ifo
BEITRÄGE
zur Wirtschaftsforschung

102
2023

Determinants and Consequences of Student Test Scores: Evidence from International, Big, and Text Data

Pietro Sancassani

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

ifo
BEITRÄGE
zur Wirtschaftsforschung

102
2023

Determinants and Consequences of Student Test Scores:
Evidence from International, Big, and Text Data

Pietro Sancassani

Herausgeber der Reihe: Clemens Fuest
Schriftleitung: Chang Woon Nam

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
https://dnb.d-nb.de
abrufbar.

ISBN Nr. 978-3-95942-125-6

Druck: Kreiter Druck, Wolfratshausen

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

Determinants and Consequences of Student Test Scores: Evidence from International, Big, and Text Data

Inaugural-Dissertation

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

eingereicht an der
Ludwig-Maximilians-Universität München
2023

vorgelegt von

Pietro Sancassani Borea Ricci

Referent: Prof. Dr. Ludger Woessmann
Koreferent: Prof. Massimo Anelli, PhD
Promotionsabschlussberatung: 12.07.2023

Datum der mündlichen Prüfung	07.07.2023
Namen der Berichterstatter:innen	Prof. Dr. Ludger Woessmann Prof. Massimo Anelli, Ph.D. Prof. Ines Helm, Ph.D.

Preface

Pietro Sancassani prepared this study while he was working at the Center for Economics of Education at the ifo Institut. The study was completed in March 2023 and accepted as a doctoral thesis by the Department of Economics at the LMU Munich. It consists of four distinct empirical essays and addresses various determinants and the consequences of student test scores. Chapter 2 investigates the impact of four teacher characteristics – whether teachers hold a Master's degree, a subject-specific qualification, a major in education, or their level of experience – on student science test scores in ten different countries. Chapter 3 shows that teacher subject-specific qualifications positively affect student science test scores in thirty countries around the world. Chapter 4 shows the association between a measure of patience derived from social media data and student test scores at the regional level. Finally, Chapter 5 shows that the salience of the education topic induced by the "PISA shock" in Germany led to an increase in the polarization of parliamentary debates about education.

Keywords: Student Test Scores, Teacher Characteristics, Teacher Qualifications, Teacher Quality, Teacher Subject-Specific Qualifications, Human Capital, Patience, Cultural Preferences, PISA shock, Polarization, Parliamentary Debates

JEL-No: I21, I24, I29, J24, H75, P16

Acknowledgement

I thank my supervisor Ludger Woessmann for his guidance and mentoring throughout this journey. I am grateful for the excellent research environment, great colleagues, and all the wonderful opportunities that being part of his team offers. It has been an honor.
Special thanks go to my second supervisor Massimo Anelli. His advice to pursue this path has borne fruit. I also want to express my gratitude to Ines Helm for being my constructive third supervisor.

I thank my coauthor Rick Hanushek for his wit, sharpness, and kindness. Working with him has taught me a lot. This journey would not have been the same without my coauthor and partner Lavinia Kinne. I have learnt a lot from her as a colleague, but even more as a partner.

Thanks to my current and former colleagues Benjamin Arold, Annika Bergbauer, Alex Bertermann, Raphael Brade, Vera Freundl, Elisabeth Grewenig, Sarah Gust, Franziska Hampf, Philipp Lergetporer, Lukas Mergele, Simon ter Meulen, Caterina Pavese, Franziska Pfaeheler, Marc Piopiunik, Sven Resnjanskij, Florian Schoner, Moritz Seebacher, Lisa Simon, Katharina Wedel, Katharina Werner, Anna Würm, and Larissa Zierow. Although most of the work in this dissertation only carries my name, you have all contributed to it. It has been a pleasure sharing these years with you and I hope our paths will cross again.
Also thanks to Franziska Binder and Ulrike Baldi-Cohrs for their support and kindness. You make ifo BI a great place.

I have shared my first three years of the PhD with the colleagues-from-afar of the OCCAM network. It has been great meeting you and sharing ideas. I also gratefully acknowledge the European Training Network OCCAM and the Smith Richardson Foundation for generous funding.

Final thanks go to my family and friends. In years of uncertainty and doubts, your unconditional love and support were a much-needed certainty.

Contents

List of Tables

List of Figures

1 Introduction

1.1 The Economics of Human Capital

Economists have investigated the determinants of economic growth and labor market success of individuals for centuries. Among the many factors that have been considered, one that has been consistently associated with both is human capital. Although there was virtually no use of the term "human capital" until the late 1950s (Goldin 2016), the "Father of Economics" Adam Smith already alluded to this concept in his eighteenth-century classic *The Wealth of Nations ([1776] 1979).* There, he identifies the "*the acquired and useful abilities"* (p. 283) of individuals as a fundamental part of the general stock of capital of any country or society. Crucially, Smith also notes that such skills can be acquired and improved upon through education and training. The importance of human capital and education in economics has therefore been recognized since the dawn of this discipline.

Despite its early recognition, the theoretical foundations of the role of human capital in economics were not laid until the late 1950s. In a series of articles, Mincer (1958), Schultz (1961) and Becker (1962) formalized the cost-benefit rationale that underlies educational investment decisions to advance individuals' skills in what became known as human capital theory, enshrined in Becker's book *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* (1964). Mincer (1958) also started to link individuals' years of schooling to their subsequent earnings. In his landmark book *Schooling, Experience, and Earnings* (1974), Mincer demonstrated the existence of such relationship by modelling the logarithm of earnings as a function of years of education and labor market experience. This equation has become the "workhorse" of empirical research on earnings determinants and one of the most widely used models in empirical economics (Lemieux 2006). Since then, numerous studies have provided causal evidence on the positive impact of individuals' human capital on wages (e.g. Card 1999; Heckman, Lochner, and Todd 2006) as well as other economically relevant outcomes, such as unemployment (e.g. Ashenfelter and Ham 1979; Nickell 1979), and health (e.g. Deaton and Paxson 2001; Cutler and Lleras-Muney 2006).

Human capital has also been at the center of the macroeconomic literature investigating the determinants of economic growth. Economic output was initially

modeled as a function of capital and labor in the Solow-Swan growth model (Solow 1956; Swan 1956), where the labor stock only included the amount of workers in an economy and the time spent working. The substantial gap observed between the actual stock of capital and labor and economic output, the "Solow residual", pushed scholars to enrich the Solow-Swan growth model with human capital. In the augmented neoclassical growth model, human capital became a fundamental input for economic growth (Mankiw, Romer, and Weil 1992). A direct implication of this model is that education, by improving individuals' skills and productivity, enhances economic growth. A distinct formalization of the role of human capital for economic growth came from the endogenous growth models. By creating new technologies, human capital increases the innovative capacity of an economy and generates economic growth (Romer 1990; Howitt and Aghion 1998).

A key passage that occurred in the last decades came from the measurement of human capital. Also thanks to their increasing availability, student cognitive skills superseded years of education as the preferred measure for human capital. By measuring what individuals learn in school rather than the time they spend at school, student cognitive skills revealed that human capital played an even more important role in the economy than previously thought. Using student cognitive skills in math and science, Hanushek and Woessmann (2008, 2012) showed that human capital is the most important determinant of long-run economic growth. Such relationship is considerably weaker if years of schooling is used as a proxy for human capital. Similarly, Altonji and Pierret (2001) showed that workers' cognitive skills are a much better predictor of wages than years of education or degree.

Emboldened by these findings, economists have increasingly focused on the education production function, that examines the relationship among the different inputs into and outcomes of the educational process (e.g., Hanushek 1986). At the school level, examples of inputs that have been linked to student outcomes are class size (e.g., Woessmann and West 2006, Angrist et al. 2019), teacher quality (e.g., Chetty, Friedman, and Rockoff 2014; Hanushek, Piopiunik, and Wiederhold 2019), teaching methods (e.g., Schwerdt and Wuppermann 2011; Bietenbeck 2014), and instruction time (e.g., Lavy 2015; Rivkin and Schiman 2015; Wedel 2021). At the institutional level, the existence of a tracking system (Hanushek and Woessmann 2006), school accountability (Bergbauer, Hanushek, and Woessmann 2021), school expenditure (e.g., Jackson, Johnson, and Persico 2016; Jackson, Wigger, and Xiong 2021), and preferences (e.g. Figlio et al. 2019; Hanushek et al. 2022) are examples of inputs that have been linked to student outcomes.

Looking ahead, human capital looks set to play an even more important role in the economy. The importance of cognitive skills for thriving in a digital and interconnected economy has been widely acknowledged (OECD 2016a, 2017). It is therefore crucial to develop a deeper understanding of what contributes to student outcomes, as these are key for long-run prosperity and the labor market success of individuals. It is also important to study what are the consequences of student outcomes for the public opinion and the political systems. As much of education worldwide is public, policies aimed at improving education necessarily need to go through a political process.

This dissertation aims at shedding light on what factors affect student test scores, and how test scores affect the political debates about education. I focus on two areas that have been shown to be important for student test scores: teachers and intertemporal preferences. In particular, I investigate the impact of various teacher characteristics on student test scores in science in an international context. I then turn to an important intertemporal preference, patience, which is crucial for education investment. This dissertation shows that patience levels in the population account for large portions of differences in student achievement both across and within countries. Finally, I turn to the impact that student test scores have on political debates. I exploit the release of the results of the first *Programme for International Student Achievement* (PISA) study in Germany, which revealed an unexpectedly low performance of German students. I show that this event increased the polarization of parliamentary debates about education in Germany.

The introduction is structured as follows: in Section 1.2, I provide an overview on the data used in the dissertation. I leverage international, large and unstructured data, which represent an important element of novelty in this field. In Section 1.3, I review the empirical methods used. To cope with the variety of data analyzed, I use methods beyond the standard econometric approaches, such as machine-learning and text analysis techniques that allow me to retrieve and analyze unstructured data. In Section 1.4, I conclude with an overview of the chapters and related policy implications.

1.2 Data

In this section, I provide an overview of the data used in this dissertation. I briefly describe the main data sources for student test scores, namely the international *Trends in International Mathematics and Science Study* (TIMSS) and PISA data, as well

as the Italian *Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione* (INVALSI) data, and U.S. *National Assessment of Educational Progress* (NAEP) data. I then introduce Facebook data on the interests of the over 3 billion Facebook users, an innovative data source that is used to derive patience and risk-taking preferences for over 200 countries, Italian regions, and U.S. states. Finally, I present the collected and digitized speeches from parliamentary debates of the 16 German states. I use these novel data for a text analysis of the education debates in Germany.

A central source of data analyzed in this dissertation consists of international and national student assessments. In Chapter 2 and Chapter 3, I use data from the *Trends in International Mathematics and Science Study* (TIMSS). TIMSS is an international large-scale assessment of students' skills in mathematics and science that has been administered every four years since 1995. Thanks to the sampling of entire classes and rich questionnaires of student, teacher, and school characteristics, TIMSS is particularly suited to study the relationship between student skills and teacher characteristics in international settings. I therefore leverage TIMSS data to provide evidence on the impact of teacher qualifications on student science achievement for over 40 countries. This represents an important contribution to this literature, which has mostly focused on national settings, thereby limiting its external validity.

Chapter 4 uses data from both international and national large-scale assessment. It combines PISA data on student test scores for over 80 countries and 2.6 million students. Similar to TIMSS, PISA is an international large-scale assessment of student skills in math, reading and science which is conducted every three years. Chapter 4 also uses national data of student math skills for Italy (INVALSI) and the United States (NAEP). These data sources are combined in an analysis that investigates the relationship between patience and student test scores.

Data to create a measure of patience for over 200 countries as well as Italian regions and the U.S. states are retrieved from Facebook. By collecting information on the interests of its over 3 billion monthly active users, Facebook has inadvertently built the largest available platform for the measurement of culture. Country and regional data on Facebook users' interests have been retrieved by systematically querying the Facebook Marketing application program interface (API), a tool offered by Facebook to configure advertisement campaigns. Together with scientifically validated measures of patience and risk-taking preferences for 76 countries from the *Global*

Preference Survey (GPS), Chapter 4 develops Facebook-derived measures of patience and risk-taking.

In Chapter 5, I have collected, digitized and analyzed the parliamentary debates of the 16 German state parliaments. I web-scraped the website of each of the 16 German states to create a novel dataset that includes all their parliamentary debates occurred in the period 2000-2008. I then used text analysis methods to parse the documents containing the debates and to extract the speeches and other relevant information such as the speaker, her role, party affiliation, state, and date of the debate. These data represent a new data source that enables me to study the impact of the release of the first PISA results in Germany on the political debates about education.

1.3 Empirical Methods

Given the variety of data sources used in this dissertation, this section provides an overview of the methods used to analyze them. I start with the microeconometric methods in Subsection 1.3.1. Microeconometric methods are the standard tools in applied microeconomics to retrieve causal estimates. I then move on to more recent techniques from the machine-learning literature in Subsection 1.3.2. I use machine-learning methods to predict patience and risk-taking preferences for countries and regions for which survey measures are not available and to classify the topics of the speeches in the parliamentary debates in the German state parliaments. Finally, I provide a brief overview of the text analysis methods used to parse these debates and analyze the speeches in Subsection 1.3.3.

1.3.1 Microeconometric Identification

A simple correlation of the relationship between teacher characteristics and student achievement is not suited to estimate causal effects. In fact, teacher characteristics are unlikely to be distributed equally among, for example, students from high and low socioeconomic status (SES). If teachers with better qualifications or higher experience are systematically assigned to high SES students, who tend to perform better in school, the estimated relationship will be biased. Linear regressions that control for observable characteristics, such as student SES, student or teacher gender, instruction time, or school location, are also unlikely to yield causal estimates. Unobservable characteristics, such as student or teacher ability, might still bias the estimates if they are correlated with teacher qualifications and they affect student achievement.

To address these concerns, in Chapter 2 and 3 I take advantage of an identification strategy that exploits the availability of student science achievement in four distinct subjects: biology, physics, chemistry, and earth science. I then include student fixed effects in a linear regression model, thereby estimating whether differences in teacher characteristics across the four science subjects are systematically related to differences in student performance across the same four subjects. This identification strategy has often been used in the literature to address concerns of student and teacher sorting and unobservable characteristics (e.g., Harris and Sass 2011; Metzler and Woessmann 2012; Bietenbeck, Piopiunik, and Wiederhold 2018). Student fixed effects control for all the unobserved student characteristics that do not vary across subjects, such as student ability, general motivation or intertemporal preferences, that are likely to affect the outcomes of interest. The effect of teacher characteristics is therefore estimated exploiting only within-student variation in student test scores and teacher characteristics.

In Chapter 3, I focus on a sample of countries where the same teacher teaches the four science subjects—biology, physics, chemistry, and earth science—to also include teacher fixed effects. This identification strategy has the additional advantage of controlling also for unobserved teacher characteristics that do not vary across subjects, such as teacher ability or motivation.

