems to deliver more causal evidence under the assumption on the exogeneity of the mediator. Mediation analysis was pioneered by Imai et al. (2011).[168] Two additional assumptions are necessary. Beyond the standard assumption of random treatment assignment across pre-treatment confounders (e.g., EU accession is exogenous to student gender), mediation analysis demands that the observed mediator is independent of potential outcomes and confounders given the actual treatment (e.g., parental occupation given EU accession and pre-treatment confounders). Thus, conditional on other confounders, the mediator is exogenous to the outcomes, i.e., student achievement. As a result, mediation analysis yields the quantity of how much of the treatment is transmitted by the

[168] The "*mediation"* package implements a command in Stata following Hicks & Tingley (2011).

mediator.[169] I report this share for each mediator in Table 5.12.[170] Similar to the traditional approach of including controls, mediation analysis suggests that the largest share of mediated effects on the student level comes from migrant status (6.8 percent). At the parental level, mediation analysis assigns a small share of mediated effects of the treatment to parental education and medium shares to parental work status and large shares to parental occupation. Yet, the largest shares expose craft and trade (13.6 percent) and plant and machine operators (11.4 percent). Mediation analysis assigns a large share of the treatment effect to single mothers (12.4 percent) and single fathers (10.6 percent). At the school level, mediation analysis suggests large mediating effects on the treatment from school location in a city with 100,000 to 1 million inhabitants (16.6 percent), a shortage of math teachers (10.8 percent), and a little and a lot of teacher absence (18.6 percent and 25.1 percent). Concerning country-level mediators, student-based tests for external comparison expose the largest share of mediated effects (39.1 percent). Overall, mediation analysis suggests similar effects to mediate the effect of EU membership on student achievement as traditional controls do: school efficiency, family wealth, and family structure.

As a result of the mediation analysis, the following mediators transmitted the largest share of the treatment (listed decreasing in size): student-based tests for external comparisons, teacher absence, school location in a medium-sized city, parental occupation in craft and trade and as plant and machine operator, single parenting father, and shortage of math teachers. Overall, conditioning on mechanisms shows no change in the coefficient of EU membership on student achievement. As bad controls suffer from endogeneity and selection bias, mediation analysis seems to deliver more causal evidence under the assumption on exogeneity of the mediator.

[169] The package allows to include each mediator separately, but not several at the same time.

[170] Note, that I executed mediation analysis in a panel on country-wave level, as computations are not possible in a panel at the individual level due to limited memory capacity. I compare estimates of the main specification in the country-wave level data to the individual level data in Table A 5.7 where I compare estimate of. Further, the mediation command does not support country- and wave-fixed effects in this setting. Therefore, I residualized the two fixed effects in the main specification following Frisch & Waugh (1933).

5.5 Robustness tests

My findings prove robust to several potential caveats. I consider selective migration, anticipation or delay of the effect from EU membership, sample composition, and alternative fixed effects.

5.5.1 Selective emigration

This sub-section aims at providing evidence on fulfilling the difference-in-differences assumption of an unchanged population. This intention is complicated by PISA's resampling of a representative set of students in every wave instead of an actual panel following the same students over time. Changes in the population pose a problem if estimates are overvalued. Usually, migration is selective towards higher ability and higher status groups of a country. Yet, the Western European labor markets demand for lower skilled labor and Eastern European emigrants seem to be of lower status. The left-behind home population may be more able and their children achieve higher student test scores. As a result, the home population would reach higher test scores due to emigration of the low performers. Then, my findings would be overestimated due to selected emigration. To address this problem, I provide descriptive evidence.

Emigration is typically directed from Eastern Europe to other Eastern European and Western European countries, especially to the direct neighbors of EU entrants, Austria and Germany (as documented in Table A 5.9). To better understand emigration patterns, family background characteristics are explored in Table A 5.8, before and after EU membership. There is no common pattern across Eastern European countries for parental education. In Estonia and Lithuania, the home population is better educated than the emigrant population. In contrast, for Hungarian and Polish emigrants, parental education of emigrants has been higher relative to the home population. For nationals of the Czech Republic, Estonia, Romania, Slovenia, and the Slovak Republic, parental home and emigrant population was educated about the same. For Czechs, this equality vanishes with EU membership - emigrants became more educated. In contrast, the Croatian emigrants did not increase educational attainment with EU membership, but their home population did. Highly educated Hungarians and Polish people emigrated to their direct neighbors, Austria and Germany, while the economically vibrant Baltic states,

Estonia and Lithuania, could retain their highly educated population. Regarding parental occupation, Eastern Europeans seem to work in lower-status occupations abroad relative to their home population and their status decreased further with EU membership, at home and abroad.

Student achievement of first-generation migrants hints at how well children fare in their new environment. For most Eastern European countries, student achievement of the emigrant population is lower than of the home population. It seems that emigrant children cannot profit from host countries and that home countries have decently developed school systems. In Romania and in the Slovak Republic, the home population performs at the level of their emigrant population, which may be due to weaker education systems at home. Comparing student achievement before and after EU membership shows that Eastern European home populations increased their achievement while emigrants decreased achievement.

To evaluate the magnitude of emigration, I report the emigrant ratio as the number of emigrants relative to their home population, which averages to 0.001 percent (Column 5).[171] As evident from Table A 5.9, I only observe very few migrant students from each single Eastern European country. Figure A 5.1 confirms that the migration ratio in each country did not react to EU accession. Employing the emigration ratio as an outcome variable in equation (5.1) yields a point estimate of EU membership of zero, see Table 5.13 Column 1. Hence, the emigrant ratio is unrelated to EU membership. In a second step, I test the emigration ratio as a potential mechanism of transmitting EU membership to student achievement. Column 3 shows that the estimate is unresponsive to conditioning on the emigration ratio, compared to the coefficient of the main specification in the reduced sample in Column 2. Yet, the point estimate suggests that increasing the number of emigrants from Eastern Europe relative to their home population by one percent decreases reading scores by forty percent of a standard deviation. This sizeable effect advocates that a larger emigration ratio of potential low performers decreases student achievement, while the effect is not well-identified.

In conclusion, parental decisions on emigration given their educational attainment does not seem to follow a common pattern across Eastern Europe. Parental occupational

[171] I disregard migrants from other countries than the Eastern EU entrants, such as Spain or France, and only regard migrants from Eastern Europe to other EU states (East and West).

status and student achievement is generally lower in the host country relative to the home country, which hints at a negative selection of emigrants leaving behind the high ability population. However, the low emigration ratio and explorative regression analysis provides evidence against an overestimation of my findings which suggest that EU membership increased student achievement.

5.5.2 Dynamics of the EU accession

A remaining confounder in the difference-in-differences model with country- and time-fixed effects is the endogeneity of EU membership. The *Process of Stabilization and Association* preceding EU accession reforms political and economic institutions in the sense that entrant countries may already be on a higher trend than non-candidates. The common-trends assumption would be violated. The data's panel structure lends itself for a placebo test. If there is no anticipation of the EU membership, there should be no effect on the achievement of students in the wave before EU membership. However, if EU membership was endogenous, I would yield significant estimates prior to achievement. Therefore, to conduct the placebo test, I create leads of the reading outcome variable relative to the EU accession by one period.[172] Table 5.14 reports the results of this placebo test. In Column 1, the point estimate of EU membership is small, negative, and not significantly related to the leading student achievement. This result advocates that EU membership is not endogenous.

Another dynamic of the EU membership effect could be enduring or delayed effects where not all institutional reforms and economic possibilities were realized at EU accession and needed time to be taken up. If there is a delay in student achievement to EU membership then the estimate of EU membership may be significant one wave after EU membership. I create lags of the reading outcome variable relative to the EU accession by one period. Column 2 reports the results of the lagged placebo test. The small point estimate of EU membership relates significantly to the lagged student achievement and suggests a continuation of positive effects of EU membership on student achievement one wave after accession. Hence, benefits from EU membership endure.

[172] PISA is surveyed every three years. Thus, one period corresponds to three years.

5.5.3 Sample composition

To ensure that my results are not driven by a specific country, I rerun the main specification excluding one country at a time. The qualitative results are insensitive to this sample alteration, with coefficients remaining significant and of similar magnitude, compare Panel A of Table 5.15.

To ensure that results are not driven by one wave, I exclude one wave at a time. In Panel B, the estimates of EU membership are unresponsive to excluding waves, except for wave 2006; where the coefficient decreases in significance and in magnitude by one third. This is not surprising, as eight out of eleven countries become EU members in that wave. This change of the coefficient suggests heterogenous treatment effects, which are stronger for the first wave of entrants as opposed to the two later waves. [173] This more intense first treatment effect is likely caused by entrants being direct neighbors to original EU members with high demand for low-skilled labor, such as Germany and Austria.

5.5.4 Specification test on fixed effects

Another robustness check validates the assumption of the absence of country-specific shocks over time in unobservables between treatment and control. I compensated for observable school quality by including various school and institutional measures, such as school resources, teacher background, school autonomy and accountability, and government funding. Results were reported in Section 5.4.3. Second, I allow for country-specific time trends. Table 5.16 shows the estimation results. The coefficient of EU membership decreases by one third but remains statistically significant in all three subjects. Hence, the model holds against country-specific time trends.

5.6 Conclusion

This chapter examined the consequences of EU membership of Eastern European countries on student achievement. I used six waves of PISA data in a country panel over fifteen years with more than one million individual observations in 32 countries.

[173] Therefore, I forego robustness checks by means of an event study, as this assumes that the three accession waves had the same effect.

Employing a difference-in-differences approach, I find that, entering the EU links to an improvement in student achievement in reading by a tenth of a standard deviation.

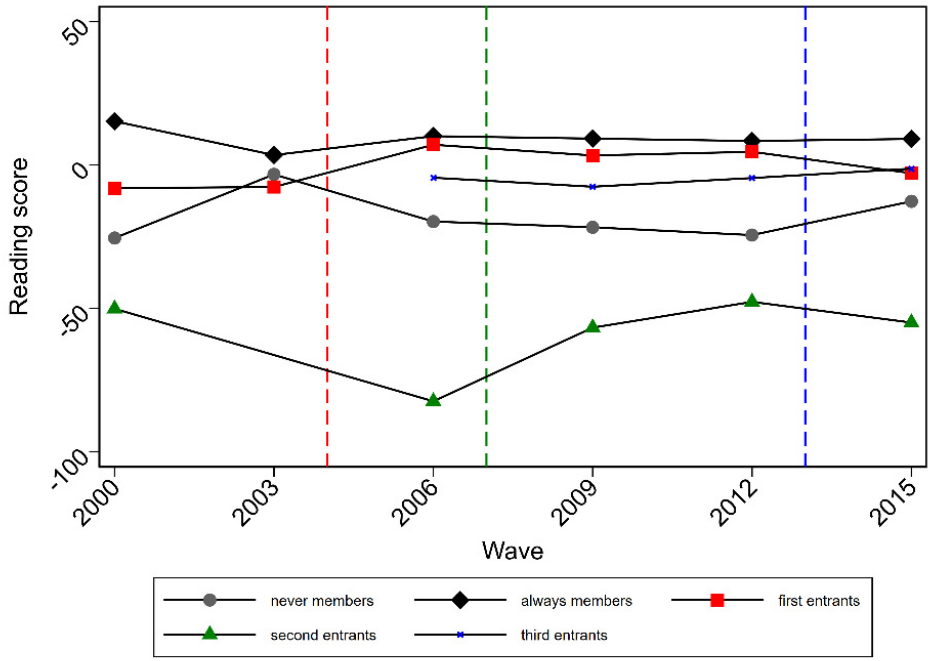
In search of mechanisms transmitting EU membership to student achievement, I test alternative outcomes from different levels of the human capital production function and find EU membership had two key effects. First, families' material wellbeing and school resources and institutions seem to have improved. For example, fathers work more, families own more status goods, schools are more rural and have more resources, and accountability shifted from internal to external testing. Second, families' social wellbeing seems to have deteriorated. For example, parental higher education decreased, parents work in lower-status jobs, intellectual goods diminished, and children live without their parents.

In a further step, I use the alternative outcomes as mechanisms to the main specification and in mediation analysis. The following mechanisms are key mediators of the treatment: tests for external comparisons, teacher absence, school location in a city, parents in lower-status occupations, single parenting, and shortage of math teachers.

Verifying the assumptions of the difference-in-differences approach, I confirm the parallel trends assumption. Robustness tests mitigate concerns on negatively selected emigration being too small in magnitude to bias estimates, absence of anticipation and an afterglow of EU accession. Results are not driven by one country but they rely on including the wave 2006. Furthermore, estimates are robust to country-specific time trends.

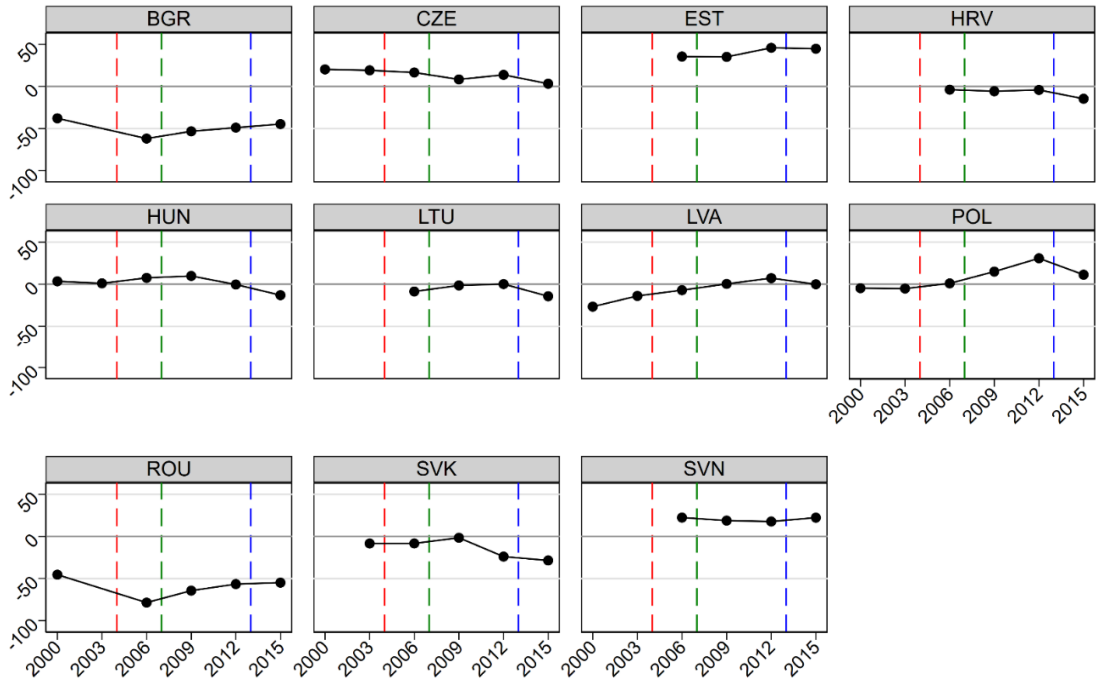
Overall, EU membership fulfilled its promise of economic and educational prosperity by increasing parental occupation and with it family wealth, and by improving school efficiency. On the downside, EU membership disrupted families with an increase of children living with one or with neither of their parents. In summary, EU membership increased student achievement.

Figure 5.1: Parallel trends of reading score by treatment and control group



Note: The graphs depict the evolution of the demeaned reading score over time in the different treatment and control groups. The reading score was demeaned by the sample average to absorb time trends. The treatment group composes of the red markers indicating the first round of EU entrants, the green markers for the second round, and the blue markers for the third round. The control group consists of the black markers standing for permanent EU members, and the grey markers for permanent non-EU members. The figure shows the reading score in each survey wave between 2000 and 2015. The red, dashed, vertical line signals the 2004 entries. The green, dashed, vertical line indicates the 2007 entries. The blue, dashed, vertical line designates the 2013 entry. The first group of EU entrants contains the Czech Republic, Estonia, Hungary, Lithuania, Latvia, Poland, the Slovak Republic, and Slovenia; the second group includes Bulgaria and Romania; and the third entrants group is formed by Croatia. Permanent EU members are Austria, Belgium, Denmark, Finland, France, the United Kingdom, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Sweden. Permanent non-EU members are Albania, Iceland, Montenegro, Norway, and Serbia.

Figure 5.2: Evolution of reading achievement in Eastern Europe



Note: The figure shows the demeaned reading score for each Eastern European entrant in each survey wave between 2000 and 2015. The score was demeaned by the average of all 32 countries in the sample. The red, dashed, vertical line signals the 2004 entries. The green, dashed, vertical line indicates the 2007 entries. The blue, dashed, vertical line designates the 2013 entry. Country abbreviations stand for: Bulgaria (BGR), the Czech Republic (CZE), Estonia (EST), Croatia (HRV), Hungary (HUN), Lithuania (LTU), Latvia (LVA), Poland (POL), Romania (ROU), the Slovak Republic (SVK), and Slovenia (SVN).

Table 5.1: Main result - The effect of EU membership on student achievement

Dep. var.	Reading score		Math score		Science score	
EU member	26.086**	9.667**	21.619	7.648	25.825**	-1.314
	(12.165)	*(3.814)*	*(14.045)*	*(4.692)*	*(12.034)*	*(4.359)*
	[0.104]	*[0.032]*	*[0.218]*	*[0.184]*	*[0.094]*	*[0.770]*
Constant	464.663***	376.124***	471.549***	383.010***	471.372***	382.940***
	(12.029)	*(2.785)*	*(13.805)*	*(3.332)*	*(11.611)*	*(3.182)*
	[0.000]	*[0.000]*	*[0.000]*	*[0.000]*	*[0.000]*	*[0.000]*
Observations	1,073,652	1,073,652	1,021,595	1,021,595	1,021,522	1,021,522
R-squared	0.013	0.088	0.009	0.100	0.013	0.090
Fixed Effects	-	x	-	x	-	x

Note: Sample mean of reading score is 487 points, of math score is 491 points, and of science score is 494 points. When indicated by x, the model controls for time- and country-fixed effects. Standard errors are clustered at the country level reported in parenthesis (*** p<0.01, ** p<0.05, * p<0.1). 1,000 times wild bootstrapped p-values are in square brackets. Least squares regression weighted by students' sampling probability.

Table 5.2: The effect of EU membership on student characteristics

Dep. var.	Female	Age	Migrant
EU member	0.002	0.035	-0.034*
	(0.007)	(0.057)	(0.020)
Constant	0.002	0.035	-0.034*
	(0.007)	(0.057)	(0.020)
Observations	1,072,650	1,072,448	1,032,304
R-squared	0.000	0.057	0.035

Note: The model controls for time- and country-fixed effects. Standard errors are clustered at the country level reported in parenthesis (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.3: The effect of EU membership on parental characteristics

Panel A: Migration and education

		Education					
Dep. var.	Migrant	No education	Primary	Lower secondary	Upper secondary I	Upper secondary I	University
EU member	-0.035***	0.004	-0.004	-0.030**	-0.050***	-0.067**	-0.040**
	(0.007)	(0.002)	(0.002)	(0.011)	(0.016)	(0.030)	(0.019)
Constant	-0.021***	0.027***	0.973***	0.899***	0.707***	0.609***	0.236***
	(0.005)	(0.002)	(0.002)	(0.009)	(0.011)	(0.015)	(0.015)
Observations	1,032,304	1,038,580	1,038,580	1,038,580	1,038,580	1,038,580	1,038,580
R-squared	0.070	0.034	0.034	0.114	0.125	0.101	0.076

Note: The education categories refer to at least having obtained this level of schooling compared to all levels below. No education and university education compare to all other categories. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Panel B: Work status

	Mother				Father			
Dep. var.	Works full time	Works part time	Looks for work	Other	Works full time	Works part time	Looks for work	Other
EU member	-0.021	0.005	-0.013	0.028	0.037*	-0.002	-0.021**	0.030*
	(0.016)	(0.013)	(0.010)	(0.018)	(0.022)	(0.007)	(0.009)	(0.017)
Constant	0.282***	0.071***	0.159***	0.500***	0.601***	0.191***	0.105***	0.324***
	(0.005)	(0.005)	(0.004)	(0.022)	(0.007)	(0.003)	(0.003)	(0.016)
Observations	1,049,909	1,049,909	1,049,909	1,043,542	1,032,847	1,032,847	1,032,847	1,035,379
R-squared	0.130	0.137	0.032	0.034	0.051	0.024	0.028	0.085

Note: All variables are dummies taking the value 0 or 1 at individual level and therefore represent the share of a country. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Panel C: Occupation

Dep. var.	Manager	Professionals	Technicians	Clerical	Services and sales	Skilled agriculture/ forestry/ fishing	Craft and trade	Plant and machine operators	Elementary
EU member	0.007	-0.027**	-0.017	0.015***	0.018**	0.006	0.003	-0.000	0.013**
	(0.018)	(0.011)	(0.013)	(0.005)	(0.007)	(0.004)	(0.008)	(0.007)	(0.005)
Constant	0.125***	0.160***	0.048***	0.036***	0.118***	0.070***	0.169***	0.070***	0.120***
	(0.008)	(0.007)	(0.004)	(0.003)	(0.005)	(0.002)	(0.004)	(0.002)	(0.003)
Observations	1,073,652	1,073,652	1,073,652	1,073,652	1,073,652	1,073,652	1,073,652	1,073,652	1,073,652
R-squared	0.017	0.023	0.012	0.008	0.008	0.014	0.023	0.007	0.013

Note: All variables are dummies taking the value 0 or 1 at individual level and therefore represent the share of a country. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.4: The effect of EU membership on family characteristics

Panel A: Family wealth

			Books at home					
Dep. var.	Consumer goods	Cultural goods	0 – 10	11 – 100	101 – 200	201 – 500	>500	Home educational resources
EU member	0.180***	-0.191***	0.017	-0.017	-0.078***	-0.056***	-0.022***	-0.044
	(0.063)	(0.049)	(0.012)	(0.012)	(0.022)	(0.016)	(0.008)	(0.034)
Constant	-1.727***	0.033	0.360***	0.640***	0.165***	0.080***	0.047***	0.248***
	(0.041)	(0.024)	(0.005)	(0.005)	(0.008)	(0.007)	(0.004)	(0.037)
Observations	1,062,602	1,055,592	1,050,897	1,050,897	1,050,897	1,050,897	1,050,897	1,063,940
R-squared	0.227	0.065	0.033	0.033	0.043	0.027	0.014	0.023

Note: The index consumer goods includes an own room, access to internet, number of phones, TVs, computers, cars, bathrooms, and DVD players. The index cultural goods includes literature books, poetry, and art work. Reference category for the number of books are 0 to 10 books. The index home educational resources contains a study desk, quiet study place, computer for school work, educational software, books for school work, technical reference books, and dictionary. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Panel B: Family structure

Dep. var.	Both parents	Single mother	Single father	Without parents
EU member	-0.058***	0.031**	0.014***	0.015**
	(0.018)	(0.014)	(0.003)	(0.006)
Constant	0.806***	0.068***	0.019***	0.040***
	(0.008)	(0.005)	(0.001)	(0.002)
Observations	1,064,038	1,025,676	1,043,859	1,058,528
R-squared	0.026	0.030	0.005	0.013

Note: All variables are dummies taking the value 0 or 1 at individual level and therefore represent the share of a country. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.5: The effect of EU membership on school characteristics

Panel A: Location

Dep. var.	Village (less 3,000)	Town (3,000-15,000)	Large town (15,000-100,000)	City (100,000-1,000,000)	Large city (>1,000,000)
EU member	0.038*	-0.039*	-0.066**	-0.007	0.000
	(0.019)	(0.019)	(0.024)	(0.012)	(0.000)
Constant	0.216***	0.783***	0.538***	0.017*	1.000
	(0.011)	(0.011)	(0.010)	(0.008)	(0.000)
Observations	1,038,771	1,029,941	1,029,941	1,029,941	59,150
R-squared	0.090	0.089	0.077	0.064	-

Note: All variables are dummies taking the value 0 or 1 at individual level and therefore represent the share of a country. The model controls for the share of ipolated school location observations and for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Panel B: Resources

						Problem with absent teachers			
Dep. var.	Private	Number of students	Government budget	Shortage of math teachers	Certified teachers	Not at all	A little	Some	A lot
EU member	0.010	-74.828**	-5.285	-0.054**	0.028	-0.002	0.010	0.012	0.010
	(0.013)	(31.154)	(3.667)	(0.024)	(0.059)	(0.074)	(0.025)	(0.014)	(0.013)
Constant	0.071***	584.156***	-6.119**	0.098***	0.390***	0.460***	0.102***	0.063***	0.071***
	(0.007)	(14.479)	(2.999)	(0.020)	(0.024)	(0.020)	(0.016)	(0.009)	(0.007)
Observations	1,033,531	1,014,384	1,017,502	1,026,860	1,029,536	1,029,536	1,029,536	1,029,536	1,033,531
R-squared	0.306	0.315	0.448	0.124	0.201	0.136	0.065	0.041	0.306

Note: All variables are dummies taking the value 0 or 1 at individual level and therefore represent the share of a country. The model controls for the share of ipolated observations of private, number of students, and share of government budget. It also conditions on time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.6: The effect of EU membership on country characteristics

				Tests			
Dep. var.	Autonomy	School-based external comparison	Student-based external comparison	Standardized monitoring	Internal testing	Internal teacher monitoring	Gov. expenditure on sec. education
EU member	-0.059	-0.094*	0.286*	-0.041	-0.053	-0.070**	-0.235
	(0.050)	(0.050)	(0.157)	(0.027)	(0.031)	(0.029)	(1.178)
Constant	0.515***	0.562***	-0.349**	0.772***	0.666***	0.755***	0.278
	(0.034)	(0.043)	(0.148)	(0.035)	(0.016)	(0.027)	(1.000)
Observations	1,073,652	1,059,186	994,129	1,059,186	1,059,186	1,059,186	782,274
R-squared	0.865	0.824	0.790	0.825	0.833	0.925	0.799

Note: Autonomy is the share of schools with academic-content autonomy. Tests also report the share of school in a country which employ standardized student tests for the different purposes. Autonomy and tests are both derived from the PISA principal questionnaire. GDP per capita is measured in international US$ in PPP, government expenditure per secondary student is a share of GDP per capita. The model controls for the ipolated share of secondary government expenditure and for time- and country-fixed effects and additionally in the last column for the ipolated share of secondary government expenditure. Standard errors are clustered at the country level.