Thanks to the identification strategies developed in Chapter 2 and 3, differences in the observed outcomes across the different science subjects can be credibly attributed to the analyzed teacher characteristics. A host of validity and robustness checks corroborate the validity of my results. In fact, results are robust across sub-samples of male or female students, high- and low-SES students, are not driven by specific countries. In Chapter 3, I also perform the analysis of unobservable selection and coefficient stability following Oster (2019), which addresses concerns about remaining confounders. Reassuringly, results from this analysis show that any bias due to unobservable characteristics should be negligible.

In Chapter 5, I estimate the impact of the increase in the salience of education induced by the release of the first PISA results in Germany—the PISA shock—on the polarization of parliamentary debates about education. Since polarization in parliamentary debates varies over time, I use a difference-in-differences approach, which controls for general underlying trends in polarization. I show that the polarization of education debates before the PISA shock was following the same trend of polarization in other topics. Hence, the main hypothesis necessary to estimate

causal effects with a difference-in-differences approach, the parallel trends, seemingly holds in my case. The main finding from this Chapter is that polarization increased as a consequence of the PISA shock, and the effect lasted for about six years. Further, I conduct a placebo test where I show that the PISA shock only affected the polarization of education debates and not the polarization of other topics, thus suggesting that the estimated effect is not biased by spillover effects. A series of robustness checks confirm that the results are not due to the specific time window nor to the specific measure of polarization used.

1.3.2 Machine-Learning Methods

Machine-learning methods are at the center of a fast growing methodological literature and are increasingly used in economics (Athey and Imbens 2019). In this subsection, I limit my description to the methods used in this dissertation, although they are suited for a wide range of applications.

In Chapter 4 and 5, I use supervised machine-learning methods with the main aim of generating predictions. Supervised machine-learning methods learn the relationship between a set of covariates and a target variable. The parameter estimates from these models can then be used to make predictions for all units for which the covariates are available, but the target variable is not. Chapter 4 uses a least absolute shrinkage and selection operator (LASSO) model to learn the relationship between the Facebook interests and the patience and risk-taking preferences of all the countries that participated in the GPS. The parameter estimates from this model are then used to make out-of-sample predictions of patience and risk-taking preferences for countries and regions for which GPS measures are not available. In Chapter 5, I use a Logistic Classifier that learns the relationship between the words used in parliamentary speeches and a label indicating whether the speech is about education or not. The set of speeches for which such label is available is only a small subset of the entire corpus of speeches. The Logistic Classifier therefore allows me to extend the classification of whether speeches are about education or not to the entire corpus of speeches by making out-of-sample predictions.

I also use unsupervised machine-learning methods. These methods are typically used for dimensionality reduction purposes or to find patterns from unlabeled data. Chapter 4 uses Principal Component Analysis (PCA) to reduce the dimensionality of the Facebook interests. In Chapter 5, I use topic modeling, a class of unsupervised machine-learning methods used to infer the underlying topics in a set of documents. Specifically, I use the Correlated Topic Modeling (CTM) (Blei and Lafferty 2007), which

allows me to classify the topics of all the speeches in the corpus of the parliamentary debates.

1.3.3 Text-Analysis Methods

I use text-analysis methods to parse and analyze parliamentary debates in Chapter 5. For parsing the documents containing the German state parliamentary debates, I use an exhaustive collection of regular expressions. Regular expressions are a sequence of characters that specify a search pattern in text. By leveraging the structure of the documents and debates, regular expressions allow me to capture all the relevant features in the debates, such as the name of a speaker, her role, party affiliation, speech, interruptions etc. This process allows me to convert the parliamentary debates into a dataset suitable for subsequent analyses.

The main outcome of interest in Chapter 5 is polarization in parliamentary debates. To measure polarization, I use a combination of standard text-analysis methods. In particular, I first convert each speech into a vector by means of the term-frequency inverse-document frequency ($tf-idf$) transformation. A $tf-idf$ representation of a speech consists of a vector where each element corresponds to a word in the corpus. The value that each element in the vector takes is given by the relative term frequency (tf) of the corresponding word in the speech weighted by the inverse of the document frequency (idf), a measure of how often the word appears in the corpus. Intuitively, a $tf-idf$ representation of a text will upweight words that appear relatively frequently in a speech and downweight words that appear frequently in the entire corpus, as these are deemed not particularly informative.

I then compute polarization as the opposite of text similarity. To this purpose, I use a standard measure of text similarity, namely the cosine similarity. The cosine similarity is computed by calculating the inner product between two vectors. To obtain a measure of polarization, I compute the cosine similarity between the $tf-idf$ vector representation of each speech in the corpus and the "average" speech from a benchmark party in the same legislative period, state, and topic. The "average" speech consists of the average between the vector representation of all the speeches from the benchmark party in one legislative period, state, and topic. To give a straightforward interpretation, I use the opposite of the cosine similarity as a polarization measure. In this way, the polarization measure of speeches that are less similar to the "average" speech from a benchmark party will be larger.

1.4 Chapter Overview

In this section, I provide a summary and the policy implications of the four essays that are part of this dissertation. The first three essays investigate aspects related to the determinants of student test scores, while the last essay analyses the consequences of student test scores for political debates about education. Each essay is self-contained and addresses a distinct research question.

Chapter 2 addresses an important factor of the education production function: teachers. Specifically, Chapter 2 investigates the impact of four teacher characteristics on student science achievement. The analyzed characteristics are whether teachers hold a Master's degree, a major in education, a subject-specific qualification, and their level of experience. It uses data from TIMSS 2015, an international large-scale assessment of student skills described in Section 1.2. The identification strategy exploits the feature that in many education systems different science subjects—physics, biology, chemistry, and earth science—are taught by different teachers. By leveraging the availability of students' test scores as well as teachers' questionnaires for each of these subjects, it implements a within-student approach which controls for unobserved student heterogeneity. Consistent with the literature investigating the impact of teacher characteristics, it finds that teachers' Masters' degree or major in education do not have a significant impact on student test scores. Similarly, teacher experience does not have a positive impact on students test scores, but students with more experienced teachers tend to report that they like studying their subjects less and find the teaching less engaging. The only teacher qualification that has a positive impact on student science achievement is whether teachers hold subject-specific qualifications in the subjects they teach.

Chapter 3 builds on the findings of Chapter 2 in that it focuses only on teacher subject-specific qualifications using TIMSS 2015 data. A fundamental difference is that Chapter 3 focuses on a distinct set of countries where the same teacher teaches the four science subjects—physics, biology, chemistry, and earth science. This allows the implementation of a model with both student and teacher fixed effects, which controls for unobserved heterogeneities in both student and teacher characteristics. Results from this model indicate that teacher subject-specific qualifications increase student science achievement and are robust to a variety of specifications, including using TIMSS 2011 data and controlling for instruction time, using OECD countries only, focusing on scarcely populated areas where teacher sorting is less likely, and are not driven by any specific subject or country.

In terms of policy implications, countries should promote prospective teachers to obtain subject-specific qualifications. By raising the standards required to become science teachers, education systems worldwide could improve students' science skills, which are crucial to address the demand for employees with a STEM background (OECD 2016b). Conversely, teacher qualifications such as Master's degrees or majors in education do not seem to be essential for students' science achievement and should therefore not be prioritized when recruiting science teachers. A similar argument can be made for teachers' experience. While it does not seem to be crucial for students' science achievement, it can negatively affect the extent to which students enjoy learning science or find the teacher engaging. Hence, teachers could benefit from professional development programs aimed at maintaining and fostering engaging teaching methods throughout their careers.

Chapter 4 investigates the extent to which a fundamental intertemporal preference, patience, accounts for differences in student achievements across Italian regions and U.S. states. This chapter is joint work with Eric A. Hanushek, Lavinia Kinne, and Ludger Woessmann. A key notion of human capital theory is that education can be considered as an investment in human capital. Hence, decisions to accrue skills should crucially depend on individuals' time preferences. However, traditional survey measures of patience are not readily available at the regional level. By leveraging the vast data available on social media—Facebook interests—Chapter 4 derives regional measures of patience within Italy and the United States. Results indicate a strong positive association of patience with student achievement across regions in both countries. Patience accounts for over two thirds of the achievement variation across Italian regions and over one third across U.S. states.

These results have important policy implications. First, they suggest that providing education systems in different regions with the same resources is not enough to address regional disparities if cultural preferences such as patience are not taken into account. Second, policies aimed at improving patience attitudes among students seem a promising way to level regional disparities in student achievement.

Chapter 5 investigates the impact of a salience shock—the PISA shock—on the polarization of parliamentary debates about education. It combines machine-learning algorithms and text-analysis methods, which are used to classify the topics of the parliamentary debates and compute the polarization measures. Exploiting the unexpectedly low performance of German students revealed by the publication of the results of the first PISA study in 2000 and the subsequent media attention that this

event received, Chapter 5 shows that the polarization of parliamentary debates on education topics in German state parliaments increased substantially after the PISA shock. It also shows that the share of speeches about education increased, and that the effect was long-lasting and faded after about six years. Additionally, the increase in salience was also accompanied by an increase in the number of initiated bills about education.

A key take-away from this Chapter is that student test scores matter for the political debates about education. International large-scale assessments such as PISA have therefore the potential of putting education topics under the spotlight and to foster related political debates. Further, results also indicate that an increase in polarization can coexist with vibrant lawmaking process, as suggested by the increased number of bills observed after the PISA shock.

2 The Effect of Teacher Characteristics on Students' Science Achievement*

2.1 Introduction

There is ample evidence that teachers have a large impact both on students' performance at school (e.g. Hanushek 1971; Murnane 1975; Rockoff 2004) as well as on a variety of outcomes later in life (Chetty, Friedman, and Rockoff 2014). However, little is known about what characteristics and teaching methods make a good teacher. The literature repeatedly demonstrates that observable teacher characteristics, especially those related to education and experience, do not tend to be good indicators of teacher quality (Hanushek 1986; Rivkin, Hanushek, and Kain 2005; Clotfelter, Ladd, and Vigdor 2007; Staiger and Rockoff 2010, among others). On the other hand, in most settings it is often difficult to credibly estimate the impact of teacher characteristics on students' performance. Unobserved student and teacher characteristics as well as sorting of students and teachers into classes and schools are only some of the most obvious threats to identification in this area.

In this paper, I investigate in an international context the impact of four teacher characteristics, namely teachers' education level, scope of experience, subject-specific qualifications, and pedagogical preparation, on students' performance. These are important characteristics as education and experience are the traditional determinants of teacher recruitment and compensation. I exploit the availability of test scores from four scientific subjects (physics, chemistry, biology and earth science) available for each 8th grade student participating in the *Trends in International Mathematics and Science Study 2015* (TIMSS 2015). Furthermore, I exploit the availability of teachers' questionnaires for each science teacher that teaches at least one science subject. My sample only includes countries in which these science subjects are taught by at least two different teachers. This is a unique setting that allows me to implement a within-student across-teachers approach by linking teachers' characteristics in one specific science subject to students' outcomes in the same subject. Using student fixed effects, I eliminate any source of unobserved student heterogeneity, such as innate abilities or effort, that is not subject-specific. To

* This chapter is based on the paper "The Effect of Teacher Characteristics on Students' Science Achievement", *ifo Working Paper* 348, 2021.

uncover some possible mechanisms through which teacher characteristics affect student performance, I also explore their impact on the extent to which students enjoy learning a subject or find teaching engaging.

In the within-student approach, other unobserved sources of student heterogeneity which are subject-specific, such as student preferences or abilities, might still bias the estimates if they are consistently associated with the mechanism through which teachers are allocated. However, this is less of a concern when the multiple outcomes belong to the same field, as in this case. A further advantage of using closely related outcomes in a within-student across-teachers approach is that this model relies on the assumption that the impact of teachers is the same across subjects. In studies using a similar approach (e.g. Metzler and Woessmann 2012; Bietenbeck, Piopiunik, and Wiederhold 2018; Hanushek, Piopiunik, and Wiederhold 2019), multiple outcomes for a single student belong to different fields (math and reading, for instance). This study uses outcomes which are more alike and, therefore, more likely to require similar skills, thereby relying on weaker assumptions.

The main result of my analysis is that teacher subject-specific qualifications have a positive and significant effect on students' science test scores. This effect is equivalent to 1.7-1.8% of a standard deviation of the students' test scores. Evidence from the US links an increase in teacher value-added by one standard deviation to an increase in student achievement by 10-20% of a standard deviation.[1] From this perspective, teacher subject-specific qualifications would explain between 9-18% of the variation in teacher effectiveness.

This effect is relatively small if compared to teacher interventions reported in other studies. For example, Taylor and Tyler (2012) report an impact of 5-11% of a formal peer evaluation program for teachers on student performance. Jackson and Makarin (2018) find an impact of 6-9% of a standard deviation of providing teachers with high-quality lesson plans on student outcomes. With respect to other instructional inputs, Lavy (2015) finds an effect of 6% of a standard deviation for an additional hour of instruction time per week. On this basis, the effect of being taught by a teacher with subject-specific qualifications corresponds to about 18 additional minutes of instruction time per week. Nevertheless, it should be kept in mind that the effect of

[1] The figure for the US is reported in Jackson, Rockoff, and Staiger (2014). The lower- and upper-bound of the estimates refer to English and math teachers, respectively. Thus, teachers seem to have a larger impact in math, which, unlike English, is mostly learned in school. In this sense, science is more similar to math.

teacher subject-specific qualifications stems from teachers teaching a science subject in which they are already specialized. Differently from the other teacher interventions mentioned previously, this effect could be achieved at virtually no cost by allocating science teachers according to their specializations.

I find a larger effect for female students and for students coming from more affluent backgrounds. I do not find a significant impact of the other teacher characteristics (education level, experience, and major in education) on students' achievements. The impact of subject-specific qualifications is robust to the addition of student indicators aiming at capturing remaining subject-specific within-student heterogeneity, namely the extent to which students enjoy learning the subject or find the teaching engaging. As such indicators are also a potential channel through which teachers can affect students' test scores, I also perform a mediation analysis. The results of this analysis show that teacher experience has a significant negative impact on the extent to which students enjoy learning a subject or find the teaching engaging. This result is robust across all subjects and model specifications. Other teacher characteristics do not have a significant impact on these indicators.

The effect of teacher subject-specific qualifications is in line with the recent literature on the effects of subject-specific teacher skills. Bietenbeck, Piopiunik, and Wiederhold (2018), for example, find an effect of 3% of a standard deviation of teacher subject knowledge on 6th-grade students' reading and math scores in Sub-Saharan Africa. Using a Peruvian 6th-grade dataset, Metzler and Woessmann (2012) find that one standard deviation in subject-specific teacher achievement increases student achievement in math by about 9% of a standard deviation, although the effects on reading are mostly insignificant. Hanushek, Piopiunik, and Wiederhold (2019) find a significant effect, equivalent to 11% of a standard deviation in students' test scores, of teachers' numeracy and literacy skills in 31 developed countries.

I do not find an effect of teacher experience on students' test scores. The literature seems to suggest that the greatest gains in teacher performance from experience occur in the early years of their careers and then quickly flatten (e.g. Rivkin, Hanushek, and Kain 2005; Clotfelter, Ladd, and Vigdor 2006; Boyd et al. 2008; Harris and Sass 2011). This might not be reflected in this analysis as the average teaching experience in my sample is relatively high. Only 5% of the teachers have less than 3 years of experience.