Table 5.7: The effect of EU membership on student achievement conditional on student characteristics

Dep. var.	Reading score	
EU member	9.222**	7.104*
	(4.323)	(3.960)
Female		37.455***
		(1.496)
Age		14.689***
		(1.805)
Migrant student		-44.945***
		(4.734)
Constant	378.015***	128.835***
	(2.653)	(28.529)
Observations	1,031,557	1,031,557
R-squared	0.090	0.137

Note: The mean of the dependent variable is 487 points. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.8: The effect of EU membership on student achievement conditional on parental characteristics

Panel A: Migration and education

Dep. Var.	Reading score			
EU member	9.098**	8.713*	9.334**	13.009***
	(4.315)	(4.325)	(3.727)	(4.172)
Migrant parent		-11.073***		
		(3.427)		
Education				
Primary				25.364***
				(6.310)
Lower secondary				38.258***
				(5.472)
Upper secondary I				63.167***
				(7.005)
Upper secondary II				79.339***
				(6.518)
University				100.844***
				(6.897)
Constant	377.796***	377.565***	380.185***	311.391***
	(2.638)	(2.664)	(2.629)	(4.840)
Observations	1,032,304	1,032,304	1,038,580	1,038,580
R-squared	0.090	0.091	0.093	0.140

Note: The mean of the dependent variable is 487 points. Reference category for parental education is no education. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Panel B: Parental labor situation

Dep. Var.	Reading score			
EU member	9.138**	8.368**	9.667**	12.553***
	(3.744)	(3.506)	(3.814)	(3.752)
Work status				
Mother works full time		18.312***		
		(2.298)		
Mother works part time		16.654***		
		(2.182)		
Mother looks for work		-18.204***		
		(2.582)		
Mother has other work status		16.519***		
		(1.805)		
Father works full time		-10.071***		
		(2.311)		
Father works part time		-13.654***		
		(2.517)		
Father looks for work		-3.868		
		(4.034)		
Father has other work status		18.312***		
		(2.298)		
Occupation				
Professionals				31.789***
				(2.476)
Technicians				5.824***
				(2.075)
Clerical				-2.350
				(1.884)
Services and sales				-30.261***
				(1.315)
Skilled agriculture/ forestry/ fishing				-41.964***
				(3.728)
Craft and trade				-46.583***
				(2.024)
Plant and machine operators				-47.516***
				(1.801)
Elementary				-71.616***
				(3.117)
Constant	380.690***	370.622***	376.124***	397.156***
	(2.582)	(3.236)	(2.785)	(3.269)
Observations	1,021,615	1,021,615	1,073,652	1,073,652
R-squared	0.090	0.104	0.088	0.172

Note: Reference category for work status is other and for occupation is manager. Elementary includes cleaner, agriculture, manufacturing, food, street. The model controls for student age and gender and its imputed shares. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.9: The effect of EU membership on student achievement conditional on family characteristics

Dep. var.		Reading score		
EU member	9.404**	15.860***	11.223**	12.062***
	(3.683)	(3.280)	(4.232)	(4.391)
Family wealth				
Consumer goods		15.860***		
		(3.280)		
Cultural goods		1.648*		
		(0.838)		
Number of books		12.529***		
11-100		(0.614)		
		49.104***		
101 - 200		(2.347)		
		80.797***		
201 - 500		(3.041)		
		100.331***		
> 500		(3.435)		
		96.850***		
Home educational resources		-2.885*		
		(1.479)		
Family structure				
Single mother				-8.498***
				(1.510)
Single father				-24.845***
				(1.679)
Without parents				-60.736
				(79.144)
Constant	382.590***	346.398***	376.697***	377.805***
	(2.515)	(2.569)	(2.689)	(2.725)
Observations	1,041,450	1,041,450	1,022,428	1,022,428
R-squared	0.087	0.225	0.089	0.091

Note: The mean of the dependent variable is 487 points. The index consumer goods includea an own room, access to internet, number of phones, TVs, computers, cars, bathrooms, and DVD players. The index cultural goods includea literature books, poetry, and art work. Reference category for the number of books are 0 to 10 books. The index home educational resources contains a study desk, quiet study place, computer for school work, educational software, books for school work, technical reference books, and dictionary. Reference group for family structure is living with both parents. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.10: The effect of EU membership on student achievement conditional on school characteristics

Dep. var.	Reading score	
EU member	9.634**	9.193*
	(4.049)	(4.799)
Location		
Town (3,000-15,000)		11.030**
		(5.167)
Large town (15,000-100,000)		9.433***
		(2.956)
City (100,000-1,000,000)		16.815***
		(4.409)
Large city (>1,000,000)		23.572***
		(5.345)
Private		31.410***
		(7.867)
Number of students		26.696***
		(3.371)
Government budget		0.035***
		(0.007)
Shortage of math teachers		0.007
		(0.055)
Certificated teachers		-5.501
		(3.363)
Teacher absence		
a little		25.775***
		(5.712)
some		4.120**
		(1.659)
a lot		-3.706
		(2.814)
Constant	377.250***	318.238***
	(3.184)	(7.534)
Observations	842,420	842,420
R-squared	0.085	0.123

Note: The mean of the dependent variable is 487 points. Reference category for location is village with less than 3,000 inhabitants. Reference category for teacher absence is not at all. The model controls for the share of ipolated observations of private, number of students, and share of government budget. It also conditions on time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.11: The effect of EU membership on student achievement conditional on country characteristics

Dep. var.	Reading score	
EU member	14.493***	12.332***
	(4.076)	(3.516)
School autonomy		1.781
		(6.971)
School autonomy x initial GDP		-0.679
		(0.524)
Tests		
School-based external comparison		17.063*
		(9.898)
Student-based external comparison		11.772***
		(3.437)
Standardized monitoring		-11.051
		(14.542)
Internal testing		37.725*
		(18.913)
Internal teacher monitoring		-22.249
		(15.217)
Expenditure on secondary education		-0.507*
		(0.281)
Constant	481.188***	495.542***
	(3.994)	(16.416)
Observations	772,067	772,067
R-squared	0.054	0.055

Note: The mean of the dependent variable is 487 points. Autonomy is the share of schools with academic-content autonomy. Tests also report the share of school in a country which employ standardized student tests for the different purposes. Autonomy and tests are both derived from the PISA principal questionnaire. GDP per capita is measured in international US$ in PPP, government expenditure per secondary student is a share of GDP per capita. The model controls for the ipolated share of secondary government expenditure and for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.12: Estimation results from mediation analysis

Dep. var.	Reading score
Student characteristics	
Female	-0.038
Age	-0.010**
Migrant student	0.068**
Parental characteristics	
Migrant parent	-0.080**
Primary	-0.008**
Lower secondary	-0.002**
Upper secondary I	0.009**
Upper secondary II	0 .008**
University	-0.053**
Mother works full time	-0.005**
Mother works part time	0.060**
Mother looks for work	0.026**
Father works full time	0.057**
Father works part time	0.062**
Father looks for work	0.012**
Professionals	0.055**
Technicians	-0.073**
Clerical	-0.042**
Services and sales	-0.087**
Skilled agriculture/ forestry/ fishing	-0.012**
Craft and trade	-0.136**
Plant and machine operators	0.114**
Elementary	0.004**
Family characteristics	
Consumer goods	0.091**
Cultural goods	0.012**
11-100 books	0.054**
101-200 books	-0.115**
201-500 books	-0.051**
>500 books	0.015**
Home educational resources	0.009**
Single mother	0.124**
Single father	0.106**
Without parents	0.033**

(Continued next page.)

School characteristics	
Town (3,000-15,000)	-0.011**
Large town (15,000-100,000)	0.044**
City (100,000-1,000,000)	0.166**
Large city (>1,000,000)	-0.007**
Private	-0.014**
Number of students	-0.004**
Government budget	0.004**
Shortage of math teachers	0.108**
Certificated teachers	0.066**
Teacher absence: a little	0.186**
Teacher absence: some	0.041**
Teacher absence: a lot	-0.251**
Country characteristics	
School autonomy	0.099**
School autonomy x initial GDP	-0.021**
School-based external comparison	0.008**
Student-based external comparison	0.391**
Standardized monitoring	0.046**
Internal testing	-0.024**
Internal teacher monitoring	0.016**
Expenditure on secondary education	-0.004**

Note: The table reports the share of the mediated effect as extracted from causal mechanism analysis. Each line represents one regression, as mediation analysis tests only one mediator per regression, but the model is residualized for time- and country-fixed effects. Due to computational limitations, I run the analysis in country-level data which produce the same main results as the individual-level data (compare Table A 5.6). Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). The mean of the dependent variable is 487 points. Least squares regression weighted by students' sampling probability.

Table 5.13: Selection test – emigration

Dep. var.	Migrant ratio	Reading score	Reading score
EU entry	0.000	9.089**	9.105**
	(0.001)	(4.283)	(4.263)
Migrant ratio			-41.620***
			(9.721)
Constant	-0.000	378.400***	378.380***
	(0.000)	(2.441)	(2.436)
Observations	976,887	976,887	976,887
R-squared	0.005	0.098	0.099

Note: The emigrant ratio represents the number of first-generation emigrants from an Eastern European country tested in another country relative to the number of students in the respective home country. The average emigrant ratio is 0.001. The mean of reading score is 487 points. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.14: Placebo test with leads and lags

Dep. var.	Leading reading score	Lagged reading score
EU member	-0.059	1.270**
	(0.557)	(0.584)
Constant	377.678***	378.062***
	(0.260)	(0.350)
Observations	1,073,620	1,073,620
R-squared	0.086	0.086

Note: The mean of the dependent variable is 487 points in reading scores. Each field represents a separate regression. The dependent variable leads or lags by one period relative to the independent variable. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.15: Robustness test - sample composition

Panel A: Omitting one country at a time

Dep. var.	Reading Score										
without:	BGR	CZE	EST	HRV	HUN	LTU	LVA	POL	ROU	SVK	SVN
EU member	8.986**	11.492***	9.728**	10.074**	10.213**	9.632**	9.183**	7.499**	8.221**	10.540**	9.646**
	(4.198)	(3.531)	(3.819)	(4.176)	(4.107)	(3.815)	(4.208)	(3.556)	(3.996)	(3.915)	(3.794)
Constant	375.781***	375.936***	376.063***	376.183***	376.232***	376.091***	376.579***	376.402***	375.725***	376.290***	375.879***
	(2.830)	(2.831)	(2.794)	(2.845)	(2.836)	(2.793)	(2.796)	(2.817)	(2.824)	(2.799)	(2.780)
Observations	1,048,780	1,037,750	1,053,694	1,052,628	1,044,437	1,053,237	1,046,736	1,046,066	1,048,979	1,045,992	1,048,585
R-squared	0.081	0.091	0.088	0.089	0.090	0.089	0.090	0.089	0.079	0.089	0.089

Note: Each cell represents a new regression based on a different sample excluding the group named in line two. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Panel B: Omitting one wave at a time

Dep. var.	Reading Score					
without:	2000	2003	2006	2009	2012	2015
EU member	8.634*	14.499***	5.765	9.837**	8.820**	11.289***
	(4.323)	(4.866)	(4.205)	(4.294)	(3.604)	(4.076)
Constant	390.500***	376.808***	375.715***	370.493***	368.149***	376.255***
	(2.272)	(2.870)	(2.865)	(2.206)	(2.058)	(2.838)
Observations	955,915	954,226	876,376	854,332	847,020	880,391
R-squared	0.081	0.094	0.082	0.086	0.090	0.094

Note: Each cell represents a new regression based on a different sample excluding the group named in line two. The model controls for time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table 5.16: Specification test on country-specific time trends

Dep. var.	Reading score	Math score	Science score
EU member	5.205**	10.114*	5.742*
	(2.220)	(5.680)	(2.911)
Constant	-1,676.036***	-7,299.901***	-3,128.500***
	(0.000)	(0.000)	(0.002)
Observations	1,021,595	1,073,652	1,021,522
R-squared	0.105	0.091	0.092

Note: Sample mean of reading score is 487 points, of math score is 491 points, and of science score is 494 points. The model controls for country-specific time trends. Standard errors are clustered at the country level reported in parenthesis (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Appendix

Table A 5.1: Summary statistics

	Treatment group		Control Group	
Variable	Mean	*Std. dev.*	Mean	*Std. dev.*
Student characteristics				
Reading score	476.9	*98.65*	490.2	*101.0*
Math score	483.7	*95.02*	490.8	*96.17*
Science score	490.8	*96.53*	493.4	*100.0*
Age	15.75	*0.319*	15.75	*0.290*
Female	0.494	*0.500*	0.499	*0.500*
Migrant student	0.020	*0.138*	0.062	*0.241*
Consumer goods	-0.397	*0.993*	0.178	*0.987*
Cultural goods	0.129	*0.971*	-0.044	*1.016*
Number of books				
0-10	0.123	*0.329*	0.126	*0.332*
11-100	0.473	*0.499*	0.461	*0.498*
101 – 500	0.183	*0.387*	0.183	*0.387*
> 500	0.135	*0.342*	0.142	*0.350*
Home educational resources	-0.025	*0.759*	-0.003	*0.971*
Parental characteristics				
Migrant parent	0.094	*0.292*	0.135	0.342
Parental education				
No education	0.001	*0.0374*	0.012	*0.107*
Primary	0.004	*0.065*	0.03	*0.170*
Lower secondary	0.034	*0.182*	0.1	*0.297*
Upper secondary I	0.138	*0.345*	0.105	*0.306*
Upper secondary II	0.395	*0.489*	0.252	*0.434*
University	0.427	*0.495*	0.504	*0.500*
Parental work status				
Mother works full time	0.660	*0.361*	0.509	*0.407*
Mother works part time	0.0891	*0.214*	0.202	*0.326*
Mother looks for work	0.0916	*0.215*	0.0598	*0.188*
Mother has other work status	0.158	*0.362*	0.232	*0.472*
Father works full time	0.761	*0.323*	0.815	*0.310*
Father works part time	0.083	*0.204*	0.0739	*0.209*
Father looks for work	0.068	*0.188*	0.0414	*0.155*
Father has other work status	0.097	*0.231*	0.0939	*0.248*
Share imputed	0.423	*0.494*	0.353	*0.478*

(Continued next page.)

	Treatment group		Control Group	
Variable	Mean	*Std. dev.*	Mean	*Std. dev.*
Parental occupation				
Manager	0.147	*0.354*	0.163	*0.369*
Professionals	0.169	*0.375*	0.217	*0.413*
Technicians	0.146	*0.353*	0.137	*0.344*
Clerical	0.065	*0.246*	0.0822	*0.275*
Services and sales	0.171	*0.376*	0.154	*0.361*
Skilled agriculture/ forestry/ fishing	0.029	*0.167*	0.022	*0.147*
Craft and trade	0.135	*0.342*	0.090	*0.286*
Plant and machine operators	0.0509	*0.220*	0.043	*0.203*
Elementary	0.0570	*0.232*	0.051	*0.220*
Family structure				
Without parents	0.0267	*0.120*	0.015	*0.098*
Single mother	0.143	*0.265*	0.124	*0.262*
Single father	0.0205	*0.105*	0.0194	*0.113*
Share imputed	0.423	*0.494*	0.353	*0.478*
School characteristics				
Location				
Village (less 3,000)	0.106	*0.308*	0.0845	*0.278*
Town (3,000-15,000)	0.207	*0.405*	0.250	*0.433*
Large town (15,000-100,000)	0.371	*0.483*	0.403	*0.490*
City (100,000-1,000,000)	0.248	*0.432*	0.189	*0.392*
Large city (>1,000,000)	0.068	*0.251*	0.0584	*0.235*
Share imputed	0.020	*0.111*	0.015	*0.120*
Private	0.0426	*0.202*	0.170	*0.375*
Share imputed	0.016	*0.125*	0.01	*0.094*
Number of students	565.7	*336.3*	693.7	*459.7*
Share imputed	0.024	*0.157*	0.036	*0.186*
Share of government budget	1.976	*10.03*	7.837	*24.89*
Share imputed	0.042	*0.189*	0.047	*0.211*
Shortage of math teachers	0.0821	*0.275*	0.183	*0.386*
Fully certificated teachers	0.919	*0.196*	0.899	*0.213*
Share imputed	0.073	*0.260*	0.040	*0.197*

(Continued next page.)

	Treatment group		Control Group	
Variable	Mean	*Std. dev.*	Mean	*Std. dev.*
Teacher absence as problem				
not at all	0.427	*0.495*	0.223	*0.416*
a little	0.442	*0.497*	0.583	*0.493*
Some	0.0706	*0.256*	0.129	*0.335*
a lot	0.0102	*0.100*	0.013	*0.114*
Share imputed	0.199	*0.399*	0.171	*0.376*
Education system				
School autonomy	0.505	*0.500*	0.607	*0.363*
School autonomy x initial GDP	2.321	*3.531*	18.46	*14.37*
Tests				
School-based external comparison	0.663	*0.142*	0.515	*0.270*
Student-based external comparison	0.499	*0.410*	0.585	*0.375*
Standardized monitoring	0.741	*0.117*	0.641	*0.166*
Internal testing	0.707	*0.123*	0.662	*0.161*
Internal teacher monitoring	0.649	*0.0987*	0.387	*0.236*
Government expenditure on secondary education as share of GDP	21.88	*4.554*	25.71	*4.524*
Share imputed	0.101	*0.301*	0.072	*0.259*
N	1,073,652			

Note: The treatment group consists of the eleven Eastern European countries accessing the EU. The control group consists of countries which either are always or never members of the EU in the period 2000 to 2015.

Table A 5.2: Number of student-level observations by country

Country	2000	2003	2006	2009	2012	2015	Total of country
Albania	4,980	-	-	4,596	4,743	-	14,319
Austria	4,745	4,597	4,927	6,590	4,755	7,007	32,621
Belgium	6,670	8,796	8,857	8,501	8,597	9,651	51,072
Bulgaria	4,657	-	4,498	4,507	5,282	5,928	24,872
Croatia	-	-	5,213	4,994	5,008	5,809	21,024
Czech Republic	5,365	6,320	5,932	6,064	5,327	6,894	35,902
Denmark	4,235	4,218	4,532	5,924	7,481	7,161	33,551
Estonia	-	-	4,865	4,727	4,779	5,587	19,958
Finland	4,864	5,796	4,714	5,810	8,829	5,882	35,895
France	4,673	-	-	-	4,613	6,108	15,394
Germany	5,073	4,660	4,891	4,979	5,001	6,504	31,108
Greece	4,672	4,627	4,873	4,969	5,125	5,532	29,798
Hungary	4,887	4,765	4,490	4,605	4,810	5,658	29,215
Iceland	3,372	3,350	3,789	3,646	3,508	3,371	21,036
Ireland	3,854	3,880	4,585	3,937	5,016	5,741	27,013
Italy	4,984	11,639	21,773	30,905	31,073	11,583	111,957
Latvia	3,893	4,627	4,719	4,502	4,306	4,869	26,916
Lithuania	-	-	4,744	4,528	4,618	6,525	20,415
Luxembourg	3,528	3,923	4,567	4,622	5,258	5,299	27,197
Montenegro	-	-	4,455	4,825	4,744	5,665	19,689
The Netherlands	-	3,992	4,871	4,760	4,460	5,385	23,468
Norway	4,147	4,064	4,692	4,660	4,686	5,456	27,705
Poland	3,654	4,383	5,547	4,917	4,607	4,478	27,586
Portugal	4,585	4,608	5,109	6,298	5,722	7,325	33,647
Romania	4,829	-	5,118	4,776	5,074	4,876	24,673
Serbia	-	-	4,798	5,523	4,684	-	15,005
Slovak Republic	-	7,346	4,731	4,555	4,678	6,350	27,660
Slovenia	-	-	6,595	6,155	5,911	6,406	25,067
Spain	6,214	10,791	19,604	25,887	25,313	6,736	94,545
Sweden	4,416	4,624	4,443	4,567	4,736	5,458	28,244
Switzerland	6,100	8,420	12,192	11,812	11,229	5,860	55,613
United Kingdom	9,340	-	13,152	12,179	12,659	14,157	61,487
Total of year	117,737	119,426	197,276	219,320	226,632	193,261	1,073,652

Note: Table shows the number of students per country and per year. "-" signifies that the country did not participate in PISA the given year or that data was officially unusable.

Table A 5.3: Mechanisms - entire set of mechanisms

Dep. var.	Reading score	
EU member	14.298**	15.736***
	(5.532)	(3.869)
Female student		33.903***
		(1.377)
Age student		12.710***
		(0.926)
Migrant student		-30.298***
		(3.141)
Migrant parent		-12.506***
		(2.207)
Parental education		
Primary		19.085***
		(3.355)
Lower secondary		20.891***
		(3.706)
Upper secondary I		31.108***
		(5.170)
Upper secondary II		35.877***
		(4.709)
University		37.174***
		(5.113)
Work status		
Mother works full time		2.030*
		(1.157)
Mother works part time		2.569
		(1.832)
Mother looks for work		-12.106***
		(1.512)
Father works full time		3.567***
		(1.179)
Father works part time		-14.253***
		(1.864)
Father looks for work		-6.642***
		(1.589)
Occupation		
Professionals		9.339***
		(0.896)
Technicians		-3.684***
		(1.024)
Clerical		-8.182***
		(1.231)
Services and sales		-23.360***
		(1.323)
Skilled agriculture/ forestry/ fishing		-22.089***
		(2.920)
Craft and trade		-31.164***
		(1.333)
Plant and machine operators		-30.065***
		(1.975)
Elementary		-47.564***
		(2.340)

(Continued next page.)

Family wealth	
Consumer goods	-2.337***
	(0.745)
Cultural goods	8.264***
	(0.460)
Number of books	
11-100	36.480***
	(2.119)
101 - 200	58.921***
	(2.818)
201 - 500	73.816***
	(3.092)
> 500	74.463***
	(3.423)
Home educational resources	-1.212
	(2.289)
Family structure	
Single mother	-1.838*
	(1.044)
Single father	-8.151***
	(1.408)
Without parents	-252.828
	(169.092)
Location	
Town (3,000-15,000)	3.509*
	(1.818)
Large town (15,000-100,000)	7.386***
	(2.326)
City (100,000-1,000,000)	9.961***
	(2.796)
Large city (>1,000,000)	12.312***
	(3.471)
Private	10.419***
	(2.563)
Number of students	0.023***
	(0.004)
Government budget	-0.009
	(0.028)
Shortage of math teachers	-5.168**
	(2.312)
Certificated teachers	16.586***
	(4.245)
Teacher absence as problem	
a little	2.410**
	(1.142)
some	-2.980
	(1.907)
a lot	-0.975
	(3.242)

(Continued next page.)

School autonomy		9.775
		(7.176)
School autonomy x initial GDP		-0.672*
		(0.387)
Tests		
School-based external comparison		19.127
		(11.314)
Student-based external comparison		12.237***
		(3.035)
Standardized monitoring		-23.309
		(13.917)
Internal testing		49.487***
		(16.759)
Internal teacher monitoring		-19.557
		(14.887)
Expenditure on secondary education		-0.541**
		(2.220)
Constant	487.863***	148.992***
	(5.478)	(28.709)
Observations	527,198	527,198
R-squared	0.065	0.298

Note: The mean of the dependent variable is 487 points. Reference category for parental education is no education. Reference category for work status is other and for occupation is manager. Elementary includes cleaner, agriculture, manufacturing, food, street. The index consumer goods include an own room, access to internet, number of phones, TVs, computers, cars, bathrooms, and DVD players. The index cultural goods include literature books, poetry, and art work. Reference category for the number of books are 0 to 10 books. The index home educational resources contain study desk, quiet study place, computer for school work, educational software, books for school work, technical reference books, and dictionary. Reference group for family structure is living with both parents. Reference category for location is village with less than 3,000 inhabitants. Reference category for teacher absence is not at all. The model controls for the share of ipolated observations of private, number of students, and share of government budget. Autonomy is the share of schools with academic-content autonomy. Tests also report the share of school in a country which employ standardized student tests for the different purposes. Government expenditure per secondary student is a share of GDP per capita. The model controls for the ipolated share of secondary government expenditure. The model also conditions on time- and country-fixed effects. Standard errors are clustered at country level (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability.

Table A 5.4: EU membership status by country

Country	2000	2003	2006	2009	2012	2015	Total of years per country
Albania	0	0	0	0	0	0	0
Austria	1	1	1	1	1	1	6
Belgium	1	1	1	1	1	1	6
Bulgaria	0	0	0	1	1	1	3
Croatia	0	0	0	0	0	1	1
Czech Republic	0	0	1	1	1	1	4
Denmark	1	1	1	1	1	1	6
Estonia	0	0	1	1	1	1	4
Finland	1	1	1	1	1	1	6
France	1	1	1	1	1	1	6
Germany	1	1	1	1	1	1	6
Great Britain	1	1	1	1	1	1	6
Greece	1	1	1	1	1	1	6
Hungary	0	0	1	1	1	1	4
Iceland	0	0	0	0	0	0	0
Ireland	1	1	1	1	1	1	6
Italy	1	1	1	1	1	1	6
Latvia	0	0	1	1	1	1	4
Lithuania	0	0	1	1	1	1	4
Luxembourg	1	1	1	1	1	1	6
Montenegro	0	0	0	0	0	0	0
Netherlands	1	1	1	1	1	1	6
Norway	0	0	0	0	0	0	0
Poland	0	0	1	1	1	1	4
Portugal	1	1	1	1	1	1	6
Romania	0	0	0	1	1	1	3
Serbia	0	0	0	0	0	0	0
Slovak Republic	0	0	1	1	1	1	4
Slovenia	0	0	1	1	1	1	4
Spain	1	1	1	1	1	1	6
Sweden	1	1	1	1	1	1	6
Switzerland	0	0	0	0	0	0	0
Total of countries per year	15	15	23	25	25	26	-

Note: 1 signifies EU membership, 0 signifies no EU membership.