It has been observed in several studies that holding a Master's degree is generally not a strong predictor of teacher performance, as summarized by Hanushek and Rivkin

(2004), among others. I also do not find a significant effect. There is no conclusive evidence regarding the impact of pedagogical preparation. This aspect, however, has received little attention in the literature so far. In line with my results, Harris and Sass (2011), for example, report no impact of teachers having majored in education on their performance as measured by student outcomes.

This paper contributes to the literature by investigating the impact of teacher characteristics on student achievement in four closely related science subjects in a unique setting. To the best of my knowledge, this is the first study that focuses on the performance of students in the natural sciences using a within-student across-teachers approach. In fact, the impact of teacher characteristics on student test scores may vary between subjects (Metzler and Woessmann 2012; Kane, Rockoff, and Staiger 2008). It is therefore important to increase our knowledge of the potentially different effects of teacher characteristics on different subjects. Furthermore, I provide additional insights into the possible mechanisms by which teacher characteristics affect student performance. Overall, the results tend to be in line with the literature and confirm that observable teacher characteristics only explain a limited amount of variation in student test scores. This can have important implications for the mechanisms by which teachers are selected and compensated, as other aspects might be more relevant.[2]

The remainder of the paper is structured as follows: Section 2.2 describes the data and provides some descriptive characteristics. Section 2.3 presents the estimation strategy. The results, mediation analysis and robustness checks are discussed in Section 2.4. Section 2.5 concludes.

2.2 Data and Descriptive Statistics

2.2.1 TIMSS 2015 and Sample Selection

I use data from TIMSS 2015, an international large-scale assessment which tests 4th and 8th grade students worldwide in math and science. TIMSS employs a two-stage clustered sampling design to draw a representative national sample from each

[2] A growing body of literature considers different forms of teachers' cognitive skills, such as teachers' scores on licensure tests (Clotfelter, Ladd, and Vigdor 2006; Goldhaber and Anthony 2007; Harris and Sass 2011), tests of teachers' subject knowledge (Metzler and Woessmann 2012; Bietenbeck, Piopiunik, and Wiederhold 2018) or country-level teachers' cognitive skills (Hanushek, Piopiunik, and Wiederhold 2019). These tend to be more consistent predictors of teacher effectiveness, but they are rarely observed.

participating country. It includes tests of entire classes within randomly selected schools in a country with sampling probabilities proportional to school size as well as background questionnaires for students, teachers, and schools. The TIMSS achievement scale was established in 1995 with a scale center point of 500 located at the mean of the combined distribution of the participating countries and a standard deviation of 100.

I focus on the achievements of 8th graders in science as this is the most suitable setting for my identification strategy. 8th graders are usually around 14 years old and their science test score is made up of four subjects: biology (35%), chemistry (20%), physics (25%) and earth science (20%).[3] Tests scores are available for each student and subject,[4] thus yielding 4 observations at most for each student in science.[5] Furthermore, there are countries in which specific science subjects are taught by different teachers, which constitutes the type of variation I exploit in this analysis. This clear distinction between closely related subjects is rather special as it typically does not occur at such an early stage of education.

In this setting, I implement a within-student across-teacher model in an international context, where the deviation of test score in one subject from the average science performance of each student is associated with the deviation of teacher characteristics in the same subject from the average science teacher characteristics of each student. Due to the design of international large-scale assessments like TIMSS, this approach is not immune to criticism (e.g. Jerrim et al. 2017). In fact, these tests typically use a matrix-sampling approach in which students complete different booklets that contain a subset of questions from a common pool. If a student's booklet does not contain any questions regarding a specific subject or domain, the score in the missing subject or domain would be derived from her performance in other subjects using item response theory. The resulting within-student variation would therefore only capture the noise caused by the imputation technique, which may be a problem for the kind of identification I use. However, each booklet of the

[3] In a typical 8th grade science curriculum, biology includes topics such as the characteristics, systems and processes of living things. Physics and chemistry topics include the study of the matter and energy, electricity and magnetism. Earth science topics are, e.g., the earth's physical features and the solar system. More information can be found in Mullis and Martin (2013).

[4] TIMSS provides 5 plausible values for each student test score. I use the first plausible valuable for each subject.

[5] Depending on countries' curricula, some exceptions are possible; students in Sweden, for instance, are not tested in earth Science as this subject does not belong to their 8th grade curriculum.

TIMSS 2015 contains two science blocks and two math blocks and each science block replicates the proportion of domains that constitute a subject as indicated in TIMSS guidelines.[6] Thus, the scores available for each student reflect the actual performance in each subject. These features make this setting suitable for my analysis.

I obtain the main variables of interest from the teacher questionnaire. I consider teachers to hold a Master's degree if they report having completed a Master's degree or higher.[7] The subject-specific qualifications of teachers are determined by whether teachers hold a major in the subject that they teach.[8] It is important to highlight that this allows me to identify whether teachers have a major in one of the four specific science subjects that are tested in TIMSS. Pedagogical preparation is captured by a variable indicating whether teachers have a major in general education or in science education.[9] These variables are all binary indicators and constitute the main features of teacher preparation. Holding a Master's degree indicates that a teacher has an advanced education level, while holding subject-specific qualifications and holding a major in education capture the content and pedagogical knowledge of a teacher, respectively. Years of experience constitute an important teacher characteristic, as more experience tends to be associated with more effectiveness in the job.

These variables provide a common metric to describe teacher preparation in an international context. Nevertheless, the actual quality of teacher preparation can be very different across countries regardless of teacher qualifications, thus making cross-country comparisons potentially misleading. However, cross-country differences are accounted for in a within-student across-teachers model which uses only the variation arising from the teacher preparation relative to the average preparation of teachers teaching in the same class.

[6] In TIMSS, biology, chemistry, physics and earth science are referred to as "domains" to distinguish them from the "subject", science, to which they belong. For simplicity, I refer to these domains as subjects. Each block in the TIMSS booklet contains between 12 and 18 items. The examination time for each student is 90 minutes. For more information concerning the assessment design, see Mullis and Martin (2013).

[7] Therefore, this category also includes teachers who have a doctoral degree or an equivalent degree, who only represent 1.5% of the sample. Excluding them does not have an impact on the results.

[8] The question is formulated as: "During your post-secondary education, what was your major or main area(s) of study?". Among other options, teachers can indicate whether they have a major in biology, physics, chemistry, and earth science, which are the subjects of interest. I will therefore consider a teacher as holding a subject-specific qualification only if she holds a major in the instruction subject.

[9] Teachers can report whether they have a major in education-science and education-general. Using only one of the two majors in the estimations has very little impact on the estimates.

Other variables of interest are the extent to which students like learning a subject, henceforth SLL, or find the teaching engaging, henceforth FTE. TIMSS 2015 provides these subject-specific indicators that are derived from the student questionnaire. The *Student Likes Learning Biology* indicator, for instance, is based on students' agreement with nine statements such as "I enjoy learning biology" or "Biology teaches me how things in the world work". Similarly, the *Students' Views on Engaging Teaching in Biology* indicator is based on ten questions, such as "I know what my teacher expects me to do" or "My teacher does a variety of things to help us learn". I standardize both indicators within subjects, so that they have a mean of 0 and a standard deviation of 1 in each subject. I also standardize student test scores within subject in order to facilitate the interpretation of the coefficients. To reduce measurement error due to the limited number of items in each subject,[10] I aggregate the normalized test scores at the class-subject level.

I impute missing values for control variables using mean imputation at the country-subject level.[11] The percentage of missing values is between 4.8 and 6.1% for all the variables in the analysis. There are no missing values for student test scores. I rescale individual weights provided by TIMSS so that each country has the same weight in the analysis. Weights within countries are therefore not affected. Throughout the analysis, I cluster standard errors at the class level as this is the level of the treatment.

In 2015, 40 countries and 285,119 students participated in the science-8th grade assessment. I select countries where a sizable part of the students is taught by at least two different teachers in the subjects of interest. This tends to be the exception across countries: in 24 out of 40 countries less than 8% of the students are taught science by at least two teachers. I drop all these countries as they contain too few (if any) observations that can be used in the subsequent analysis. I also exclude 6 additional countries[12] for which I am unable to link different teachers to the science subect (s) they teach.[13] In the remaining 10 countries, I exclude cases where students are taught science by only one teacher, where the teacher's characteristics of interest are missing

[10] For example, the individual student test score for physics, which constitutes 25% of the science test, is based on 6 to 9 items.

[11] I only use complete cases with respect to the main teacher variables of interest. Whenever school mean is unavailable, I impute missing values by country mean. Although not reported, the main results are robust to the exclusion of imputed values.

[12] Dubai, United Arab Emirates, Israel, Japan, Korea and the US.

[13] This occurs whenever the variable provided by TIMSS indicating the "Subject Code" of the teacher does not refer to a particular subject but is coded as "Integrated Science".

or where I am unable to link teachers to a specific subject.[14] The final sample consists of 39,827 students and 5,709 teachers in 10 countries: Armenia, England, Georgia, Hungary, Kazakhstan, Lithuania, Malta, Russia, Slovenia and Sweden.

2.2.2 Descriptive Statistics

All countries participating in TIMSS 2015 are reported in Table 2.1 in descending order of performance. Countries that are part of the analysis are in bold. Countries on the left side of the table are above the international median, while those on the right side are below the international median. A large variation in the average score of the considered countries can be observed. The top performer, Slovenia, has an average score of 551 while the average score of Georgia, the lowest in the sample, is 443. This means that the difference between the country with the highest and the country with the lowest test score is larger than one standard deviation. Many of the countries in which science subjects are taught separately are former soviet countries, while this is not the case for most of the other countries participating in TIMSS 2015. Nevertheless, the large variation in average test scores of the countries that are part of the analysis speaks in favor of the external validity of this study.

It is important to keep in mind that TIMSS selects representative samples of the students within countries, which does not necessarily yield a representative sample of teachers. Nevertheless, evidence from TALIS (OECD 2014), an international survey of the teacher population, does not indicate large discrepancies between the teachers included in the descriptive statistics of TIMSS and the population of teachers in a country.[15]

Descriptive statistics of the sample are reported in Table 2.2. The total number of observations (148,751) is given by the student-subject combination. It can be noted that, on average, each student is observed 3.74 times. Students' teachers are highly educated: 91% of the students are taught by teachers who have at least a Bachelor's degree. The share of students taught by teachers who have a Master's degree is 48%. In their report covering 20 years of TIMSS, Mullis, Martin, and Loveless (2016) acknowledge that since 1995, the first year in which TIMSS was conducted, countries have increased the requirements for becoming a teacher.

[14] These cases account for 4% of the sample in the 10 countries.

[15] To verify this, I compare the descriptive statistics of interest for the 13 countries that participated both in TIMSS 2015 (8th grade) and in TALIS 2013.

With an average experience of almost 20 years, the teachers in the sample are considerably older than the average teacher in TIMSS who has around 15 years of experience.[16] It can also be noted that most teachers are female.

The *Home Resources* indicator is a comprehensive measure of the socioeconomic status (SES) of the students. It is based on questions regarding parents' education, number of books at home and number of home study supports available for students (such as an own room or internet connection).

The descriptive statistics by subjects for the main teacher variables of interest are presented in Table 2.3. Physics teachers have, on average, a slightly lower level of education and specialization, while earth science teachers are less likely to have majored in education. Biology teachers are, on average, less experienced and earth science teachers are less likely to have majored in education. It can also be noted that there are fewer observations for chemistry and earth science. This is because students are not tested in subjects that are not taught in the current school year. For example, Swedish students did not take the earth science test. Therefore, only 3 test scores are available for Swedish students. Further descriptive statistics at the country level can be found in Table A2.1 in the Appendix. Overall, the descriptive statistics by subject do not reveal great differences. It is important to highlight that, while substantial differences of teacher characteristics across subjects do not represent a concern for the identification strategy *per se,* they might signal different selection mechanisms for teachers in different science subjects. However, this does not seem to be supported by the data as descriptive statistics by subject do not reveal great differences.

A major threat to the identification strategy arises from subject-specific non-random allocation of teachers and students. With respect to students' socioeconomic status (SES), the literature suggests that the allocation of teachers is unlikely to be random. On the one hand, more wealthy parents try to secure better resources for their children by choosing better schools (Clotfelter, Ladd, and Vigdor 2006).[17] On the other hand, countries try to improve the conditions in disadvantaged schools through

[16] Such a difference is due to the prevalence of countries in which teachers typically work as teachers throughout their entire career. The high average experience might make it harder to capture the effect of experience on students' achievements if it is concentrated in the first years of teachers' careers, as the literature suggests.

[17] There is evidence that in Malta, Russia, Slovenia, and the United Kingdom disadvantaged schools are significantly worse off than advantaged schools in terms of the proportion of teachers with a major in science; the same applies to Georgia with respect to the proportion of fully certified teachers (OECD 2018).

smaller classes or lower student-teacher ratios.[18] While all student background characteristics are held constant in a within-student model, subject-specific non-random allocation of teachers and students might still bias the estimates. However, there is no clear indication that such patterns apply to specific subject. To uncover possible non-random patterns of subject-specific allocation of teachers, I present the relevant average teacher characteristic by subject and the socioeconomic background of the students in Table 2.4. I also provide test statistics for differences in average teacher characteristics between high- and low-SES students. High-SES students are those who are above the median of the *Home Resources* indicator in their respective country. The figures highlight two important patterns in the sample. First, the hypothesis that teachers are not allocated randomly with respect to students' SES is confirmed. In all subjects, low-SES students are on average less likely to be taught by teachers with a Masters' degree but more likely to be taught by teachers who majored in education. Similarly, low-SES students are more likely to be taught by more experienced teachers. All these within-subject differences are highly statistically significant. As for subject-specific qualifications teachers, high-SES students are more likely to be taught by such teachers only in biology and earth science.

The second important pattern is that the differences between the characteristics of teachers of high- and low-SES students always point to the same direction. This suggests that, despite the allocation of teachers with respect to student background characteristics being non-random, it is consistent across subjects. This is relevant since a major threat to identification in a within-student across-teachers model lies in systematic differences in teacher allocation across subjects, a pattern that is not supported by the data.

2.3 Empirical Strategy

As a first step, I estimate the following OLS model including a rich set of controls:

$$A_{icdk} = \beta' T_{cdk} + \gamma' X_{ick} + \delta' C_{cdk} + \tau' S_{ck} + \theta_k + \varepsilon_{icdk} \qquad (2.1)$$

[18] In Georgia, for example, classes in the most disadvantaged schools have, on average, 10 students less than the classes in the most advantaged schools. In Hungary, Malta, Russia and Sweden the classes in disadvantaged schools are also significantly smaller than in advantaged schools. Furthermore, in Georgia, Hungary, Malta and Russia, the student-teacher ratio in the most disadvantaged schools is more than 30% lower than in the most advantaged schools (OECD 2018). However, it has also been shown that increasing the number teachers often comes at the expense of the quality of the teaching staff (Jepsen and Rivkin 2009; Dieterle 2015; OECD 2018).

where A_{icdk} is the achievement of student *i* in class *c* in subject *d* in country *k*, T_{cdk} is the vector of student *i*'s teacher characteristics of interest, X_{ick} is a vector of student subject-invariant variables that control for student and family background, C_{cdk} is a vector of subject-specific variables related to student preferences, instruction time and other teacher traits, S_{ck} is a vector of class-specific variables, such as the number of students or the school location, θ_k is a vector of country fixed effects that accounts for country-specific heterogeneity, and ε_{icdk} is the idiosyncratic error term.