Table A 5.5: Individual characteristics by family structure

	Both parents				Single parent				Without parents			
	before EU entry		after EU entry		before EU entry		after EU entry		before EU entry		after EU entry	
Variable	Mean	*Std. Dev.*	Mean	*Std. Dev.*	Mean	*Std. Dev.*	Mean	*Std. Dev.*	Mean	*Std. Dev.*	Mean	*Std. Dev.*
Reading score	474.88	*100.67*	499.68	*93.75*	478.30	*97.40*	490.81	*95.63*	453.12	*107.33*	486.22	*100.79*
Age	15.73	*0.34*	15.77	*0.29*	15.75	*0.33*	15.76	*0.29*	15.77	*0.29*	15.78	*0.29*
Female	0.50	*0.50*	0.50	*0.50*	0.51	*0.50*	0.52	*0.50*	0.48	*0.50*	0.49	*0.50*
Consumer goods	-0.20	*1.07*	0.14	*0.85*	-0.34	*1.05*	-0.21	*0.88*	-0.14	*1.03*	-0.02	*0.89*
Cultural goods	0.18	*0.98*	0.07	*0.98*	0.09	*1.00*	-0.08	*1.00*	0.14	*0.99*	0.04	*0.97*
Home educational resources	-0.058	*0.002*	-0.027	*0.728*	-0.059	*0.002*	-0.026	*0.748*	-0.059	*0.002*	0.033	*1.256*
Books	3.63	*1.59*	3.54	*1.47*	3.71	*1.56*	3.34	*1.48*	3.25	*1.48*	3.32	*1.44*
Parental education	4.55	*1.28*	4.44	*1.46*	4.50	*1.28*	4.39	*1.47*	4.51	*1.38*	4.46	*1.46*
Parental occupation	6,055.91	*2,649.18*	6,325.452	*2,660.03*	5,852.52	*2,613.94*	6,242.49	*2,638.18*	6,364.75	*2,628.40*	6,515.72	*2,612.23*
Maternal work status	2.02	*1.23*	2.01	*1.53*	1.55	*1.02*	1.55	*1.01*	1.69	*1.09*	1.59	*1.02*
Paternal work status	1.42	*0.91*	1.32	*0.80*	1.85	*1.14*	1.84	*1.13*	2.28	*1.29*	2.18	*1.27*

Note: The table shows descriptives of key individual characteristics across the three family structures. Consumer and cultural goods and home educational resources are standardized to mean zero and standard deviation of one. Books is the number of books at home. Parental education is classified in ISCED coding ranging from to zero to six with lower values representing lower education levels. Parental occupation is classified in ISCO codes between 1,000 and 9,996 with lower values representing higher occupational status. Maternal and paternal work status is coded in four categories with higher values representing less working time, i.e., lower work status.

Table A 5.6: Pairwise correlations of parental involvement

	Reading score	Consumer goods	Cultural goods	Edu. resources	Books
Academic interest	0.154	0.020	0.310	0.121	0.230
(N =115,335 in 2000)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*
Study time	-0.131	-0.030	0.047	0.049	0.001
(N = 219,341 in 2012)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.685)*
Talks about school	0.089	0.041	0.049	0.094	0.036
(N = 171,468 in 2015)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*
Emotional support	0.114	0.146	0.144	0.214	0.109
(N = 182,725 in 2015)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*
Learning support	-0.022	0.023	0.132	0.125	0.077
(N = 49,246 in 2015)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*

Note: The table reports the correlation coefficients of student achievement and family wealth with measures of parental involvement with their children. P-values are reported in parenthesis below. Parents' academic interest in their children is a WLE index ranging between -2.2 and 2.72 with higher values representing higher interest. Time parents study with their children ranges from zero to 30. Talking about school takes values of 0 or 1 representing the answer option "no" and "yes". Emotional support is an index ranging between -3.1 and 1.1, and learning support is an index ranging between -5.8 and 3.7 with higher values representing higher support. The indices consumer goods, cultural good, educational resources and books were standardized to mean zero and standard deviation of one.

Table A 5.7: Main results on country-wave level

Dep. var.	Reading score	
Aggregation level:	Individual	Country
EU member	9.667**	9.667**
	(3.814)	(4.323)
Constant	376.124***	376.124***
	(2.785)	(3.157)
Observations	1,073,652	168
R-squared	0.088	0.915

Note: This table shows regression results from the main specification in the individual-level data (Column 1) and in the country-level data (column 2). This table provides additional information to Table 5.12. Sample mean of reading score is 487 points, the model controls for time- and country-fixed effects. Standard errors are clustered at the country level reported in parenthesis (*** p<0.01, ** p<0.05, * p<0.1). Least squares regression weighted by students' sampling probability in individual data.

Table A 5.8: Difference between emigrants and their home population

	Parental Educational Attainment				Parental Occupation				Student Test Score in Reading			
	Home population		Emigrant population		Home population		Emigrant population		Home population		Emigrant population	
	before EU	after EU	before EU	after EU	before EU	after EU	before EU	after EU	before EU	after EU	before EU	after EU
BGR	4.9	4.7	.	4.6	5,976.0	6,569.2	.	8,145.7	415.9	435.4	.	455.2
CZE	4.6	4.5	4.6	5.4	5,324.2	6,289.8	5,540.1	6,419.2	501.8	503.8	470.8	433.9
EST	.	4.9	.	4.9		6,014.6	.	6,940.5	.	511.5	.	454.3
HRV	4.5	4.7	4.7	4.0	6,794.2	6,706.9	6,511.3	7,689.3	478.9	487.7	433.4	418.9
HUN	4.5	4.5	.	4.9	5,988.2	6,773.6	.	6,350.8	481.0	489.6	.	428.5
LTU	.	5.0	.	3.5		6,129.1	.	9,132.0	.	469.9	.	371.3
LVA	5.1	4.9	.	.	5,664.3	6,711.0	.		478.6	489.6	.	.
POL	4.4	4.1	.	4.5	5,831.8	6,633.0	.	7,100.2	482.9	511.6	.	462.4
ROU	4.7	4.6	.	4.6	6,795.9	7,508.5	.	6,874.7	420.0	432.7	.	433.8
SVK	4.3	4.4	4.6	4.5	6,530.0	6,959.2	6,250.2	6,342.5	475.2	466.7	528.1	468.9
SVN	.	4.2	.	4.2	.	6,683.7	.	7,793.6	.	468.8	.	488.7

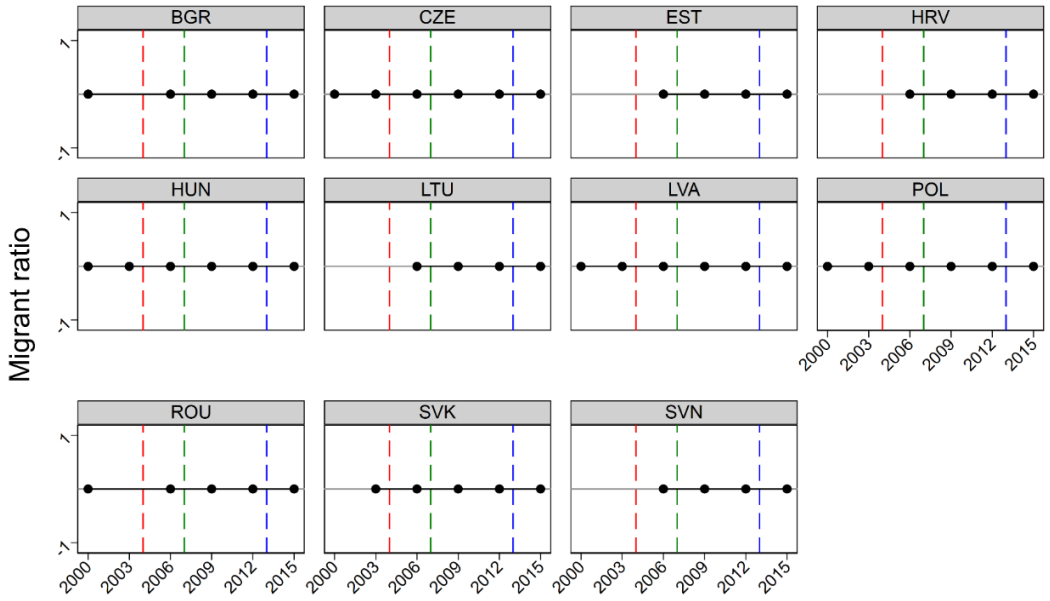
Note: This table shows descriptive statistics of the home population in the eleven Eastern European entrant countries compared to emigrants from those countries not living in their home country anymore. The country means are displayed. Parental education is classified in ISCED coding ranging from to zero to six with lower values representing lower education levels. Parental occupation is classified in ISCO codes with lower values representing higher occupational status. Student test score has a mean of 500 points with a standard deviation of 100 points. Missing data is due to Eastern European countries joining PISA wave by wave. Country abbreviations stand for: Bulgaria (BGR), the Czech Republic (CZE), Estonia (EST), Croatia (HRV), Hungary (HUN), Lithuania (LTU), Latvia (LVA), Poland (POL), Romania (ROU), the Slovak Republic (SVK), and Slovenia (SVN).

Table A 5.9: Summary statistics on emigrants and natives by home country

Nationality	Country of Residence	N home country	N host country	Emigrant ratio	SD
Bulgaria	Greece, the Netherlands	24,872	11	0.000	*0.021*
Croatia	Austria, Germany, Montenegro	21,024	150	0.007	*0.084*
Czech Rep.	Austria, Slovak Rep.	35,902	51	0.001	*0.038*
Estonia	Finland, Ireland	19,958	175	0.009	*0.093*
Hungary	Austria, Slovak Rep., Slovenia	29,215	54	0.002	*0.043*
Lithuania	Ireland	20,415	3	0.000	*0.012*
Latvia	-	26,916	0	0.000	*0.000*
Poland	Austria, Germany, United Kingdom, Ireland, the Netherlands	27,586	330	0.012	*0.108*
Romania	Austria, Ireland	24,673	96	0.004	*0.062*
Slovak Rep.	Austria, Czech Rep.	27,660	194	0.007	*0.083*
Slovenia	Austria, Germany	25,067	14	0.001	*0.024*

Note: The table shows origin, destination, and number of first generation migrants in the sample.

Figure A 5.1: Evolution of the migrant ratio



Note: The migrant ratio relates the number of first generation emigrants from Eastern Europe living in another country to the number of students in each Eastern European country. The red, dashed, vertical line signals the 2004 entries. The green, dashed, vertical line indicates the 2007 entries. The blue, dashed, vertical line designates the 2013 entry. Country abbreviations stand for: Bulgaria (BGR), the Czech Republic (CZE), Estonia (EST), Croatia (HRV), Hungary (HUN), Lithuania (LTU), Latvia (LVA), Poland (POL), Romania (ROU), the Slovak Republic (SVK), and Slovenia (SVN).

Bibliography

Acemoglu, D & Dell, M 2010, 'Productivity Differences Between and Within Countries', *American Economic Journal: Macroeconomics*, vol. 2, no. 1, pp. 169–188.

Acemoglu, D, Johnson, S & Robinson, JA 2001, 'The Colonial Origins of Comparative Development: An Empirical Investigation', *The American Eonomic Review*, vol. 91, no. 5, pp. 1369–1401.

Aghion, P & Howitt, P 1998, *Endogenous Growth Theory,* MIT Press, Cambridge, Mass.

Altonji, JG, Elder, TE & Taber, CR 2005, 'Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools', *Journal of Political Economy*, vol. 113, no. 1, pp. 151–184.

Anderson, RC 1962, 'The Case for Teacher Specialization in the Elementary School', *The Elementary School Journal*, vol. 62, no. 5, pp. 253–260.

Andrabi, T, Das, J & Khwaja, AI 2017, 'Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets', *The American economic review*, vol. 107, no. 6, pp. 1535–1563.

Andrews, P & coauthors 2014, *OECD and Pisa tests are damaging education worldwide*, The Guardian. Available from: https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics [20 June 2018].

Angrist, JD & Lavy, V 1999, 'Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement', *The Quarterly Journal of Economics*, vol. 114, no. 2, pp. 533–575.

Angrist, JD, Lavy, V, Leder-Luis, J & Shany, A 2017, 'Maimonides Rule Redux', *NBER Working Paper*, vol. 23486.

Angrist, JD & Pischke, J-S 2009, *Mostly harmless econometrics*. *An empiricist's companion,* Princeton University Press, Princeton.

Anniste, K, Tammaru, T, Pungas, E & Paas, T 2012, *Emigration after EU Enlargement: Was There a Brain Drain Effect in the Case of Estonia?*

Araujo, P de & Lagos, S 2013, 'Self-esteem, education, and wages revisited', *Journal of Economic Psychology*, vol. 34, pp. 120–132.

Balart, P, Oosterveen, M & Webbink, D 2018, 'Test scores, noncognitive skills and economic growth', *Economics of Education Review*, vol. 63, pp. 134–153.

Baldwin, RE & Sheghezza, E 1996, 'Growth and European Integration: Towards an Empirical Assessment', *CEPR Discussion Paper*, no. 1393.

Baldwin, RE & Wyplosz, C 2012, *The economics of European integration,* McGraw-Hill Education Higher Education, London.

Banerji, R, Berry, J & Shotland, M 2013, *The Impact of Mother Literacy and Participation Programs on Child Learning: Evidence from a Randomized Evaluation in India,* Cambridge, MA.

Barrios, T, Diamond, R, Imbens, GW & Kolesár, M 2012, 'Clustering, Spatial Correlations, and Randomization Inference', *Journal of the American Statistical Association*, vol. 107, no. 498, pp. 578–591.

Baumert, J & Demmrich, A 2001, 'Test motivation in the assessment of student skills: The effects of incentives on motivation and performance', *European Journal of Psychology of Education*, vol. 16, no. 3, pp. 441–462.

Bazzi, S, Gaduh, A, Rothenberg, AD & Wong, M 2016, 'Skill Transferability, Migration, and Development. Evidence from Population Resettlement in Indonesia', *The American Eonomic Review*, vol. 106, no. 9, pp. 2658–2698.

Benhabib, J & Spiegel, MM 1994, 'The role of human capital in economic development evidence from aggregate cross-country data', *Journal of Monetary Economics*, vol. 34, no. 2, pp. 143–173.

Bergbauer, A, Hanushek, E & Woessmann, L 2018, 'Testing', *NBER Working Paper*, vol. 24,836, July.

Bergman, P 2015, 'Parent-child information frictions and human capital investment. Evidence from a field experiment', *CESifo Working Paper Series*, vol. 5,391, June 2015.

Bertrand, M 2011, 'New Perspectives on Gender', *Handbook of Labor Economics*, 4b, pp. 1543–1590.

Bietenbeck, J 2014, 'Teaching practices and cognitive skills', *Labour Economics*, vol. 30, pp. 143–153.

Bietenbeck, J, Piopiunik, M & Wiederhold, S 2018, 'Africa's Skill Tragedy. Does Teachers' Lack of Knowledge Lead to Low Student Performance?', *Journal of Human Resources*, vol. 53, no. 3, pp. 553–578.

Bils, M & Klenow, PJ 2000, 'Does Schooling Cause Growth?', *The American Eonomic Review*, vol. 90, no. 5, pp. 1160–1183.

Bishop, J 2006, *Chapter 15 Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn – Externalities, Information Problems and Peer Pressure*.

Bishop, JH 1997, 'The Effect of National Standards and Curriculum-Based Exams on Achievement', *The American Eonomic Review*, vol. 87, no. 2, pp. 260–264.

Bishop, JH & Wößmann, L 2004, 'Institutional Effects in a Simple Model of Educational Production', *Education Economics*, vol. 12, no. 1, pp. 17–38.

Bitzer, J & Gören, E 2018, *Foreign Aid and Subnational Development: A Grid Cell Analysis,* Department of Economics, University of Oldenburg, V-407-18.

Björklund, A, Ginther, DK & Sundström, M 2007, 'Family structure and child outcomes in the USA and Sweden', *Journal of population economics*, vol. 20, no. 1, pp. 183–201.

Björklund, A & Sundström, M 2006, 'Parental Separation and Children's Educational Attainment. A Siblings Analysis on Swedish Register Data', *Economica*, vol. 73, no. 292, pp. 605–624.

Bold, T, Filmer, D, Martin, G, Molina, E, Stacy, B, Rockmore, C, Svensson, J & Wane, W 2017, 'Enrollment without Learning. Teacher Effort, Knowledge, and Skill in Primary Schools in Africa', *Journal of Economic Perspectives*, vol. 31, no. 4, pp. 185–204.

Borghans, L & Schils, T 2012, *The leaning tower of Pisa: Decomposing achievement test scores into cognitive and noncognitive components*.

Botezat, A & Pfeiffer, F 2014, *The Impact of Parents Migration on the Well-Being of Children Left Behind Initial Evidence from Romania*.

bpb 2016, *Fünf Jahre Arbeitnehmerfreizügigkeit in Deutschland*. Available from: http://www.bpb.de/politik/hintergrund-aktuell/226107/arbeitnehmerfreizuegigkeit, http://www.bamf.de/DE/Infothek/Statistiken/Wanderungsmonitor/Freizuegigkeit/freizuegigkeit-node.html.

Bruederle, A & Hodler, R 2017, 'Nighttime Lights as a Proxy for Human Development at the Local Level', *CESifo Working Paper*, no. 6555.

Buddin, R & Zamarro, G 2009, 'Teacher qualifications and student achievement in urban elementary schools', *Journal of Urban Economics*, vol. 66, no. 2, pp. 103–115.

Bulman, G, Fairlie, R, Goodman, S & Isen, A 2017, *Parental Resources and College Attendance: Evidence from Lottery Wins*.

Burgess, S, Wilson, D & Worth, J 2013, 'A natural experiment in school accountability: The impact of school performance information on pupil progress', *Journal of Public Economics*, vol. 106, pp. 57–67.

Busby, J, Smith, TG, Krishnan, N & Bekalo, M 2013, *Climate Security Vulnerability Model, Version 3.0*. Available from: https://www.strausscenter.org/ccaps-content/climate-vulnerability-model.html [07 November 2018].

Bütikofer, A, Dalla-Zuanna, A & Salvanes, KG 2017, *Breaking the Links. Natural Resource Booms and Intergenerational Mobility,* Munich.

Cameron, AC, Gelbach, JB & Miller, DL 2008, 'Bootstrap-Based Improvements for Inference with Clustered Errors', *Review of Economics and Statistics*, vol. 90, no. 3, pp. 414–427.

Campos, NF, Coricelli, F & Moretti, L 2014, *Economic Growth and Political Integration: Estimating the Benefits from Membership in the European Union Using the Synthetic Counterfactuals Method*.

Card, D 1999, 'The causal effect of education on earnings', *Handbook of Labor Economics*, 3A, pp. 1801–1863.

Carneiro, P & Heckman, JJ 2002, 'The Evidence on Credit Constraints in Post-Secondary Schooling', *The Economic Journal*, vol. 112, no. 482, pp. 705–734.

Carnoy, M & Loeb, S 2002, 'Does External Accountability Affect Student Outcomes? A Cross-State Analysis', *Educational Evaluation and Policy Analysis*, vol. 24, no. 4, pp. 305–331.

Castelló-Climent, A, Chaudhary, L & Mukhopadhyay, A 2017, 'Higher Education and Prosperity. From Catholic Missionaries to Luminosity in India', *The Economic Journal*, vol. 2, no. 1, p. 169.

Center for International Earth Science Information Network (CIESIN) 2017, *Gridded Population of the World, Version 4 (GPWv4): Administrative Unit Center Points with Population Estimates, Revision 10,* Palisades, NY. Available from: https://doi.org/10.7927/H46H4FCT [09 March 2018].

Cervellati, M, Esposito, E & Sunde, U 2017, 'Long Term Exposure to Malaria and Development. Disaggregate Evidence for Contemporaneous Africa', *Disaggregate Evidence for Contemporaneous Africa*, vol. 83, no. 1, pp. 129–148.

Chen, X & Nordhaus, WD 2011, 'Using luminosity data as a proxy for economic statistics', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 21, pp. 8589–8594.

Chetty, R, Friedman, JN & Rockoff, JE 2017, 'Measuring the Impacts of Teachers: Reply', *The American economic review*, vol. 107, no. 6, pp. 1685–1717.

CIA World Factbook 2017, *GDP -- Composition, by sector of origin*. Available from: https://www.cia.gov/library/publications/the-world-factbook/fields/214.html [14 March 2019].

CIESIN 2016, *Africa Continental Population Datasets (2000 - 2020)*, Center for International Earth Science Information Network - CIESIN - Columbia University. Available from: http://www.worldpop.org.uk/data/summary/?doi=10.5258/SOTON/WP00004.

Clionadh, R, Linke, A, Hegre, H & Karlsen, J 2010, 'Introducing ACLED-Armed Conflict Location and Event Data', *Journal of Peace*, vol. 47, no. 5, pp. 651–660.

Clotfelter, CT, Ladd, HF & Vigdor, JL 2010, 'Teacher Credentials and Student Achievement in High School', *Journal of Human Resources*, vol. 45, no. 3, pp. 655–681.

Cohen, D & Soto, M 2007, 'Growth and human capital: good data, good results', *Journal of Economic Growth*, vol. 12, no. 1, pp. 51–76.

Condie, S, Lefgren, L & Sims, D 2014, 'Teacher heterogeneity, value-added and education policy', *Economics of Education Review*, vol. 40, pp. 76–92.

Crespo Cuaresma, J, Ritzberger-Grünwald, D & Silgoner, MA 2008, 'Growth, Convergence and EU Membership', *Applied Economics*, vol. 40, no. 5, pp. 643–656.

Cunha, F & Heckman, J 2007, 'The Technology of Skill Formation', *American Economic Review*, vol. 97, no. 2, pp. 31–47.

Dahl, GB & Lochner, L 2012, 'The Impact of Family Income on Child Achievement. Evidence from the Earned Income Tax Credit', *The American Eonomic Review*, vol. 102, no. 5, pp. 1927–1956.

Darvas, Z 2018, *What is the financial balance of EU membership for central Europe?*, Bruegel. Available from: http://bruegel.org/2018/02/what-is-the-financial-balance-of-eu-membership-for-central-europe/ [21 January 2019].

Dee, T 2007, 'Teachers and the Gender Gaps in Student Achievement', *Journal of Human Resources*, vol. 42, no. 3.

Dee, TS 2005, 'A Teacher like Me: Does Race, Ethnicity, or Gender Matter?', *The American Eonomic Review*, vol. 95, no. 2, pp. 158–165.

Dee, TS & Jacob, B 2011, 'The impact of no Child Left Behind on student achievement', *Journal of Policy Analysis and Management*, vol. 30, no. 3, pp. 418–446.

Deere, D & Strayer, W 2001, 'Putting schools to the test: School accountability, incentives, and behavior', *Working Paper*, vol. 113.

DeMars, CE & Wise, SL 2010, 'Can Differential Rapid-Guessing Behavior Lead to Differential Item Functioning?', *International Journal of Testing*, vol. 10, no. 3, pp. 207–229.

Deming, DJ, Cohodes, S, Jennings, J & Jencks, C 2016, 'School Accountability, Postsecondary Attainment, and Earnings', *Review of Economics and Statistics*, vol. 98, no. 5, pp. 848–862.

Dixit, A 2002, 'Incentives and Organizations in the Public Sector: An Interpretative Review', *Journal of Human Resources*, vol. 37, no. 4, p. 696.

Dobson, JR 2009, 'Labour mobility and migration within the EU following the 2004 Central and East European enlargement', *Employee Relations*, vol. 31, no. 2, pp. 121–138.

Donald, SG & Lang, K 2007, 'Inference with Difference-in-Differences and Other Panel Data', *Review of Economics and Statistics*, vol. 89, no. 2, pp. 221–233.

Duflo, E, Hanna, R & Ryan, SP 2012, 'Incentives Work. Getting Teachers to Come to School', *The American Eonomic Review*, vol. 102, no. 4, pp. 1241–1278.

Dustmann, C, Ku, H & Kwak, DW 2018, 'Why Are Single-Sex Schools Successful?', *Labour Economics*, vol. 54, pp. 79–99.

Easterly, W & Levine, R 1997, 'Africa's Growth Tragedy: Policies and Ethnic Divisions', *The Quarterly Journal of Economics*, vol. 112, no. 4, pp. 1203–1250.

Easterly, W & Levine, R 2003, 'Tropics, germs, and crops: how endowments influence economic development', *Journal of Monetary Economics*, vol. 50, no. 1, pp. 3–39.

Eklöf, H 2010, 'Skill and will: test-taking motivation and assessment quality', *Assessment in Education: Principles, Policy & Practice*, vol. 17, no. 4, pp. 345–356.

Elvidge, C, Ziskin, D, Baugh, K, Tuttle, B, Ghosh, T, Pack, D, Erwin, E & Zhizhin, M 2009, 'A Fifteen Year Record of Global Natural Gas Flaring Derived from Satellite Data', *Energies*, vol. 2, no. 4, pp. 595–622.

Enzi, B 2017, 'The Effect of Pre-Service Cognitive and Pedagogical Teacher Skills on Student Achievement Gains: Evidence from German Entry Screening Exams', *ifo Working Paper Series*, vol. 234.

Ermisch, JF & Francesconi, M 2001, 'Family structure and children's achievements', *Journal of population economics*, vol. 14, no. 2, pp. 249–270.

Eros, JM & Candelario-Quintana, L 2006, *Mineral Facilities of Africa and the Middle East*. Available from: https://pubs.usgs.gov/of/2006/1135/ [07 November 2018].

European Commission 2013, *European Social Fund*. Available from: http://ec.europa.eu/esf/main.jsp?catId=51&langId=en [26 February 2019].

European Council 2017, *Setting the EU's Political Agenda*. Available from: https://www.consilium.europa.eu/en/european-council/role-setting-eu-political-agenda/ [26 February 2019].

European Union 1998-2019, *EUR-Lex: Open method of coordination*. Available from: https://eur-lex.europa.eu/summary/glossary/open_method_coordination.html [14 March 2019].

Eurostat 2018a, *Annual net earnings [earn_nt_net]*, European Union. Available from: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=earn_nt_net&lang=en [03 December 2018].

Eurostat 2018b, *Employment and activity - annual data*. *From 15 to 64 years*. Percentage of total population. Available from: https://ec.europa.eu/eurostat/data/database.

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

Falck, O, Mang, C & Woessmann, L 2018, 'Virtually No Effect? Different Uses of Classroom Computers and their Effect on Student Achievement', *Oxford Bulletin of Economics and Statistics*, vol. 80, no. 1, pp. 1–38.

Falk, A, Becker, A, Dohmen, T, Enke, B, Huffman, D & Sunde, U 2018, 'Global Evidence on Economic Preferences*', *The Quarterly Journal of Economics*, vol. 133, no. 4, pp. 1645–1692.

Farchy, E 2009, *The Impact Of EU Accession On Human Capital Formation*. *Can Migration Fuel A Brain Gain?,* The World Bank.

Figlio, D & Getzler, L 2009, 'Accountability, ability and disability: Gaming the system? Improving school accountability: Check-ups or choice', vol. 14.

Figlio, D & Loeb, S 2011, 'School accountability.' in *Handbook of the Economics of Education,* eds EA Hanushek, Eric A., S Machin & L Woessmann, North Holland, San Diego, CA, USA.

Figlio, DN & Winicki, J 2005, 'Food for thought: the effects of school accountability plans on school nutrition', *Journal of Public Economics*, vol. 89, 2-3, pp. 381–394.

Fleisher, BM, Sabirianova, K & Wang, X 2005, 'Returns to skills and the speed of reforms. Evidence from Central and Eastern Europe, China, and Russia', *Journal of Comparative Economics*, vol. 33, no. 2, pp. 351–370.

Fraja, G de, Oliveira, T & Zanchi, L 2010, 'Must Try Harder: Evaluating the Role of Effort in Educational Attainment', *Review of Economics and Statistics*, vol. 92, no. 3, pp. 577–597.