The vector of interest, β, captures the association between teacher characteristics and student achievement. However, unobservable characteristics that are both correlated with student achievement and teacher characteristics might bias the estimates. In the previous section I provided evidence of non-random allocation of teacher characteristics with respect to students' SES. However, such non-random allocation might also occur along other unobserved student dimensions which cannot be accounted for in this model. For instance, teachers with subject-specific qualifications might be systematically assigned to classes with more motivated and better performing students. Therefore, teacher characteristics might still not be allocated randomly conditional on observable student characteristics, which would bias the OLS estimates of the teacher characteristics.

As I observe the results of each student in at least three different subjects, I can eliminate bias due to unobservable student characteristics that do not vary across science subjects. Multiple observations for each student allow me to implement a within-student across-teacher model which controls for unobserved and subject-invariant student traits. The only variation that is left in order to capture the effect of teacher characteristics is the within-student and across-subjects variation. This can be achieved empirically by estimating the following student fixed effects model:

$$A_{icdk} = \beta' T_{cdk} + \delta' C_{cdk} + \mu_i + \mu_d + \varepsilon_{icdk} \qquad (2.2)$$

where A_{icdk} is the achievement of student *i* in class *c* in subject *d* and country *k*, T_{cdk} is the vector of student i's teacher characteristics of interest, namely whether a teacher holds a Master's degree, the years of experience, whether a teacher holds subject-specific qualifications in the subject being taught and whether a teacher majored in education . The vector β captures the parameters of interest. C_{cdk} are subject-specific controls, such as teacher gender and instruction time, which account for the remaining subject-specific heterogeneity. Finally, μ_i and μ_d are student and subject fixed effects, respectively, so that all coefficients are estimated using only

within-student variation, thus controlling for every variable that does not vary across subjects. ε_{icdk} is the idiosyncratic error.

Student fixed effects control for a variety of characteristics that are known to largely affect student achievement, such as socioeconomic status and subject-invariant innate abilities. They also control for all subject-invariant school and class features, such as class size or the school environment. Subject fixed effects eliminate subject-specific test score heterogeneities as well as other unobserved factors that are specific to one subject. For example, they account for the fact that the test might be more difficult on average in one subject or that teachers in one subject might be, on average, better prepared.

Estimates could still be biased if the association between unobservable student and teacher characteristics differs between subjects. This might be the case if physics teachers with subject-specific qualifications were more likely to be placed in a class with more motivated students but the same would not apply to biology teachers. Although this cannot be ruled out entirely, Table 2.4 in the previous section does not indicate different patterns of student-teacher matching across subjects.

The model relies on the assumption that the impact of teacher characteristics is homogenous across subjects. Compared to studies examining different subjects, this analysis relies on a weaker assumption as the multiple outcomes belong to the same field. Furthermore, I provide suggestive evidence that this does not seem to be the case. The OLS analysis in the following section demonstrates that the relationship between teacher characteristics and student achievement is not substantially different across subjects. On the other hand, the fact that the multiple outcomes are so closely related to each other makes it difficult to pin down the actual impact of a single teacher in the taught subject. There is indeed a potential for the impact of a teacher to spill over into adjacent subjects. Furthermore, the amount of variation in outcomes that can be exploited should be a priori smaller as performances in related subjects should not be too different. Therefore, it is likely that this analysis yields conservative estimates of the impact of teacher characteristics on student outcomes.

Student fixed effects also account for general science knowledge and therefore for the impact of characteristics of previous teachers. In fact, it should be kept in mind that students' performance in science is the result of several years of schooling during which students were potentially taught by many different teachers. Furthermore, it is likely that the allocation mechanisms between teachers and students remain in place throughout all years of schooling, which could exacerbate pre-existing differences.

For these reasons, an excessive portion of the variation in student achievement might be attributed to the characteristics of current teachers, leading to a bias in the estimates. By capturing each student's stock of knowledge in the sciences, student fixed effects limit the amount of variation that can be falsely attributed to the current teacher. This might come at the cost of increasing the attenuation bias that is due to the fact that the binary indicators I use are a rough measure of teacher preparation. For all of these reasons, the estimated coefficients should be considered as a lower bound.

2.4 Results

2.4.1 Main Results

OLS results of a model that includes a large set of control variables and country fixed effects to account for country heterogeneity are reported in Table A2.2 in the Appendix. In the pooled regression that includes all science subjects in Column 1, only the major in education is positive and marginally significant. This association is equivalent to 3% of a standard deviation in student achievement. The magnitude of the teacher subject-specific qualifications' coefficient is virtually identical but due to a larger standard error, it is not significant. The results in Columns 2 to 5 are not significant, except for the result for the major in education in Column 3, which is positively and statistically significant. Figures in this table do not show substantial heterogeneity across subjects. However, due to the possible correlation between teacher characteristics and unobservable student traits that might affect students' test scores, OLS estimates are likely to yield biased estimates.

To circumvent such possible bias, I implement the within-student across-teachers model of Eq. (2.2). Results are reported in Table 2.5. In Columns 1 to 4, I present the relationship between teacher characteristics and student science test scores controlling for teacher gender and instruction time once student and subject fixed effects have been accounted for, separately for each characteristic. In Column 5, I include all the teacher characteristics of interest simultaneously. Results underline a positive and significant effect of teacher subject-specific qualifications on student achievement, equivalent to between 1.7%-1.8% of a standard deviation. The magnitude of this coefficient is considerably smaller than the one observed in the OLS model, although the parameter is estimated more precisely. All other characteristics considered do not seem to have a significant impact.

The impact of having majored in education is virtually zero, which suggests that the parameter estimated with the OLS model was substantially biased upwards even after controlling for student background characteristics. The small magnitude of the observed coefficients might also be due to the fraction of total variation that remains in the students' test scores. Once student and subject fixed effects are accounted for, the within-student standard deviation in the test scores is 0.33, or one-third of the standard deviation of the full sample. This can be considered as the amount of variation that can realistically be influenced by teachers, as it already takes into account the impact of important factors such as the socioeconomic status, gender or innate abilities. From this perspective, the observed impact of specialized teachers amounts to 5.1%-5.6% of the within-student standard deviation.[19]

I explore heterogeneities by students' characteristics in Table 2.6.[20] In Columns 1-2, I explore heterogeneities in the impact of teachers according to students' gender. The impact of teacher subject-specific qualifications on female students' test scores is positively significant and is equivalent to 2.2% of a standard deviation, while it is positive but insignificant for male students. Such a difference is sizable but not statistically significant. The impact of experience is positively significant for female students, although the magnitude is rather small and only marginally significant. Similarly, the impact of teachers who have majored in education is marginally significant and negative for males, with a magnitude smaller than 1% of a standard deviation.

As most teachers are female, the higher impact of teacher subject-specific qualifications on female students might be due to positive classroom interactions between female teachers and female students. This is not new to the literature and several studies find that having a female teacher improves female students' educational outcomes (e.g. Dee 2005, 2007; Winters et al. 2013; Gong, Lu, and Song 2018). However, including an interaction term between teacher subject-specific qualifications and teacher gender in Equation (2.2) with female students does not support this interpretation. In fact, the coefficient of the interaction between the teacher subject-specific qualifications and teacher being a female is negative but not significant (not shown).

[19] For consistency with the existing literature, I only consider effects relative to the full standard deviation of the model in the remainder of the paper.

[20] I only report the specifications including all the explanatory variables of interest as there is very little additional value in presenting the bivariate specifications as in Table 2.5.

In Columns 3-4, I divide the sample between low- and high-SES students, i.e. students whose SES is below or above the median in their respective country. Teacher subject-specific qualifications have a positive and significant effect only on students coming from more affluent backgrounds, with an estimated impact of 2.8% of a standard deviation. For teachers with subject-specific qualifications, the difference between the coefficients of the two samples is significant. It is plausible to assume that teachers find an environment better suited for learning in schools attended by high-SES students and can therefore deploy their knowledge more effectively. Furthermore, teachers with subject-specific qualifications might be able to work more efficiently with students who have more subject knowledge from the beginning.[21] This is captured to a large extent by students' SES, with a difference in the average test scores between high- and low-SES students equivalent to 45% of a standard deviation. Although this difference includes current school input, a large part of it is probably due to knowledge accrued before the current school year.

2.4.2 Mediation Analysis

In this section, I explore potential channels through which teacher characteristics affect student achievement. There are two student indicators described in Section 2.2, which capture the extent to which students like learning the subject (SLL) and find the teaching engaging (FTE). As a first step, I include these indicators as additional subject-specific controls in the within-student model with student test scores as the dependent variable.

While including potential channels of the treatments in the regressions might come at the cost of over-controlling, this step ensures that the potential channels are relevant and that there is no omitted variable bias left from remaining subject-specific endogeneity.[22] Results in Table 2.7 do not seem to provide evidence of bias due to the

[21] To substantiate this hypothesis, I also divide the sample between low- and high-achievers, i.e. students whose average science test score is below or above the median science test score in their respective country. The results (not shown) are virtually identical to those obtained when I divide the sample between low- and high-SES students. For high-achieving students, the effect of teacher subject-specific qualifications is positive and significant, while for low-achieving students is positive but not significant. However, dividing the sample between low- and high-achieving students is likely to be endogenous to the treatments. I therefore stick to the previous specification (dividing the sample between low- and high-SES students) as the preferred one.

[22] In fact, one possible remaining concern is that students will perform better in one specific subject simply because they have a preference for it, and, therefore, will enjoy learning it and find the teaching more engaging. Thus, omitting the SLL and FTE indicators might cause an omitted variable bias if, for example, science teachers with subject-specific qualifications tend to be assigned to classes where students have a preference for their subject.

omission of subject-specific controls. The impact of teacher characteristics is in fact robust to the inclusion of these subject-specific indicators. In particular, the impact of teacher subject-specific qualifications remains significant in all specifications but slightly decreases in its magnitude. Both indicators are positively associated with student test scores, but the results should not be interpreted causally.[23] The magnitude of their coefficients is virtually identical when they are included separately (Columns 2 and 3), but the SLL indicator seems to be more relevant for student achievement when both indicators are included (Column 4) in a horse-race regression. While the two indicators are strongly correlated to each other (0.70, p-value < 0.01), the SLL indicator is a clearer indicator of student preferences and thus more suitable to account for this aspect. Conversely, the FTE indicator seems to be better suited as a mediator, as it is more likely to be affected by teacher characteristics.

I explore the role of the SLL and FTE indicators as potential mechanisms in Table 2.8 and 9, where I use them as outcomes of teacher characteristics using the same models of Equation (2.1) and (2.2). I report results from the pooled OLS model with various sets of controls and fixed effects (Columns 1, 2 and 3) as well as from the within-student model (Column 4). The OLS model should not include major biases in this context. In fact, there is no obvious way in which a non-random sorting of teachers and students might be based on students' appreciation for a subject or how engaging they find the teaching. The results seem to support such hypotheses, as the coefficients are virtually identical regardless of the model used. Only the parameter associated to teachers' Master's degrees becomes positive in the within-student specification, but it remains statistically insignificant in all specifications.

The main result illustrated in these tables is that teachers' experience has a clear and significantly negative impact on whether students like the subject or find the teaching engaging. The results are robust to the inclusion of student, teacher, and school controls as well to the inclusion of student fixed effects.[24] The impact of an additional year of experience leads to a decrease in both indicators equivalent to roughly 0.4% of a standard deviation. While this analysis does not provide consistent estimates of the impacts of both the SLL and FTE indicators on student achievement, it is

[23] Reverse causation is likely to be an issue for these controls, as students who perform better in one subject are probably more likely to enjoy the subject and the teaching more. However, a causal analysis of these controls lies outside the scope of this paper.

[24] A separate OLS regression for each subject (not shown) also confirms these results.

reasonable to assume that students will learn more if they are more engaged or enjoy a subject. These are also desirable outcomes *per se*.

The negative impact of teacher experience on the SLL and FTE indicators might help explain a pattern that is frequently discussed in the literature, namely that the largest gain in experience is concentrated in the very first years of teachers' careers.[25] In fact, it is possible that net impact of teacher experience is a combination of factors that improve with increasing experience, such as classroom management or subject knowledge, and other factors that worsen with increasing experience, such as enthusiasm for the subject or for teaching in general. The marginal benefit of an additional year of experience might therefore fade out as the latter factors offset the former ones.

As a final remark, it can be observed that the coefficients of teacher subject-specific qualifications are quite large in both tables, especially in the within-student model, although they never reach statistical significance. Hence, although the point estimates consistently point in the positive direction, I cannot reject the null hypotheses that students do not enjoy learning a subject more nor find the teaching more engaging when taught by specialized teachers.

2.4.3 Robustness Checks

To ensure that results are not driven by single countries, where, for example, teachers with subject-specific qualifications are particularly effective compared to teachers

[25] This suggests a non-linear relationship between student test scores and teacher experience. To explore this aspect, I implement several non-linear specifications of experience in the within-student model of Equation (2.2) with science test scores as outcome, namely experience squared, logarithm of experience and a piecewise specification, (i.e. having 2, 3-5 or 6 or more years of experience). However, the impact of teacher experience is not significant in any these specifications. As a further step, I restrict the sample to the youngest cohort of teachers (25 years or less) or teachers with less than 4 years of experience. Although the resulting samples are too small to draw reliable conclusions (1,024 and 4,028 observations, respectively), the impact of teacher experience is positively significant in this context, with a magnitude between 1.9-4.2% of a standard deviation for one additional year of experience. The positive impact disappears when the second-to-youngest cohort is included (25 to 30 years old teachers) or teachers with less than 5 years of experience are included, thus suggesting diminishing marginal returns to experience. This is also shown in Boyd et al. 2008, who report gains between 5-7% of a standard deviation during the first year of experience, with these gains accounting for more than half of the cumulative experience effect.

without them, I repeat the analysis excluding one country at a time.[26] The results are reported in Table A2.3. While the effect remains largely positive in all columns, it does not reach any conventional level of statistical significance when Malta, Slovenia or Sweden are excluded (Column 7,9 and 10). On the other hand, excluding Hungary yields the largest estimate of the impact of teacher subject-specific qualifications, suggesting that they are not particularly effective in this country.[27]

When Armenia, Hungary or Lithuania are excluded (Columns 1, 4 and 6), the results for the major in education become marginally significant and negative, with a point estimate of around 1.1% of a standard deviation. Overall, the coefficient for the major in education always points to the negative direction in the within-student models with student test scores as outcome variable. A possible interpretation for this is that pedagogical and subject-specific knowledge are substitutes in the preparation of teachers. In fact, the correlation between being a specialized teacher and having majored in education is significantly negative (-0.29, p-value < 0.01). Therefore, the major in education might also be capturing the effect of a lower level of subject knowledge.

I also perform a further robustness check in which I omit one subject at a time. Table A2.4 shows that the impact of teacher subject-specific qualifications is stronger when earth science and especially physics are dropped. This suggests that teachers with teacher subject-specific qualifications are less effective in these subjects. Conversely, their impact fades when biology is excluded from the analysis, indicating that the effect is driven by biology teachers. A possible explanation for this comes from the design of the test. As described in Section 2.2, biology constitutes the largest part of the science test (35%). Therefore, test scores in this subject should be considered more reliable and less noisy than test scores in other subjects. Omitting biology from the within-student model might therefore leave only test scores that are too noisy to detect a relatively small effect.

[26] In principle, it is possible to run a separate regression for each single country. However, some countries contribute very little to the identification due to very large or small shares of the variables of interest (e.g. only 3% of the teachers in Kazakhstan have a Master's degree). Thus, single-country regressions might not be particularly informative.

[27] Hungary is the country with the lowest share of teachers with subject-specific qualifications (26% of the teachers hold these, see Table A2.1 in the Appendix). This might suggest that content knowledge is not a priority in the training of lower secondary science teachers in this country. Nevertheless, the overall performance of Hungarian students in science is well above the international TIMSS average, suggesting that other factors contribute to a country's students achieving good results.

As also observed when omitting some countries, the coefficient for the major in education becomes significantly negative when physics and, in particular, biology are dropped. Again, this might be due to the fact that a major in education might capture part of the effect of lower subject knowledge.

2.5 Conclusion

It is widely acknowledged that teachers play a fundamental part in student education and that education systems worldwide should strive to ensure teacher quality. Nevertheless, what constitutes teacher quality remains relatively unresolved. Available teacher characteristics such as education and experience tend to be weak predictors of teachers' effectiveness. This paper complements previous studies using within-student across-subject analyses in that it focuses exclusively on science achievement in a group of countries in which 8th graders are taught sciences by different teachers.

The main result of the analysis is that science teachers who are hold subject-specific qualifications in the subject that they teach have a positive and significant impact on students' science performance, while neither having a Master's degree nor holding a major in education or the number of years of experience have a significant impact on students' performance. This result confirms that subject knowledge tends to be a stronger predictor of teacher effectiveness than, for example, the general education level or experience. A related policy implication is that subject knowledge should play a key role in the recruitment and compensation of teachers in lower secondary schools. Furthermore, the benefit of teacher subject-specific qualifications could be reaped at no additional cost by allocating science teachers according to their specialization.

In the mediation analysis, I find that teacher experience negatively affects the indicators that measure how much students like a subject and find the teaching engaging. This result might help to explain a pattern which has often been observed in the literature, namely that most of the gains from teaching experience in terms of student performance seem to be concentrated in the very first years of the teaching career. A possible implication of this result is that teachers should be incentivized to update their teaching methods throughout their career in order to keep their students engaged.

Tables

Table 2.1: Average Science Score in TIMSS 2015, Entire Sample

Country	Average Scale Score	(SE)	Country	Average Scale Score	(SE)
Singapore	597	(3.2)	Turkey	493	(4.0)
Japan	571	(1.8)	**Malta**	**481**	**(1.6)**
Chinese Taipei	569	(2.1)	United Arab Emirates	477	(2.3)
Korea, Rep. of	556	(2.2)	Malaysia	471	(4.1)
Slovenia	**551**	**(2.4)**	Bahrain	466	(2.2)
Hong Kong SAR	546	(3.9)	Qatar	457	(3.0)
Russian Federation	**544**	**(4.2)**	Iran, Islamic Rep. of	456	(4.0)
England	**537**	**(3.8)**	Thailand	456	(4.2)
Kazakhstan	**533**	**(4.4)**	Oman	455	(2.7)
Ireland	530	(2.8)	Chile	454	(3.1)
United States	530	(2.8)	**Armenia***	**452**	**(-)**
Hungary	**527**	**(3.4)**	**Georgia**	**443**	**(3.1)**
Canada	526	(2.2)	Jordan	426	(3.4)
Sweden	**522**	**(3.4)**	Kuwait	411	(5.2)
Lithuania	**519**	**(2.8)**	Lebanon	398	(5.3)
New Zealand	513	(3.1)	Saudi Arabia	396	(4.5)
Australia	512	(2.7)	Morocco	393	(2.5)
Norway (9)	509	(2.8)	Botswana (9)	392	(2.7)
Israel	507	(3.9)	Egypt	371	(4.3)
Italy	499	(2.4)	South Africa (9)	358	(5.6)

Note: The figure has been obtained from TIMSS 2015 8th grade Science Achievement. Standard errors of the average country science achievement are in parentheses. Countries that are part of the analyzed sample are in bold. *Armenia took the test one year later and was not included in the original

Table 2.2: Descriptive Statistics

	Mean	SD	Min	Max
Bachelors' Teachers	0.43	0.49	0.0	1.0
Masters' Teachers	0.48	0.49	0.0	1.0
Experience (y)	19.90	11.18	0.0	45.0
Subject-Specific Qual. Teachers	0.83	0.36	0.0	1.0
Major in Education	0.49	0.49	0.0	1.0
Female Teachers	0.80	0.39	0.0	1.0
Instruction Time (h)	1.58	0.71	0.0	10.0
Home Resources	10.73	1.54	4.2	13.9
# Observations		148,751		
# Students		39,827		
# Teachers		5,709		

Note: The unit of observation is given by the student-subject combination. The table reports weighted descriptive statistics of the main variables of interest. Bachelors' Teachers hold only a Bachelors' degree, while Masters' Teachers also hold a Masters' degree. Experience is measured in years. Subject-specific qualification teachers are those who have a major in their instruction subjects. The Home Resources indicator provided by TIMSS captures the socioeconomic status of the students and is based on parents' education, number of books at home and home study supports available for students.

Table 2.3: Descriptive Statistics by Subject

Variables	Physics		Biology	
	Mean	(SD)	Mean	(SD)
Bachelors' Teachers	0.45	(0.50)	0.43	(0.50)
Masters' Teachers	0.45	(0.49)	0.48	(0.49)
Experience (y)	20.23	(11.56)	18.95	(11.11)
Subject-Specific Qual. Teachers	0.80	(0.39)	0.85	(0.35)
Major in Education	0.50	(0.48)	0.53	(0.48)
Instruction Time (h)	1.73	(0.80)	1.52	(0.69)
# Students	39,169		38,069	
# Teachers	1,722		1,710	

Variables	Chemistry		Earth Science	
	Mean	(SD)	Mean	(SD)
Bachelors' Teachers	0.42	(0.49)	0.41	(0.49)
Masters' Teachers	0.49	(0.49)	0.51	(0.49)
Experience (y)	19.90	(10.90)	20.59	(11.03)
Subject-Specific Qual. Teachers	0.82	(0.38)	0.87	(0.33)
Major in Education	0.52	(0.49)	0.39	(0.48)
Instruction Time (h)	1.60	(0.63)	1.46	(0.64)
# Students	37,487		33,896	
# Teachers	1,636		1,360	

Note: The table reports weighted descriptive statistics by subject. For each subject, the number of distinct students and teachers observed is also reported.

Table 2.4: Teacher Characteristics by Subject and Student SES

	Physics		Biology		Chemistry		Earth Science	
	Low-SES	High-SES	Low-SES	High-SES	Low-SES	High-SES	Low-SES	High-SES
Masters' Teachers	0.44	0.47	0.47	0.49	0.49	0.50	0.49	0.55
t-test statistic	*(4.47)* ***		*(5.40)* ***		*(3.28)* ***		*(11.07)* ***	
Experience (y)	20.66	19.61	19.30	18.44	19.95	19.83	20.77	20.31
t-test statistic	*(-8.86)* ***		*(-7.46)* ***		*(-1.00)*		*(-3.78)* ***	
Subject-Specific Qual. Teachers	0.80	0.80	0.84	0.87	0.81	0.82	0.86	0.89
t-test statistic	*(0.34)*		*(7.53)* ***		*(1.59)*		*(6.66)* ***	
Major in Education	0.50	0.49	0.54	0.53	0.53	0.50	0.39	0.39
t-test statistic	*(-2.16)* **		*(-2.49)* **		*(-6.61)* ***		*(-0.82)*	

Note: The table reports the weighted means of the main independent variables of interest by student SES and subject. High-SES students are students who fall above the median SES level within their country. For each variable, I report the t-statistic associated with the difference in the means between High- and Low-SES students. *** p<0.01, ** p<0.05, * p<0.1.

Table 2.5: The Effect of Teacher Characteristics on Students' Test Scores

	(1)	(2)	(3)	(4)	(5)
Masters' Teachers	0.0011				0.0015
	(0.0057)				*(0.0057)*
Experience		0.0003			0.0002
		(0.0002)			*(0.0002)*
Subject-Specific Qual. Teachers			0.0182**		0.0172*
			(0.0088)		*(0.0089)*
Major in Education				-0.0088	-0.0076
				(0.0054)	*(0.0055)*
Observations	148,751	148,751	148,751	148,751	148,751
Students, Subject FE	YES	YES	YES	YES	YES

*Note:*The table reports the results for the within-student across-teachers model that includes four science subjects (physics, biology, chemistry, earth science). The number of observations is given by all the student-subject combinations. All specifications control for instruction time and teacher gender and include student and subject fixed effects. Test scores have been standardized within subjects and aggregated at the classroom-subject level to reduce measurement error. Standard errors are clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1.

Table 2.6: Main Results by Gender and SES

	Student Gender		SES	
	Male	Female	Low-SES	High-SES
	(1)	(2)	(3)	(4)
Masters' Teachers	-0.0029	0.0056	-0.0004	0.0052
	(0.0059)	*(0.0061)*	*(0.0061)*	*(0.0065)*
Experience	0.0001	0.0005*	0.0002	0.0003
	(0.0002)	*(0.0003)*	*(0.0003)*	*(0.0003)*
Subject-Specific Qual. Teachers	0.0106	0.0224**	0.0115	0.0280**
	(0.0087)	*(0.0104)*	*(0.0086)*	*(0.0118)*
Major in Education	-0.0097*	-0.0049	-0.0077	-0.0063
	(0.0058)	*(0.0058)*	*(0.0060)*	*(0.0061)*
Observations	76,350	72,401	85,538	63,213
Students, Subject FE	YES	YES	YES	YES

Note: The table reports the results for the within-student across-teachers model that includes four science subjects (physics, biology, chemistry, earth science). The number of observations is given by all the student-subject combinations. All specifications control for instruction time and teacher gender and include student and subject fixed effects. Each column reports the estimated coefficient in the indicated sub-sample. High-SES students are those above the median SES level within their country. Test scores have been standardized within subjects and aggregated at the classroom-subject level to reduce measurement error. Standard errors are clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table 2.7: Additional Subject-Specific Controls

	(1)	(2)	(3)	(4)
Masters' Teachers	0.0015	0.0014	0.0015	0.0014
	(0.0057)	*(0.0056)*	*(0.0056)*	*(0.0056)*
Experience	0.0002	0.0003	0.0003	0.0003
	(0.0002)	*(0.0002)*	*(0.0002)*	*(0.0002)*
Subject-Specific Qual. Teachers	0.0172*	0.0168*	0.0168*	0.0167*
	(0.0089)	*(0.0088)*	*(0.0088)*	*(0.0088)*
Major in Education	-0.0076	-0.0078	-0.0079	-0.0078
	(0.0055)	*(0.0054)*	*(0.0054)*	*(0.0054)*
SLL		0.0139***		0.0105***
		(0.0013)		*(0.0013)*
FTE			0.0135***	0.0059***
			(0.0016)	*(0.0016)*
Observations	148,751	148,751	148,751	148,751
Students, Subject FE	YES	YES	YES	YES

Note: The table reports the results for the within-student across-teachers model that includes four science subjects (physics, biology, chemistry, earth science). The number of observations is given by all the student-subject combinations. SLL stands for the *Students Like Learning* indicator, while FTE stands for the *Students Find the Teaching Engaging* indicator. I have standardized the SLL and FTE indicators to have mean 0 and standard deviation 1 within each subject. All specifications control for instruction time and teacher gender and include student and subject fixed effects. Test scores have been standardized within subjects and aggregated at the classroom-subject level to reduce measurement error. Standard errors are clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table 2.8: Teacher Characteristics and the *Student Likes Learning* Indicator

	OLS (1)	OLS (2)	OLS (3)	Within- (4)
Masters' Teachers	-0.0011	-0.0155	-0.0076	0.0116
	(0.0152)	*(0.0148)*	*(0.0148)*	*(0.0172)*
Experience	-0.0034***	-0.0032***	-0.0031***	-0.0038***
	(0.0006)	*(0.0006)*	*(0.0006)*	*(0.0006)*
Subject-Specific Qual. Teachers	0.0298	0.0253	0.0298	0.0262
	(0.0240)	*(0.0230)*	*(0.0221)*	*(0.0214)*
Major in Education	0.0140	0.0118	0.0130	0.0136
	(0.0144)	*(0.0139)*	*(0.0137)*	*(0.0151)*
Observations	148,751	148,751	148,751	148,751
R^2	0.0887	0.1151	0.1185	0.5593
Country FE	YES	YES	YES	NO
Student Controls	NO	YES	YES	NO
Class, School Controls	NO	NO	YES	NO
Student, Subject FE	NO	NO	NO	YES

Note: The table reports the results for an OLS model (Column 1,2,3) and a within-student across-teachers model (Column 4) that include four science subjects (physics, biology, chemistry, earth science). The number of observations is given by all the student-subject combinations. The dependent variable is the "Student Likes Learning the Subject" indicator standardized within subjects. Student controls are student SES, gender, language spoken at home, whether parents have foreign origins and expectations in educational achievement. Class controls are class size, share of students with language difficulties, class SES and the share of native speakers. School controls are the school location, whether science instruction is hindered by shortage of resources, school discipline problems and school emphasis on academic success. Subject-specific controls are teacher gender and instruction time. Country fixed effects are included in Columns 1-3, student and subject fixed effects are included in Column 4. Standard errors are clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table 2.9: Teacher Characteristics and the *Student Finds the Teaching Engaging* Indicator

	OLS (1)	OLS (2)	OLS (3)	Within- (4)
Masters' Teachers	-0.0239	-0.0333**	-0.0247	0.0045
	(0.0165)	*(0.0164)*	*(0.0161)*	*(0.0172)*
Experience	-0.0040***	-0.0039***	-0.0039***	-0.0039***
	(0.0006)	*(0.0006)*	*(0.0006)*	*(0.0006)*
Subject-Specific Qual. Teachers	0.0026	0.0006	0.0054	0.0243
	(0.0247)	*(0.0243)*	*(0.0233)*	*(0.0217)*
Major in Education	0.0248*	0.0230	0.0249*	0.0227
	(0.0147)	*(0.0146)*	*(0.0143)*	*(0.0153)*
Observations	148,751	148,751	148,751	148,751
R^2	0.1058	0.1179	0.1236	0.6388
Country FE	YES	YES	YES	NO
Student Controls	NO	YES	YES	NO
Class, School Controls	NO	NO	YES	NO
Student, Subject FE	NO	NO	NO	YES