Francesconi, M, Jenkins, SP & Siedler, T 2010, 'Childhood family structure and schooling outcomes. Evidence for Germany', *Journal of population economics*, vol. 23, no. 3, pp. 1073–1103.

Frisch, R & Waugh, FV 1933, 'Partial Time Regressions as Compared with Individual Trends', *Econometrica*, vol. 1, no. 4, p. 387.

Fryer, RG 2018, 'The “Pupil” Factory: Specialization and the Production of Human Capital in Schools', *The American Eonomic Review*, vol. 108, no. 3, pp. 616–656.

Fryer, RG & Levitt, SD 2010, 'An Empirical Analysis of the Gender Gap in Mathematics', *American Economic Journal: Applied Economics*, vol. 2, no. 2, pp. 210–240.

GADM 2016, *Database of Global Administrative Areas*. Available from: https://gadm.org/maps.html.

Gallup, JL, Sachs, JD & Mellinger, AD 1999, 'Geography and Economic Development', *International Regional Science Review*, vol. 22, no. 2, pp. 197–232.

Ganimian, A & Murnane, R 2016, 'Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Impact Evaluations', *NBER Working Paper*, vol. 20284.

Gennaioli, N, La Porta, R, Lopez-de-Silanes, F & Shleifer, A 2013, 'Human Capital and Regional Development *', *The Quarterly Journal of Economics*, vol. 128, no. 1, pp. 105–164.

Gerretson, H, Bosnick, J & Schofield, K 2008, 'A Case for Content Specialists as the Elementary Classroom Teacher', *The Teacher Educator*, vol. 43, no. 4, pp. 302–314.

Gershman, B & Rivera, D 2018, 'Subnational diversity in Sub-Saharan Africa. Insights from a new dataset', *Journal of Development Economics*, vol. 133, pp. 231–263.

Ghosh, T, Powell, RL, Elvidge, CD & Baugh, KE 2010, 'Shedding Light on the Global Distribution of Economic Activity', *The Open Geography Journal*, vol. 3, pp. 147–160.

Glewwe, P, Maïga, E & Zheng, H 2014, 'The Contribution of Education to Economic Growth: A Review of the Evidence, with Special Attention and an Application to Sub-Saharan Africa', *World Development*, vol. 59, pp. 379–393.

Gneezy, U, Leonard, KL & List, JA 2009, 'Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society', *Econometrica*, vol. 77, no. 5, pp. 1637–1664.

Gneezy, U, List, JA, Livingston, JA, Sadoff, S, Qin, X & XU, Y 2017, 'Measuring success in education: The role of effort on the test itself', *NBER Working Paper*, vol. 24004.

Gneezy, U, Niederle, M & Rustichini, A 2003, 'Performance in Competitive Environments: Gender Differences', *Quarterly Journal of Economics*, vol. 118, no. 3, pp. 1049–1074.

Gruber, J 2004, 'Is Making Divorce Easier Bad for Children? The Long-Run Implications of Unilateral Divorce', *Journal of Labor Economics*, vol. 22, no. 4, pp. 799–833.

Guiso, L, Monte, F, Sapienza, P & Zingales, L 2008, 'Diversity. Culture, gender, and math', *Science (New York, N.Y.)*, vol. 320, no. 5880, pp. 1164–1165.

Halász, G 2015, 'Education and Social Transformation in Central and Eastern Europe', *European Journal of Education*, vol. 50, no. 3, pp. 350–371.

Hanushek, E 1970, *The Production of Education, Teacher Quality, and Efficiency*. *in Do Teachers Make a Difference?,* Government Printing Office, Washington, D.C.

Hanushek, EA, 'Do Teachers Make a Difference?' in *Teacher Quality, and Efficiency,* pp. 79–99.

Hanushek, EA 1979, 'Conceptual and empirical issues in the estimation of educational production functions', *The Journal of Human Resources*, vol. 14, no. 3, pp. 351–388.

Hanushek, EA 1997, 'Assessing the Effects of School Resources on Student Performance. An Update', *Educational Evaluation and Policy Analysis*, vol. 19, no. 2, p. 141.

Hanushek, EA 2003, 'The Failure of Input-Based Schooling Policies', *The Economic Journal*, vol. 113, no. 485, F64-F98. Available from: https://www.jstor.org/stable/3590139.

Hanushek, EA 2006, *Chapter 14 School Resources*. *Handbooks of the Economics of Education*.

Hanushek, EA, Kain, JF & Rivkin, SG 2009, 'New Evidence about Brown v. Board of Education. The Complex Effects of School Racial Composition on Achievement', *Journal of Labor Economics*, vol. 27, no. 3, 349-383.

Hanushek, EA, Link, S & Woessmann, L 2013, 'Does school autonomy make sense everywhere? Panel estimates from PISA', *Journal of Development Economics*, vol. 104, pp. 212–232.

Hanushek, EA, Piopiunik, M & Wiederhold, S 2018, 'The Value of Smarter Teachers. International Evidence on Teacher Cognitive Skills and Student Performance', *Journal of Human Resources,* 0317-8619R1.

Hanushek, EA & Raymond, ME 2005, 'Does school accountability lead to improved student performance?', *Journal of Policy Analysis and Management*, vol. 24, no. 2, pp. 297–327.

Hanushek, EA & Rivkin, SG 2006, 'Teacher Quality', *Handbook of the Economics of Education*, II, pp. 1051–1078.

Hanushek, EA & Rivkin, SG 2012, 'The Distribution of Teacher Quality and Implications for Policy', *Annual Review of Economics*, vol. 4, no. 1, pp. 131–157.

Hanushek, EA, Ruhose, J & Woessmann, L 2017, 'Knowledge Capital and Aggregate Income Differences: Development Accounting for US States', *American Economic Journal: Macroeconomics*, vol. 9, no. 4, pp. 184–224.

Hanushek, EA, Schwerdt, G, Wiederhold, S & Woessmann, L 2015, 'Returns to skills around the world: Evidence from PIAAC', *European Economic Review*, vol. 73, pp. 103–130.

Hanushek, EA & Woessmann, L, 'The Economics of International Differences in Educational Achievement' in *Handbook of the Economics of Education,* pp. 89–200.

Hanushek, EA & Woessmann, L 2008, 'The Role of Cognitive Skills in Economic Development', *Journal of Economic Literature*, vol. 46, no. 3, pp. 607–668.

Hanushek, EA & Woessmann, L 2012, 'Schooling, educational achievement, and the Latin American growth puzzle', *Journal of Development Economics*, vol. 99, no. 2, pp. 497–512.

Hanushek, EA & Woessmann, L 2015, *The knowledge capital of nations. Education and the economics of growth,* MIT Press, Cambridge Mass. u.a.

Hanushek, EA & Woessmann, L 2016, 'Knowledge capital, growth, and the East Asian miracle', *Science (New York, N.Y.)*, vol. 351, no. 6271, pp. 344–345.

Hanushek, EA & Zhang, L 2009, 'Quality-Consistent Estimates of International Schooling and Skill Gradients', *Journal of Human Capital*, vol. 3, no. 2, pp. 107–143.

Harden, JJ 2011, 'A Bootstrap Method for Conducting Statistical Inference with Clustered Data', *State Politics & Policy Quarterly*, vol. 11, no. 2, pp. 223–246.

Harris, DN & Sass, TR 2011, 'Teacher Training, Teacher Quality and Student Achievement', *Journal of Public Economics*, vol. 95, 7-8, pp. 798–812.

Heckman, JJ & Rubinstein, Y 2001, 'The Importance of Noncognitive Skills: Lessons from the GED Testing Program', *The American Eonomic Review*, vol. 91, no. 2, pp. 145–149.

Henderson, JV, Storeygard, A & Weil, DN 2011, 'A Bright Idea for Measuring Economic Growth', *The American Eonomic Review*, vol. 101, no. 3, pp. 194–199.

Henderson, VJ, Storeygard, A & Weil, DN 2012, 'Measuring Economic Growth from Outer Space', *The American Eonomic Review*, vol. 102, no. 2, pp. 994–1028.

Hicks, R & Tingley, D 2011, 'Causal mediation analysis', *The Stata Journal*, vol. 11, no. 4, pp. 605–619.

Hodler, R & Raschky, PA 2014, 'Regional Favoritism', *The Quarterly Journal of Economics*, vol. 129, no. 2, pp. 995–1033.

Hoffmann, L & Richter, D 2016, 'Aspekte der Aus- und Fortbildung von Deutsch- und Englischlehrkräften im Ländervergleich' in *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich,* eds P Stanat, K Böhme, S Schipolowski & N Haag, Waxmann, Münster/ New York, pp. 481–482.

Holmstrom, B & Milgrom, P 1991, 'Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design', Journal of Law, Economics and Organization, no. 7, pp. 24–52.

Hout, M & Elliott, SW 2011, *Incentives and test-based accountability in education,* National Academies Press, Washington, D.C.

Hoxby, CM 2000, 'The Effects of Class Size on Student Achievement: New Evidence from Population Variation', *The Quarterly Journal of Economics*, vol. 115, no. 4, pp. 1239–1285.

Imai, K, Keele, L, Tingley, D & Yamamoto, T 2011, 'Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies', *American Political Science Review*, vol. 105, no. 04, pp. 765–789.

IQB 2013, *IQB-Ländervergleich 2012*. *Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I,* Waxmann, Münster, Berlin.

IQB 2016, *IQB-Bildungstrend 2015*. *Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich,* Waxmann, Münster.

Jacob, BA 2005, 'Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools', *Journal of Public Economics*, vol. 89, 5-6, pp. 761–796.

Jacob, BA & Lefgren, L 2008, 'Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education', *Journal of Labor Economics*, vol. 26, no. 1, pp. 101–136.

Jacob, BA & Rockoff, JE 2011, 'Organizing Schools to Improve Student Achievement: Start Times, Grade Configurations, and Teacher Assignments', *Discussion Paper*, vol. 08.

Jansen, M, Schroeders, U, Lüdtke, O & Anand Pant, H 2014, 'Interdisziplinäre Beschulung und die Struktur des akademischen Selbstkonzepts in den naturwissenschaftlichen Fächern', *Zeitschrift für Pädagogische Psychologie*, vol. 28, 1-2, pp. 43–49.

Jerven, M 2013, *Poor numbers. How we are misled by African development statistics and what to do about it,* Cornell Univ. Press, Ithaca, NY [u.a.].

Jürges, H, Schneider, K & Büchel, F 2005, 'The Effect of Central Exit Examinations on Student Achievement: Quasi-Experimental Evidence from Timss Germany', *Journal of the European Economic Association*, vol. 3, no. 5, pp. 1134–1155.

Kling, JR, Liebman, JB & Katz, LF 2007, 'Experimental Analysis of Neighborhood Effects', *Econometrica*, vol. 75, no. 1, pp. 83–119.

Koedel, C, Mihaly, K & Rockoff, JE 2015, 'Value-added modeling: A review', *Economics of Education Review*, vol. 47, pp. 180–195.

Koning, P & van der Wiel, K 2012, 'School Responsiveness to Quality Rankings: An Empirical Analysis of Secondary Education in the Netherlands', *De Economist*, vol. 160, no. 4, pp. 339–355.

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

Koretz, DM 2009, *Measuring up. What educational testing really tells us,* Harvard University Press, Cambridge, Mass.

La Fuente, A de & Doménech, R 2006, 'Human Capital in Growth Regressions: How Much Difference Does Data Quality Make?', *Journal of the European Economic Association*, vol. 4, no. 1, pp. 1–36.

Laffont, J-J & Martimort, D ca. 2003, *The theory of incentives. The principal-agent model,* Princeton University Press, Princeton, N.J.

Lavecchia, AM, Liu, H & Oreopoulos, P 2016, 'Behavioral economics of education: Progress and possibilities', Handbook of the Economics of Education, no. 5, pp. 1–74.

Lavy, V & Schlosser, A 2011, 'Mechanisms and Impacts of Gender Peer Effects at School', *American Economic Journal: Applied Economics*, vol. 3, no. 2, pp. 1–33.

Lenski, AE, Hecht, M, Penk, C, Milles, F, Mezger, M, Heitmann, P, Stanat, P & Pant, HA 2016, *IQB-Ländervergleich 2012. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente.,* Berlin.

Leschnig, L, Schwerdt, G & Zigova, K 2017, *Central school exams and adult skills: Evidence from PIAAC*.

Link, S 2012, 'Single-Sex Schooling and Student Performance. Quasi-Experimental Evidence from South Korea', *ifo Working Papers*, vol. 146.

Lippmann, Q & Senik, C 2018, 'Math, Girls and Socialism', *IZA Discussion Paper Series*, no. 11532.

Lucas, RE 1988, 'On the mechanics of economic development', *Journal of Monetary Economics*, vol. 22, no. 1, pp. 3–42.

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

Mankiw, NG, Romer, D & Weil, DN 1992, 'A Contribution to the Empirics of Economic Growth', *The Quarterly Journal of Economics*, vol. 107, no. 2, pp. 407–437.

Metzler, J & Woessmann, L 2012, 'The impact of teacher subject knowledge on student achievement. Evidence from within-teacher within-student variation', *Journal of Development Economics*, vol. 99, no. 2, pp. 486–496.

Michalopoulos, S & Papaioannou, E 2013, 'Pre-colonial Ethnic Institutions and Contemporary African Development', *Econometrica : journal of the Econometric Society*, vol. 81, no. 1, pp. 113–152.