Note: The table reports the results for an OLS model (Column 1,2,3) and a within-student across-teachers model (Column 4) that include four science subjects (physics, biology, chemistry, earth science). The number of observations is given by all the student-subject combinations. The dependent variable is the "Student Finds the Teaching Engaging" indicator standardized within subjects. Student controls are student SES, gender, language spoken at home, whether parents have foreign origins and expectations in educational achievement. Class controls are class size, share of students with language difficulties, class SES and the share of native speakers. School controls are the school location, whether science instruction is hindered by shortage of resources, school discipline problems and school emphasis on academic success. Subject-specific controls are teacher gender and instruction time. Country fixed effects are included in Columns 1-3, student and subject fixed effects are included in Column 4. Standard errors are clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Appendix

Table A2.1: Descriptives by Country

	Armenia		England		Georgia		Hungary		Kazakhstan	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Students' Science	452.4	(104.43)	568.92	(85.64)	437.54	(96.75)	526.49	(95.27)	530.15	(106.29)
Bachelors' Teachers	0.13	(0.34)	0.62	(0.49)	0.09	(0.29)	0.65	(0.48)	0.93	(0.25)
Masters' Teachers	0.79	(0.38)	0.25	(0.39)	0.89	(0.31)	0.33	(0.46)	0.03	(0.17)
Experience (y)	22.96	(10.51)	12.83	(9.37)	22.39	(11.29)	23.23	(10.20)	19.38	(11.22)
Subject-Specific Qual. Teachers	0.96	(0.18)	0.78	(0.38)	0.96	(0.19)	0.26	(0.43)	0.97	(0.18)
Major in Education	0.29	(0.43)	0.53	(0.46)	0.39	(0.48)	0.86	(0.34)	0.25	(0.43)
Instruction Time (h)	1.72	(0.44)	-		1.69	(0.65)	1.39	(0.61)	1.77	(0.7)
# Students	5,002		819		4,035		4,893		4,887	
# Teachers	588		224		645		599		791	

	Lithuania		Malta		Russia		Slovenia		Sweden	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Students' Science	516.38	(84.19)	502.62	(112.92)	543.93	(87.73)	553.42	(82.95)	518.78	(91.4)
Bachelors' Teachers	0.55	(0.5)	0.7	(0.46)	0.24	(0.43)	0	(0.06)	0.5	(0.5)
Masters' Teachers	0.41	(0.48)	0.22	(0.4)	0.74	(0.43)	0.61	(0.48)	0.38	(0.47)
Experience (y)	24.38	(10.19)	10.99	(7.98)	22.95	(11.05)	21.98	(10.17)	12.57	(8.37)
Subject-Specific Qual. Teachers	0.95	(0.22)	0.91	(0.27)	0.97	(0.16)	0.93	(0.25)	0.63	(0.46)
Major in Education	0.55	(0.48)	0.52	(0.48)	0.53	(0.5)	0.22	(0.4)	0.77	(0.39)
Instruction Time (h)	1.45	(0.65)	2.19	(1.25)	1.58	(0.43)	1.45	(0.53)	1.13	(0.45)
# Students	4,347		2,756		4,780		4,257		4,051	
# Teachers	904		335		749		572		302	

Note: Each column reports weighted descriptive statistics by country. The number of distinct students and teachers are also reported.

Table A2.2: OLS Regressions

	All	Physics	Biology	Chemistry	Earth
	(1)	(2)	(3)	(4)	(5)
Masters' Teachers	0.0133	0.0163	-0.00587	0.0161	0.0238
	(0.0169)	*(0.0253)*	*(0.0246)*	*(0.0266)*	*(0.0313)*
Experience	0.000820	-0.00189	0.00114	0.00189	-0.000109
	(0.000705)	*(0.00121)*	*(0.00110)*	*(0.00121)*	*(0.00122)*
Subject-Specific Qual. Teachers	0.0297	-5.73e-05	0.0382	-0.0131	0.0663
	(0.0239)	*(0.0362)*	*(0.0337)*	*(0.0434)*	*(0.0521)*
Major in Education	0.0304*	-0.0170	0.0585**	0.0313	0.0532
	(0.0182)	*(0.0254)*	*(0.0268)*	*(0.0290)*	*(0.0325)*
Observations	148,751	39,193	38,070	37,555	33,933
R^2	0.451	0.478	0.481	0.455	0.514
Country FE	YES	YES	YES	YES	YES
Student, Class, School Controls	YES	YES	YES	YES	YES

Notes: Each column includes an OLS regression for the specified subjects. Column 1 includes all subjects. All specifications include country fixed effects, student, subject-specific, class and school controls. Student controls are student SES, gender, language spoken at home, whether parents have foreign origins and expectations in educational achievement. Subject-specific controls are teacher gender, whether students enjoy learning the subject, find the teaching engaging and instruction time. Class controls are class size, share of students with language difficulties, class SES and the share of native speakers. School controls are the school location, whether science instruction is hindered by shortage of resources, school discipline problems and school emphasis on academic success. Test scores have been standardized within subjects and aggregated at the class-subject level to reduce measurement error. Standard errors are clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A2.3: Leave-One-Country-Out

	Armenia (1)	England (2)	Georgia (3)	Hungary (4)	Kazakhstan (5)	Lithuania (6)	Malta (7)	Russia (8)	Slovenia (9)	Sweden (10)
Masters' Teachers	-0.0012	-0.0011	0.0021	0.0015	0.0013	0.0026	0.0039	0.0039	-0.0009	0.0023
	(0.0056)	*(0.0058)*	*(0.0059)*	*(0.0064)*	*(0.0056)*	*(0.0060)*	*(0.0060)*	*(0.0061)*	*(0.0065)*	*(0.0059)*
Experience	0.0003	0.0002	0.0003	0.0003	0.0001	0.0003	0.0001	0.0002	0.0002	0.0003
	(0.0003)	*(0.0002)*	*(0.0003)*	*(0.0003)*	*(0.0002)*	*(0.0003)*	*(0.0002)*	*(0.0003)*	*(0.0003)*	*(0.0002)*
Subject-Specific Qual. Teachers	0.0146*	0.0153*	0.0250***	0.0324***	0.0187**	0.0165*	0.0110	0.0165*	0.0113	0.0115
	(0.0087)	*(0.0092)*	*(0.0092)*	*(0.0112)*	*(0.0082)*	*(0.0095)*	*(0.0091)*	*(0.0091)*	*(0.0094)*	*(0.0099)*
Major in Education	-0.0090*	-0.0078	-0.0091	-0.0112**	-0.0074	-0.0117**	-0.0059	-0.0097	0.0002	-0.0043
	(0.0054)	*(0.0057)*	*(0.0060)*	*(0.0057)*	*(0.0055)*	*(0.0058)*	*(0.0059)*	*(0.0060)*	*(0.0059)*	*(0.0054)*
Observations	129,058	146,286	132,975	129,253	129,277	131,506	142,003	129,908	131,877	136,616
Students, Subject FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The table reports the results for the within-student across-teachers model that includes four science subjects (physics, biology, chemistry, earth Science). The number of observations is given by all the student-subject combinations. The country indicated in each column has been dropped for the estimation. All specifications control for instruction time and teacher gender and include student and subject fixed effects. Test scores have been standardized within subjects and aggregated at the classroom-subject level to reduce measurement error. Standard errors are clustered at the classroom level. ***p<0.01, **p<0.05, *p<0.1

Table A2.4: Leave-One-Subject-Out

	All	Physics	Biology	Chemistry	Earth
	(1)	(2)	(3)	(4)	(5)
Masters' Teachers	0.0015	0.0031	0.0019	0.0028	-0.0007
	(0.0057)	*(0.0079)*	*(0.0072)*	*(0.0064)*	*(0.0064)*
Experience	0.0002	0.0004	0.0004	0.0000	0.0003
	(0.0002)	*(0.0003)*	*(0.0003)*	*(0.0003)*	*(0.0003)*
Subject-Specific Qual. Teachers	0.0172*	0.0312**	-0.0004	0.0148	0.0206**
	(0.0089)	*(0.0131)*	*(0.0111)*	*(0.0107)*	*(0.0091)*
Major in Education	-0.0076	-0.0144*	-0.0195***	0.0029	0.0014
	(0.0055)	*(0.0076)*	*(0.0068)*	*(0.0065)*	*(0.0067)*
Observations	148,751	107,779	110,377	111,042	113,247
Students, Subject FE	YES	YES	YES	YES	YES

Note: The table reports the results for the within-student across-teachers model. The number of observations is given by all the student-subject combinations. In Column 1, all the subjects are included. In Columns 2-5, the indicated subject has been dropped for the estimation. All specifications control for instruction time and teacher gender and include student and subject fixed effects. Test scores have been standardized within subjects and aggregated at the classroom-subject level to reduce measurement error. Standard errors are clustered at the classroom level. ***p<0.01, **p<0.05, *p<0.1

3 The Effect of Teacher Subject-Specific Qualifications on Student Science Achievement*

3.1 Introduction

What makes a good teacher? This question has been at the center of a large literature spanning several decades. Although a definitive answer remains elusive, a consensus has seemingly emerged on some facts. Many studies have shown that generic teacher qualifications, such as teacher degree level, advanced degrees, or certification status are not good predictors of teacher quality (Hanushek 1986; Rivkin, Hanushek, and Kain 2005; Clotfelter, Ladd, and Vigdor 2007; Buddin and Zamarro 2009; Staiger and Rockoff 2010; Ladd and Sorensen 2015). Conversely, subject-specific teacher qualifications tend to better predict teacher quality (Monk and King 1994; Goldhaber and Brewer 1997, 2000; Croninger et al. 2007; Clotfelter, Ladd, and Vigdor 2010), although findings in this field are more mixed and less abundant.

Yet, a striking feature of this literature calls for caution when interpreting results: the vast majority of studies uses US data. A recent survey of high-quality studies from 2003 to 2018 investigating the effect of any teacher characteristics on student scores features no studies investigating teacher subject-specific qualifications outside the US (Coenen et al. 2018).[1] A concurrent survey of the literature on teacher effectiveness and student outcomes highlights the same issue, thus questioning the extent to which existing evidence applies to other contexts (Burroughs et al. 2019). As teacher education programs vary greatly from country to country (Blömeke, Kaiser, and Lehmann 2010; Tatto et al. 2012), policymakers worldwide should be wary of existing evidence when devising policies concerning teachers. This deficit of evidence is even more critical for developing countries, which likely benefit the most from improving student achievement (Hanushek and Woessmann 2015).

* This chapter is based on the paper "The Effect of Teacher Subject-Specific Qualifications on Student Science Achievement", *Labour Economics*, 2023.

[1] Among the reviewed studies, only 9 out of the 58 studies considered do not use US data. Further, a previous review of this literature for the period until 2003 only covered US studies. The rationale for doing so was that the authors were aware of only one study not conducted in the US (Wayne and Youngs 2003).

Chapter 3: Teacher Subject-Specific Qualifications

In this paper, I investigate the impact of subject-specific teacher qualifications—as captured by teachers holding a major in science subjects—on student science test scores in an international setting. In most contexts, estimating the impact of teacher characteristics on student outcomes is challenging. Non-random assignment of teachers to students as well as unobservable student and teacher characteristics are the most obvious concerns from an econometric standpoint. I tackle these issues in a novel way by using within-student within-teacher across-subjects variation. I exploit the availability of test scores and teacher qualifications in four science subjects—biology, chemistry, physics, and earth science—available for each 8th-grade student participating in the *Trends in Mathematics and Science Study 2015* (TIMSS 2015). I focus on 30 countries where science is taught as an integrated subject, namely where all science subjects are taught by the same teacher, which constitute most of the countries in TIMSS 2015. Estimates obtained using the within-student within-teacher variation are not biased by unobserved student or teacher characteristics that do not vary across subjects, thus mitigating the most serious sources of bias.

I find that teacher subject-specific qualifications have a positive and statistically significant impact on student test scores. The magnitude of the impact is equivalent to 3.5% of a standard deviation (SD). Putting this figure into perspective, evidence from the US links an increase in teacher effectiveness by one SD to an increase in student math achievement by 20% SD (Jackson, Rockoff, and Staiger 2014). If the variation in teacher effectiveness in the international sample that I use is similar to that in the US, teacher subject-specific qualifications would explain approximately 17.5% of the variation in teacher effectiveness. Similarly, a student would gain approximately $6,825 on average in cumulative lifetime income from being taught by a teacher with subject-specific qualifications in a single grade.[2] Compared to other educational inputs, the effect of teacher subject-specific qualifications is equivalent to an increase of 2 hours and 10 minutes of weekly classroom instruction.[3]

[2] I obtain this figure by multiplying the average gain in cumulative lifetime income from a one SD improvement in teacher value-added in a single grade ($39,000) calculated in Chetty, Friedman, and Rockoff (2014) in the US, by the share of teacher value-added "explained" by teacher subject-specific qualifications (17.5%).

[3] This estimate is obtained by dividing the estimated effect of teacher subject-specific qualifications (3.5% SD) by the average of the impact of a one-hour increase in weekly instruction time on student test scores (1.6% SD) computed in Bietenbeck and Collins (2023) using six waves of TIMSS and PISA data, weighted by the number of countries in each wave.

Heterogeneity analyses reveal that the impact is stronger for female students, especially when they are taught by female teachers, and for students with a lower socio-economic status (SES). Concerning teacher characteristics, the impact of teacher subject-specific qualifications is stronger for teachers who also hold a major in education and follows a concave path with respect to years of teacher experience. The analysis of cross-country heterogeneities suggests that students in lower-achieving countries benefit more from being taught by teachers with subject-specific qualifications. These findings, together with the previously described larger impact of teacher with subject-specific qualifications for students with lower SES, suggest that students in more disadvantaged contexts might benefit the most from having such teachers. To shed light on the possible mechanisms through which teachers with subject-specific qualifications affect student achievement, I conduct a mediation analysis. I find that up to 20% of the impact of subject-specific qualifications is explained by teachers being more confident to teach subjects in which they hold a major.

This paper contributes to the literature of the impact of teacher characteristics on student test scores in three ways. First, it contributes to the existing evidence on subject-specific teacher qualifications as a determinant of student achievement in an international setting. Previous studies have generally found positive effects of subject-specific teacher qualifications on student test scores, especially for math (Monk and King 1994; Goldhaber and Brewer 1997, 2000; Clotfelter, Ladd, and Vigdor 2010), although other studies have not found any effect (Aaronson, Barrow, and Sander 2007; Harris and Sass 2011).[4] However, all the evidence in this field comes from studies conducted in the US. I enrich this literature by providing first evidence that teacher subject-specific qualifications positively affect student test scores in an international setting. Further, I find a stronger effect in developing and low-performing countries. This result suggests that the current consensus of the literature on teacher qualifications may underestimate the benefits that teacher qualifications bring to students around the world. Nudging teachers to acquire subject-specific qualifications is therefore likely to be beneficial for countries worldwide, especially in developing countries.

[4] A related strand of this literature has focused on teacher subject knowledge measured with subject-specific test scores rather than qualifications, showing that these are a consistent determinant of student test scores, especially in math (Clotfelter, Ladd, and Vigdor 2007; Boyd et al. 2008; Kukla-Acevedo 2009; Metzler and Woessmann 2012), and also in international settings (Bietenbeck, Piopiunik, and Wiederhold 2018; Hanushek, Piopiunik, and Wiederhold 2019).

Second, I analyze the impact of subject-specific teacher qualifications in a novel way by using a within-student within-teacher across-subjects approach, with subjects belonging to the same field. Much like the more commonly used within-student across-subjects approach[5], it accounts for subject-invariant student characteristics that are known to affect student achievement, such as student ability or socioeconomic background. However, it has the additional advantage of holding constant any teacher characteristics that do not differ across subjects.[6] Further, to the best of my knowledge, this is the first study that applies this approach in a context where the subjects belong to the same field, i.e., science, as opposed to different fields, such as math and reading. A key assumption of all the approaches that exploit within-student across-subjects variation is that unobserved sources of subject-specific student or teacher heterogeneity do not bias the estimates. Given the relatedness of the subjects, this assumption is more likely to hold in this case.

The third contribution of this paper is that I focus exclusively on an important yet understudied subject: science. In the recent survey of the effect of teacher characteristics on student test scores by Coenen et al. (2018), science was among the subjects analyzed in only 11 of the 58 reviewed studies, while the majority of studies focused on math and/or reading. The lack of interest in science is at odds with the current educational and political debate. Calls to nurture science skills in school to address the need for employees with a STEM background and for a scientifically literate public have been pervasive in the last decade (Carnevale, Smith, and Melton 2011; President's Council of the Advisors on Science and Technology 2012; OECD 2016c; European Commission: DG Employment Social Affairs and Inclusion 2020). The literature has also shown that the impact of teacher qualifications on student test scores varies across subjects. For example, the US study by Clotfelter, Ladd, and Vigdor (2010) finds that the effects of teacher subject-specific certifications are, on average, positive, but very heterogeneous. Test scores of students taught by teachers with math or English certification are 11% SD and 10% SD higher, respectively, while it finds no effect for biology. Similarly, Monk and King (1994) and Goldhaber and

[5] The within-student across-subjects approach has been used extensively in the literature to study the impact of teacher characteristics (Clotfelter, Ladd, and Vigdor 2010; Bietenbeck, Piopiunik, and Wiederhold 2018; Hanushek, Piopiunik, and Wiederhold 2019; Sancassani 2021) as well as other educational inputs, such as instruction time (Lavy 2015; Wedel 2021; Bietenbeck and Collins 2023) or teaching practices (Bietenbeck 2014) on student outcomes.

[6] Among the studies investigating the impact of teacher qualifications on student test scores, only Harris and Sass (2011) includes one specification with teacher fixed effects. However, they exploit within-teacher variation over time rather than over subjects.

Brewer (2000) find that teacher subject-specific qualifications have a positive impact on student math test scores, but little or no effect in science.[7] Harris and Sass (2011) does not find evidence of the impact of teacher subject knowledge in math and reading acquired through undergraduate coursework on students' math or reading test scores, but it speculates that in other areas, such as science in secondary school, teacher subject knowledge may be a determinant of student test scores. I provide evidence in favor of this hypothesis.

The remainder of the paper is structured as follows: Section 3.2 describes the data and provides some descriptive statistics. Section 3.3 presents the estimation strategy. The main results, heterogeneities, international evidence, and robustness checks are discussed in Section 3.4. The mediation analysis is discussed in Section 3.5. Section 3.6 concludes.

3.2 Data and Descriptive Statistics

3.2.1 TIMSS 2015 and Sample Construction

I use data from TIMSS 2015, an international large-scale assessment of math and science skills of 4th- and 8th-grade students, which was the latest wave available at the start of this project. I replicate my main results also using data from the previous TIMSS wave, namely TIMSS 2011. TIMSS includes mathematics and science questions aimed at measuring students' grade-specific curriculum knowledge and a rich set of background questionnaires about students, teachers and schools that gather information about the educational and social contexts of students. The grade-specific focus of the TIMSS assessment makes it more suitable to study the impact of teacher subject-specific qualifications, as these are more likely to affect students' knowledge in a specific grade.[8] TIMSS employs a two-stage random sample design. In the first stage, a random sample of schools is drawn from each participating country with sampling probabilities proportional to school size. In the second stage, one or more entire classes of students are randomly selected from each school.[9] By sampling entire

[7] Using teacher math and reading test scores, Metzler and Woessmann (2012) finds that an increase of one SD in teacher math test scores raises 6th-grade students' math test scores by 9% SD, but has no effect on reading test scores in Peru.

[8] Conversely, the better-known Programme for International Student Assessment (PISA) tests 15-year-old students' general problem-solving ability in math, science, and reading, regardless of students' curriculum and school grade.

[9] In the sample of my analysis, in 76% of the cases only one class per school was sampled in each school.

classes, TIMSS offers the ideal setting to study the relationship between teacher characteristics and student outcomes. The TIMSS achievement scale was established in 1995 by setting the mean of the average score of all participating countries in TIMSS 1995 to 500 and the standard deviation to 100.

I focus on 8th graders in science. I exclude 4th graders since teachers in primary school are typically trained as generalist teachers (Tatto et al. 2012), which raises important questions regarding the representativeness of the minority of teachers with subject-specific qualifications.[10] In 2015, 40 countries took part in the TIMSS 8th-grade study. 8th graders are around 14 years old and their TIMSS science assessment is made up of the following four subjects, with the share of questions concerning each domain reported in parentheses: biology (35%), chemistry (20%), physics (25%) and earth science (20%).[11] On top of the students' overall science test scores, TIMSS provides test scores for the above-mentioned four science subjects.[12] This is crucial for my identification strategy, which exploits within-student across-subjects variation. I consider the 30 countries where a single teacher teaches all four science subjects, i.e., countries where science is taught as an integrated subject, which allows me to exploit the within-teacher variation.

While compelling from an econometric standpoint, the within-student approach has recently received some criticism due to the design of international large-scale assessments (Jerrim et al. 2017). These tests typically use a matrix-sampling approach that involves splitting the entire pool of test questions into achievement booklets. Students are then randomly assigned to complete only one booklet. This approach ensures a comprehensive picture of the achievement of the student population while keeping the length of the test for each student manageable. Focusing on the Programme for International Student Assessment (PISA), Jerrim et al. (2017) highlights that if a student's booklet does not contain any questions regarding a

[10] In TIMSS 2015, 79% of 8th graders have science teachers with subject-specific qualifications in science, while only 38% of 4th graders have such teachers (Martin et al. 2016).

[11] TIMSS distinguishes between "subjects", i.e., math and science, and the "domains" that constitute each subject, such as biology, chemistry, physics, and earth science for science. To ease exposition, I refer to biology, chemistry, physics, and earth science as subjects. For 4th graders, the TIMSS assessment does not make a distinction for biology and chemistry, which are grouped together under the name "life science". This distinction does not map directly into teachers' majors and is a further reason to exclude 4th graders from the analysis.

[12] For each test score, TIMSS provides five plausible values. Throughout the analysis, I use the first plausible value for each subject. As a robustness check, I replicate the main result of the analysis using all five plausible values for all science subjects. Results are robust to this specification (see Table A3.10).

specific subject or domain, multiple imputation is used to create the missing test scores. The resulting within-student variation would then mostly capture the noise induced by the imputation technique. However, unlike PISA, each TIMSS 2015 booklet contains both math and science questions, and, importantly for my application, each science block replicates the proportion of science subjects that constitute science (see Mullis and Martin (2013) for further details about the TIMSS 2015 assessment design), thus limiting this concern.

The explanatory variable of interest comes from the teacher questionnaire, where teachers are asked to indicate their major(s) during their post-secondary education in a pre-specified set of subjects.[13] I construct the sample of interest so that each observation consists of a student-subject combination, which yields four observations for each student. Teacher subject-specific qualifications, the explanatory variable, is a dummy variable that takes value one if the science teacher reports holding a major in the corresponding subject and zero otherwise. For example, if a student's teacher reports holding a major in biology but not in other science subjects, the teacher-subject-specific-qualifications variable will take value one for that student-biology observation and zero for the other student's observations (student-physics, student-chemistry, student-earth science). This constitutes the source of variation in the explanatory variable that I exploit in the within-student within-teacher across-subjects approach.

Teacher subject-specific qualifications might affect student achievement through different channels. For example, teachers might be more prepared to teach subjects in which they have a major. Using the teacher questionnaire, I construct a variable to substantiate this hypothesis.[14] For each science subject, the teacher questionnaire includes a list of topics (5.5 on average) addressed by the TIMSS science test.[15] For

[13] The original wording of the question is: "During your <post-secondary> education, what was your major or main area(s) of study?". And the possible subjects are: Mathematics, Biology, Physics, Chemistry, Earth Science, Education-Mathematics, Education-Science, Education-General, Other. Teachers can indicate as many majors are they see fit.

[14] Potentially, other channels might also be relevant, such as subject knowledge, motivation, or teaching methods. Unfortunately, the data at hand do not allow any investigation of these or other channels.

[15] For the full list of topics and the exact wording of the question, see Table A3.1 in the Appendix. An example of a topic for biology is *"Cells, their structure and functions, including respiration and photosynthesis as cellular processes".* For chemistry, *"Physical and chemical properties of matters".* For physics, *"Energy forms, transformations, heat, and temperature".* For earth science, *"Earth's structure and physical feature […]".*

each of these topics, teachers can indicate whether they feel *very well prepared, somewhat prepared,* or *not well prepared* to teach it.[16] If a teacher holds a subject-specific qualification in a subject, she might feel more prepared and, therefore, confident to teach topics that belong to that subject, and this might raise student test scores. I define a variable that captures the level of preparedness of teachers as the share of topics in each subject that a teacher feels *very well prepared to teach* and test whether it is a mediator of the effect of teacher subject-specific qualifications in the mediation analysis.

In 2015, 40 countries and 285,119 8th graders participated in the TIMSS science assessment. While in most countries science at the 8th grade is taught as an integrated subject, with a single teacher teaching all science subjects, this is not the case in 10 of the countries participating in TIMSS 2015 (namely Armenia, Georgia, Hungary, Kazakhstan, Lebanon, Lithuania, Malta, Russia, Slovenia, and Sweden). In countries where science is taught as separate subjects, a teacher only teaches one of the four science subjects in a classroom. I exclude the 10 countries where science is taught as separate subjects, thus excluding 47,292 students (16.6% of the original sample), as they are not suitable for the within-student within-teacher approach. In the remaining countries, I also exclude 13,383 students (4.7% of the original sample) who are taught science by more than one teacher.[17] The resulting sample consists of 224,454 students, 11,243 teachers and 30 countries. As each student is observed in four science subjects and the unit of observation is the student-subject combination, the total number of observations is 897,760. Throughout the analysis, I use the student sampling weights.

I standardize all test scores within-subject so that the average test score has mean zero and standard deviation one in each subject. Regression coefficients can be therefore interpreted in terms of percentage of a standard deviation. Missing values in the explanatory variable of interest as well as in the controls are imputed using country-level mean imputation. For the main explanatory variable of interest, teacher subject-specific qualifications, I include an imputation dummy in all the regressions.

[16] Teachers can also select the option "*not applicable*" if the topic is not in the 8th grade curriculum or they are not responsible for teaching that topic. For the same list of topics, teachers are also asked whether they taught the topic this year, before this year or not (see Table A3.1, Panel B). However, the topics taught might reflect differences in curricula rather than being an outcome of teacher qualifications and are therefore not included in the mediation analysis.

[17] The only exception is Morocco, where students are taught physics and chemistry by one teacher and biology and earth science by another teacher. This framework also yields within-teacher variation as I observe each teacher in two subjects.

11.8% of values in the teacher subject-specific qualifications variable are missing. All regression results are robust to the exclusion of observations where teacher subject-specific qualifications are missing.

3.2.2 Descriptive Statistics

I report the main descriptive statistics for the sample of interest in Table 3.1. Concerning the main explanatory variable, biology is the most common teacher subject-specific qualification, with 42% of the students taught by a teacher with a major in biology, followed by chemistry (36% of the students), physics (31%), and earth science (20%). It is important to remind that teachers can report more than one subject-specific qualification; in fact, students are taught on average by teachers with 1.24 subject-specific qualifications in science. The modal student is taught by a science teacher with one subject-specific qualification.[18] This figure varies substantially across countries, with the highest average number of teacher subject-specific qualifications in Israel and the lowest in Ontario (Canada) (see Column 1 in Table A3.3 in the Appendix). On average, 73% of the students are taught by teachers who hold at least one subject-specific qualification. Again, this figure masks important cross-country heterogeneities, with the highest share of such students being in England and Morocco and the lowest in Iran (see Column 2 in Table A3.3 in the Appendix). Overall, these data suggest that most 8th-grade science teachers have acquired university-level content knowledge in at least one of the science subjects that they teach.[19] Even teachers without a major in a certain subject likely received some form of training in the content of the subject that they teach. In fact, according to the international teacher survey TALIS 2018 led by the OECD, 92% of a representative sample of lower secondary education teachers in 48 countries report having received training in the content of the subject that they teach (OECD 2019). The source of variation in the explanatory variable that I exploit for the preferred identification strategy stems from students being taught by teachers having at least

[18] For the distribution of the number of subject-specific qualifications, see Table A3.2 in the Appendix (Column 3). Along with subject-specific qualifications, teachers can also indicate whether they have majors in other subjects, including education, education-science, or education-math. I also report the distribution of the number of subject-specific qualifications by whether teachers also hold any major in education in Table A3.2 (Column 1 and 2).

[19] The observed cross-country heterogeneities might be due to how teachers are trained and selected in different countries. Another explanation is that the concept of majoring in one subject differs across countries. Thus, the subject knowledge acquired by majoring in one subject might also vary accordingly, affecting the independent variable's cross-country comparability. Nonetheless, this concern is not an issue for my estimates as I do not exploit variation stemming from cross-country variation in the independent variable.

one and less than four science subject-specific qualifications. It is therefore important that a considerable number of students are taught by teachers who satisfy this requirement. This is in fact the case, as 66% of the students are taught by such teachers (see Column 3 in Table A3.3 in the Appendix).