Michalopoulos, S & Papaioannou, E 2014, 'National Institutions and Subnational Development in Africa', *Quarterly Journal of Economics*, vol. 129, no. 1, pp. 151–213.

Min, B 2008, 'Democracy and Light. Electoral Accountability and the Provision of Public Goods', *mimeo*.

Mincer, J 1967, *Schooling, Experience, and Earnings,* NBER, New York.

Muralidharan, K & Sundararaman, V 2011, 'Teacher Performance Pay: Experimental Evidence from India', *Journal of Political Economy*, vol. 119, no. 1, pp. 39–77.

Musset, P 2010, 'Initial Teacher Education and Continuing Training Policies in a Comparative Perspective: Current Practices in OECD Countries and a Literature Review on Potential Effects', *OECD Education Working Papers*, vol. 48.

Neal, D & Schanzenbach, DW 2010, 'Left Behind by Design: Proficiency Counts and Test-Based Accountability', *Review of Economics and Statistics*, vol. 92, no. 2, pp. 263–283.

Nelson, RR & Phelps, ES 1966, 'Investment in Humans, Technological Diffusion, and Economic Growth', *The American economic review*, vol. 56, 1/2, pp. 69–75.

Niederle, M & Vesterlund, L 2010, 'Explaining the Gender Gap in Math Test Scores: The Role of Competition', *Journal of Economic Perspectives*, vol. 24, no. 2, pp. 129–144.

Nikolova, M & Nikolaev, B 2017, 'Does Joining the EU Make You Happy? Evidence from Bulgaria and Romania', *Journal of Happiness Studies*, vol. 18, pp. 1593–1623.

NOAA's National Geophysical Data Center 2016, *DMSP Data*. *Version 4 DMSP-OLS Nighttime Lights Time Series*.

Nollenberger, N, Rodríguez-Planas, N & Sevilla, A 2016, 'The Math Gender Gap: The Role of Culture', *American Economic Review*, vol. 106, no. 5, pp. 257–261.

Nunes, LC, Reis, AB & Seabra, C 2015, 'The publication of school rankings: A step toward increased accountability?', *Economics of Education Review*, vol. 49, pp. 15–23.

Nunn, N & Puga, D 2012, 'Ruggedness. The Blessing of Bad Geography in Africa', *Review of Economics and Statistics*, vol. 94, no. 1, pp. 20–36.

Nunn, N & Wantchekon, L 2011, 'The Slave Trade and the Origins of Mistrust in Africa', *The American Eonomic Review*, vol. 101, no. 7, pp. 3221–3252.

OECD 2015, *Education Policy Outlook*. *Poland*.

OECD 2016a, *Education Policy Outlook*. *Estonia*.

OECD 2016b, *Education Policy Outlook*. *Slovenia*.

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

Pant, HA, Stanat, P, Hecht, M, Heitmann, P, Jansen, M, Lenski, AE, Penk, C, Pöhlmann, C, Roppelt, A, Schroeders, U & Siegle, T 2017, *IQB Ländervergleich in Mathematik und den Naturwissenschaften 2012 (IQB-LV 2012),* IQB - Institute for Educational Quality Improvement.

Park, H, Behrman, JR & Choi, J 2018, 'Do single-sex schools enhance students' STEM (science, technology, engineering, and mathematics) outcomes?', *Economics of Education Review*, vol. 62, pp. 35–47.

Petrakis, PE & Stamatakis, D 2002, 'Growth and educational levels: a comparative analysis', *Economics of Education Review*, vol. 21, no. 5, pp. 513–521.

Pfeifer, G, Wahl, F & Marczak, M 2017, *Illuminating the World Cup Effect: Night Lights Evidence from South Africa*.

Pritchett, L 2001, 'Where has all the Education Gone?', *The World Bank Economic Review*, vol. 15, no. 3, pp. 367–391.

Pritchett, L 2015, 'Creating education systems coherent for learning outcomes: Making the transition from schooling to learning', *RISE Working Paper*, 15/005.

Ramirez, FO, Schofer, E & Meyer, JW 2018, 'International Tests, National Assessments, and Educational Development (1970–2012)', *Comparative Education Review*, vol. 62, no. 3, pp. 344–364.

Reback, R 2008, 'Teaching to the rating: School accountability and the distribution of student achievement', *Journal of Public Economics*, vol. 92, 5-6, pp. 1394–1415.

Reback, R, Rockoff, J & Schwartz, HL 2014, 'Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind', *American Economic Journal: Economic Policy*, vol. 6, no. 3, pp. 207–241.

Reuben, E, Wiswall, M & Zafar, B 2017, 'Preferences and Biases in Educational Choices and Labour Market Expectations: Shrinking the Black Box of Gender', *The Economic Journal*, vol. 127, no. 604, pp. 2153–2186.

Rivera-Batiz, LA & Romer, PM 1991, 'Economic Integration and Endogenous Growth', *The Quarterly Journal of Economics*, vol. 106, no. 2, p. 531.

Rockoff, J & Turner, LJ 2010, 'Short-Run Impacts of Accountability on School Quality', *American Economic Journal: Economic Policy*, vol. 2, no. 4, pp. 119–147.

Roger, L 2018, 'Blinded by the Light? Heterogeneity in the LuminosityGrowth Nexus and the 'African Growth Miracle'', *CREDIT Research Paper*, 18/04.

Romer, PM 1990, 'Endogenous Technological Change', *Journal of Political Economy*, vol. 98, 5, Part 2, S71-S102.

Rouse, CE, Hannaway, J, Goldhaber, D & Figlio, D 2013, 'Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure', *American Economic Journal: Economic Policy*, vol. 5, no. 2, pp. 251–281.

Sachs, JD & Warner, AM 2001, 'The curse of natural resources', *European Economic Review*, vol. 45, 4-6, pp. 827–838.

SACMEQ 2007, *III*. SPSS. Available from: http://www.sacmeq.org/?q=seacmeq-data.

Sanz-de-Galdeano, A & Vuri, D 2007, 'Parental Divorce and Students? Performance: Evidence from Longitudinal Data', *Oxford Bulletin of Economics and Statistics*, vol. 69, no. 3, pp. 321–338.

Schipolowski, S, Haag, N, Milles, F, Pietz, S. & Stanat, P 2018, *IQB-Bildungstrend 2015. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente in den Fächern Deutsch und Englisch,* Berlin.

Schütz, G, Ursprung, HW & Woessmann, L 2008, 'Education Policy and Equality of Opportunity', *Kyklos*, vol. 61, no. 2, pp. 279–308.

Schwerdt, G & Woessmann, L 2017, 'The information value of central school exams', *Economics of Education Review*, vol. 56, pp. 65–79.

Schwerdt, G & Wuppermann, AC 2011, 'Is traditional teaching really all that bad? A within-student between-subject approach', *Economics of Education Review*, vol. 30, no. 2, pp. 365–379.

Shewbridge, C, Ehren, M, Santiago, P & Tamassia, C 2012, 'OECD Reviews of Evaluation and Assessment in Education: Luxembourg 2012'.

Shewbridge, C, Jang, E, Matthews, P & Santiago, P 2011, 'OECD Reviews of Evaluation and Assessment in Education: Denmark 2011'.

Sinn, H-W, Flaig, G, Werding, M, Munz, S, Düll, N & Hofmann, H 2001, *EU-Erweiterung und Arbeitskräftemigration : Wege zu einer schrittweisen Annäherung der Arbeitsmärkte,* München.

Smith, A 1776, *An Inquiry into the Nature and Causes of the Wealth of Nations*, Strahan and Cadell.

Stanat, P, Böhme, K, Schipolowski, S, Haag, N, Weirich, S, Sachse, K, Hoffmann, L, Federlein, F, Institut zur Qualitätsentwicklung im Bildungswesen & Humboldt-Universität zu Berlin 2018, *IQB-Bildungstrend 2015 Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich (IQB-BT 2015),* IQB - Institute for Educational Quality Improvement.

Staneva, AV & Abdel-Latif, H 2016, 'From Soviet to Europe. Returns to Education Puzzle in Bulgaria', *Labour*, vol. 30, no. 3, pp. 347–367.

Statista 2012, *Beamtenstatus und Anzahl der Lehrer an Allgemeinbildenden- und Berufsschulen im Jahr 2010 nach Bundesländern*. Available from: https://de.statista.com/statistik/daten/studie/215871/umfrage/beamtenstatus-und-anzahl-der-lehrer-an-allgemeinbildenden-und-berufsschulen/ [05 March 2019].

StatSilk 2016, *SACMEQ Shapefiles*. Available from: https://www.statsilk.com/maps/download-free-shapefile-maps.

Stecher, BM 2008, *Pain and gain*. *Implementing No Child Left Behind in three states, 2004-2006,* Rand, Santa Monica, Calif.

Tartari, M 2015, 'Divorce and the cognitive achievement of children', *International Economic Review*, vol. 56, no. 2, pp. 597–645.

Terhart, E 2007, 'Strukturprobleme der Lehrerausbildung in Deutschland', *Óhidy A., Terhart E., Zsolnai J. (eds) Lehrerbild und Lehrerbildung. VS Verlag für Sozialwissenschaften,* pp. 45–65.

Tiedemann, J 2000, 'Gedner-related Beliefs of Teachers in Elementary School Mathematics', *Educational Studies in Mathematics*, vol. 41, pp. 191–207.

Tiedemann, J & Faber, G 1998, 'Mädchen im Mathematikunterricht. Selbstkonzept und Kausalattributionen im Grundschulalter', *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, XXVII, no. 1, pp. 61–71.

Todd, PE & Wolpin, KI 2003, 'On the Specification and Estimation of the Production Function for Cognitive Achievement', *The Economic Journal*, vol. 113, no. 485, F3-F33.

UIS 2018, *Drop-out rate in primary education*. *Cumulative drop-out rate to the last grade of primary education, both sexes (%)*, UNESCO Institute for Statistics. Available from: http://data.uis.unesco.org/Index.aspx?queryid=156# [19 October 2018].

Woessmann, L 2016, 'The importance of school systems: Evidence from international dierences in student achievement', *Journal of Economic Perspectives*, vol. 30, no. 3, pp. 3–31.

Woessmann, L 2018, 'Central exit exams improve student outcomes', *IZA World of Labor*.

Woessmann, L, Luedemann, E, Schuetz, G & West, MR 2009, *School accountability, autonomy, and choice around the world,* Cheltenham, UK: Edward Elgar.

World Bank 2018a, *World Bank Indicators*, World Bank Open Data. Available from: https://data.worldbank.org/ [23 April 2018].

World Bank 2018b, *World Bank Indicators*, World Bank Open Data. Available from: https://data.worldbank.org/ [23 April 2018].

World Bank 2018c, *World Development Indicators*. *Agriculture, forestry, and fishing, value added (% of GDP)*. Available from: https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS [30 October 2018].

World Bank 2018d, 'World Development Report 2018. Learning to Realize Education's Promise'. Available from: http://www.worldbank.org/en/publication/wdr2018.

World Bank 2019, *GDP per capita, PPP (current international $)*. Available from: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?end=2007&locations=ZG&start=1990 [22 February 2019].

Wossmann, L 2003, 'Schooling Resources, Educational Institutions and Student Performance: the International Evidence', *Oxford Bulletin of Economics and Statistics*, vol. 65, no. 2, pp. 117–170.

Wuertz Rasmussen, A 2009, *Family structure changes and children's health, behavior, and educational outcomes*, 09-15.

Zamarro, G, Hitt, C & Mendez, I 2016, 'When Students Don’t Care: Reexamining International Differences in Achievement and Non-cognitive Skills', *EDRE Working Paper*, 2016-18.

Zapryanova, G & Esipova, N 2016, *Most in Eastern Europe Positive About EU Membership*, Gallup World Survey. Available from: https://news.gallup.com/poll/210083/eastern-europe-positive-membership.aspx [21 January 2019].

Annika B. Bergbauer

Winzererstraße 49 – 80797 München – Tel.: 0177 85 62 990

Email: a.b.bergbauer@gmail.com

Praktische Erfahrung

seit 01/ 2016	ifo Institut e. V., Zentrum für Bildungsökonomik, München, Deutschland Wissenschaftliche Mitarbeiterin
03/ 2015 – 12/ 2015	Research on Socio-Economic Policy (RESEP) der Universität Stellenbosch, Südafrika Assoziierte Wissenschaftlerin
08/ 2010 – 07/ 2015	Internationaler Bund Göttingen, Deutschland Pädagogische Referentin
07/ 2014 – 09/ 2014	giz – Gesellschaft für internationale Zusammenarbeit Nachhaltige Infrastruktur, Transport und Mobilität, Eschborn, Deutschland Praktikantin
02/ 2012 – 04/ 2012	KfW Entwicklungsbank Bildung in Nordafrika und Naher Osten, Frankfurt, Deutschland Praktikantin

Ausbildung

01/ 2016 – 07/2019	Ludwig-Maximilians-Universität (LMU), München, Deutschland Promotion in Volkswirtschaftslehre
04/ 2013 – 10/ 2015	Georg-August-Universität Göttingen, Deutschland Master of Arts in Development Economics
01/ 2014 – 06/ 2014	Stellenbosch University, Südafrika Studienaufenthalt
07/ 2013 – 12/ 2013	Delhi School of Economics, Indien Studienaufenthalt
10/ 2009 – 04/ 2013	Georg-August-Universität Göttingen, Deutschland Bachelor of Arts in Economics
10/ 2011 – 01/ 2012	École Supérieure de Commerce International, Marne-la-Vallée, Frankreich Studienaufenthalt