Apart from the subject-specific qualifications, the TIMSS background questionnaires provide a wealth of information on teachers' and students' backgrounds, which I now briefly describe. On average, science teachers of 8th-grade students report high levels of education. 62% of the students are taught by teachers with a Bachelor's degree and 22% by teachers with a Master's degree. These figures are in line but slightly smaller than those reported for the entire TIMSS 2015 8th-grade science sample, in which 92% of students are taught by teachers with at least a Bachelor's degree. Teachers report having, on average, 14.54 of experience, in line with the figure reported for the entire TIMSS 2015 sample of 15 years of experience. The share students taught by teachers who report having a major in education is 61%[20] and having a major in education is negatively correlated with also having a subject-specific qualification in science (-.28, *p*-value < .001). The share of female teachers is 58%. The average weekly instruction time for students in science is 5.65 hours. On average, students are taught by teachers who feel confident to teach 54% of the topics tested in TIMSS.

To explore country heterogeneities, I include country-level data from a variety of sources. For the distinction between developed and developing countries, I use the World Economic Situation and Prospects (WESP) 2014 classification of the United Nations (United Nations 2014). For GNI per-capita measures of countries in 2015, I use the World Bank data (World Bank 2021). The large variation in average science performance of the considered countries as well as other factors such as geographical location or economic development speaks in favor of the external validity of this study.

[20] This figure includes teachers that report having either a major in education, education-science or education-mathematics. The figure for teachers who report having a major in education-science is 51%, for teachers who report having a major in education is 27%, and for education-mathematics is 9%. According to the TALIS 2018 survey, 92% of teachers across OECD countries and all the countries participating in TALIS received training in general pedagogy and in the pedagogy of the subjects that they teach (OECD 2019). It is therefore unlikely that teachers in my sample do not have any pedagogical preparation, regardless of whether they report holding any major in education.

3.3 Empirical Strategy

To causally estimate the effect of teacher subject-specific qualifications on test scores, one would need to assume that teachers are randomly assigned to students and subject-specific qualifications to teachers. In practice, however, this is unlikely to be the case. First, the allocation of teachers is typically non-random, as, for example, wealthy parents tend to secure better resources for their children by choosing better schools (Clotfelter, Ladd, and Vigdor 2006). Second, teachers' decision to obtain subject-specific qualifications might depend on preferences or ability. If the teacher subject-specific qualifications are correlated with determinants of student test scores, the estimated effect of teacher subject-specific qualifications will be biased.

To address these concerns, I first implement a standard OLS approach estimating an education production function with a rich set of controls, which account for observable heterogeneities. I then implement a within-student within-teacher approach which also accounts for unobserved student and teacher heterogeneity that are subject invariant.

I first estimate the following linear model using OLS with a rich set of controls:

$$A_{istk} = \alpha S_{istk} + \gamma' X_{tk} + \delta' Z_{itk} + \theta_k + \varphi_s + \varepsilon_{istk} \quad (3.1)$$

where A_{istk} denotes the test score of student *i* in subject $s \in$ (*biology, chemistry, physics, earth science*), taught by teacher *t* in country *k*. A_{istk} is determined by the teacher subject-specific qualifications of student *i*'s teacher *t* in subject *s*, S_{istk}, a vector of teacher as well as class and school characteristics X_{tk}, a vector of students characteristics Z_{itk}, country fixed effects θ_k and subject fixed effects, φ_s, with ε_{istk} being the idiosyncratic error. This model accounts for several factors that are known to affect students' outcome, such as students' socioeconomic status or gender (included in the vector Z_{itk}), teachers' experience (included in the vector X_{tk}) as well as country and subject heterogeneities (captured by the fixed effects θ_k and φ_s, respectively).

The main identifying assumption to obtain an unbiased estimate of the parameter of interest, α, is that teacher subject-specific qualifications, S_{istk}, are uncorrelated with the error term conditional on the included regressors. While controlling for observable student, teacher and class characteristics alleviates some of the concerns mentioned previously, unobservable determinants of students' test scores that are correlated with teacher subject-specific qualifications might still lead to a violation of the

identifying assumption. If, for example, higher ability students are systematically sorted into classes with teachers with subject-specific qualifications, the estimated α in Eq. (3.1) is potentially upward biased. Conversely, α could be downward biased if teachers with subject-specific qualifications tend to be assigned to classrooms with lower-ability students. Similarly, more motivated, or higher-ability teachers might be more likely to hold a subject-specific qualification. Thus, both student and teacher unobserved characteristics can potentially bias the estimate of α and can do so independently from each other. It is therefore important to develop an identification strategy that can tackle both sources of bias.

To this purpose, I estimate a within-student within-teacher across-subjects model. As I observe the results of each student in four distinct science subjects, I can eliminate the heterogeneity due to unobservable student characteristics that do not vary across science subjects by including student fixed effects in Eq. (3.1). Further, I also observe every teacher in the same four subjects. I therefore include teacher fixed effects in Eq. (3.1), which control for all unobserved teacher characteristics that do not vary across subjects.[21] Essentially, student and teacher fixed effects account for all the observable and unobservable characteristics at the student, teacher, class, and school level that do not vary across subjects. Empirically, I estimate the following linear model:

$$A_{ist} = \beta S_{ist} + \sigma_i + \tau_t + \varphi_s + \varepsilon_{ist} \tag{3.2}$$

Where the subject-specific test score A_{ist} is determined by the teacher subject-specific qualifications S_{ist} and the student, teacher, and subject fixed effects (σ_i, τ_t, and φ_s respectively). Student and teacher fixed effects make the inclusion of all the subject-invariant student (Z_{itk}), teacher and classroom (X_{tk}) variables as well as country fixed effects (θ_k) redundant and are therefore omitted from Eq. (3.2).

Student fixed effects control for many subject-invariant characteristics that are known to affect student achievement, such as the socioeconomic status, general motivation, innate abilities, as well as classroom and school characteristics. Similarly, teacher fixed effects control for the subject-invariant components of observables teacher characteristics, such as teacher experience, education level or gender, as well as the subject-invariant components of unobserved teacher characteristics, such as motivation or ability. Finally, subject fixed effects eliminate subject-specific test score

[21] This represents the main difference with respect to the identification strategy in Sancassani (2021), where teachers are observed in only one science subject, thus preventing the exploitation of the within-teacher variation.

heterogeneities and other subject-specific unobserved factors, such as different curriculum coverage in different subjects. The estimation of β in Eq. (3.2) is therefore unlikely to be biased by the two main sources of bias mentioned: the unobserved subject-invariant student and teacher characteristics.

The main threat to the identification strategy consists of unobserved subject-specific heterogeneities. In fact, the estimated β might still be biased if unobserved subject-specific determinants of student outcomes, such as subject-specific instruction time, student or teacher ability or passion for the subject are correlated with the teacher subject-specific qualifications. To alleviate such concerns, I show that the results are robust to the inclusion of subject-specific instruction time and to restricting the sample to schools where student sorting is unlikely. Furthermore, following Oster's bounding exercise (Oster 2019), I show that any remaining bias due to unobserved factors should be negligible. Another concern for my identification strategy is that the estimated β might capture the effect of being taught by a teacher with subject-specific qualifications in the 8th grade and in previous years. Unfortunately, the data at hand do not allow to control for the qualifications of teachers in previous years. Nonetheless, the focus on the grade-specific knowledge of the curriculum of the TIMSS assessment ensures that any bias through this channel is most likely small. Finally, it is worth reminding that the more likely sorting of student and teachers based on student SES, general ability or interest for science is accounted for by student and teacher fixed effects.

A further assumption of this model is that the impact of teacher subject-specific qualifications is homogenous across subjects. Compared to studies that use a similar within-student identification strategy but using different subjects, it is a far weaker assumption in this setting, as the student test scores belong to the same field. Other things being equal, it is unlikely that science subject-specific qualifications might have a larger or smaller impact in different science fields, and I provide evidence of this. [22] Further, I show that the effect of teacher subject-specific qualifications is robust and stable with respect to the individual exclusion of each science subject in the robustness checks, which alleviates this concern.

[22] I directly test this by estimating the linear model in Eq. (3.1) including an interaction term between teacher subject-specific qualifications and subjects. I then perform a Wald test of equality of all the coefficients of the interaction terms, which I cannot reject (p-value = .77, F-statistic = .26); pairwise tests of equality of the coefficients also rule out heterogeneity in the coefficients.

A potential downside of using closely related outcomes is that the effect of teacher subject-specific qualifications in one science subject might spill over into other subjects. Relatedly, being the subjects so closely related to each other, the amount of variation that can be exploited should not be too large, as they probably require a similar set of student innate abilities. Considering these points, my estimates likely reflect a lower bound of the true effect.

3.4 Results

3.4.1 Main Results

Table 3.2 presents the main results of the impact of teacher subject-specific qualifications on student test scores. I first report results of the linear model described in Eq. (3.1) pooling the student test scores in the four science subjects—biology, chemistry, physics, and earth science—with an increasingly rich set of control variables (Columns 1-3). I then report the result using the within-student within-teacher across-subjects approach described in Eq. (3.2) (Column 4). The impact of teacher subject-specific qualifications on student test scores is positive and statistically significant and varies between 3.3% SD to 3.6% SD. The preferred estimate, the one obtained with the within-student within-teacher across-subjects approach (Column 4), lies between the coefficients of the pooled linear models. It is positive and statistically significant at the 1% level and implies that teacher subject-specific qualifications raise student test scores in the subject in which a teacher holds a subject-specific qualification by 3.5% SD. The estimated coefficient in Column 1 changes very little when including controls and fixed effects in the regressions, despite a substantial increase in the R^2. This suggests that the remaining bias due to unobserved subject-specific factors is likely small. I substantiate this claim formally in Section 3.4.4, where I perform an analysis of unobservable selection and coefficient stability following Oster (2019).

Results show that teacher subject-specific qualifications matter for student science test scores. The magnitude of the effect, equivalent to 3.5% SD, is relatively small for a single school year but can become substantial if considered over a school cycle of six years, the average duration of secondary education worldwide (UNESCO 2021).

3.4.2 Heterogeneity – Student and Teacher Characteristics

I explore heterogeneities of the impact of teacher subject-specific qualifications in Table 3.3 using the within-student within-teacher across-subjects approach in Eq. (3.2). Several studies have found that student and teacher gender matters for

educational achievement, especially for female students (Dee 2005; Paredes 2014; Lim and Meer 2017; Sansone 2017). This is even more important in science and, more in general, STEM subjects, where females have been historically underrepresented. I interact the teacher subject-specific qualifications separately with student and teacher gender (Column 1 and 2, respectively) to tease out heterogeneities in the effect of teacher subject-specific qualifications with respect to student and teacher gender. Estimates suggest that female students benefit more from being taught by a teacher with subject-specific qualifications (Column 1), whereas teacher gender alone does not seem to play a role for the effectiveness of teacher subject-specific qualifications (Column 2). As a further step, I explore whether female students, who already benefit more from being taught by teachers with subject-specific qualifications, benefit even more when these teachers are also females. The rationale for this analysis follows the role-model effect of teachers observed in the literature (Dee 2005; Paredes 2014), according to which girls benefit from being assigned to female teachers without negative effects for boys. Such effect is possibly because female students might be more confident in learning science if the role-model to which they are exposed is a female teacher. I therefore test whether the interaction term between the teacher subject-specific qualifications and student gender varies by teacher gender.[23] I find that female students taught by teachers with subject-specific qualifications perform significantly better when their teachers are also females (table not shown), in line with the teacher role-model effect mentioned previously.

Teacher subject-specific qualifications may have a different impact on students with different SES, which, to a large extent, also captures student prior achievement. Theoretically, the marginal increase in teacher subject knowledge induced by teachers acquiring subject-specific qualifications might have different returns based on students' prior knowledge. Differences in the impact of teacher's subject-specific qualifications with respect to student SES might therefore reveal different functional forms that characterize the relationship between teacher subject knowledge and students' achievement. I explore such heterogeneity in Column 3, where I interact

[23] Empirically, I include an interaction between teacher gender and the interaction between teacher subject-specific qualifications and student gender to the model estimated in Column 1, but without including the main effects for the triple interaction. This is equivalent to estimating the interaction term between teacher subject-specific qualifications and student gender separately for the sample of female and male teachers. The coefficient associated with the triple interaction, which captures the effect for female students taught by female teachers with subject-specific qualifications, is positive and statistically significant (.018, *p*-value < .10). Similarly, the effect of teacher subject-specific qualifications for female students is larger when estimated for the sample of female teachers as opposed to the sample of male teachers (.072 and .053, respectively; with *p*-value < .01 for both terms).

teacher subject-specific qualifications with an indicator for student SES. I find that the effect of teacher subject-specific qualifications decreases as student SES increases. This finding suggests a steeper relationship between teacher subject knowledge and student achievement for lower SES students.[24] It also has important equity implications, as students from more disadvantaged contexts benefit the most from having teachers with subject-specific qualifications.

A similar theoretical argument can be made for other teacher qualifications. Teacher subject-specific qualifications could affect students' test scores differently based on teachers' general educational attainment or pedagogical knowledge. A steeper relationship between teacher subject knowledge and student test scores for teachers who also have a Master's degree or a major in education might indicate a complementarity between these additional qualifications. I explore these hypotheses in Column 4 and 5. I do not find a statistically significant interaction between teacher subject-specific qualifications and teacher holding a Master's degree (Column 4). I therefore do not find supporting evidence for the complementarity between such qualifications. Conversely, the interaction between teacher subject-specific qualifications and whether the teacher holds a major in education is positive and statistically significant (Column 5), which implies that the effect of teacher subject-specific qualifications is larger for teachers who also have a major in education. This result suggests that teacher pedagogical knowledge, captured by the major in education, and teacher subject knowledge, captured by the teacher subject-specific qualifications, are complementary ingredients for effective teaching.

Finally, I explore the role that teacher experience plays in the effectiveness of teacher subject-specific qualifications (Column 6). I include both a linear and quadratic term for teacher experience[25] to tease out the largely documented non-linear relationship between teacher experience and student test scores (Rivkin, Hanushek, and Kain 2005; Boyd et al. 2008; Clotfelter, Ladd, and Vigdor 2010). The coefficients suggest a concave relationship between the effect of teacher subject-specific qualifications

[24] I find similar results by interacting teacher subject-specific qualifications with a more direct measure of student prior achievement, student math test scores (not shown). Students in the lower part of the distribution of math test scores benefit the most from teachers with subject-specific qualifications. The student SES indicator correlates highly with the math test scores, but due to the potential endogeneity of the math test scores, I stick to the interaction with student SES as the main specification for this analysis.

[25] Following the existing literature on the (non-linear) effect of teacher experience on student test scores, I also define teacher experience in bins (namely 0-1 year, 2-5 years, 6-9 years, 10-12 years, 13-16, 17-23, 24+). Results from this specification (not shown) are qualitatively the same.

interacted with teacher experience and students' achievement. I provide a graphical representation of this result in Figure 3.1, which shows that the effect of subject-specific qualifications reaches its peak around the midpoint of teacher experience (at roughly 18 years of experience), after which it declines. It is important to remind that teacher experience is collinear to teacher age. It is possible that the observed pattern is due to an experience effect, meaning that teachers improve their effectiveness in the first part of their career by, for example, learning by doing. Alternatively, this pattern could also be due to a cohort effect, meaning that the ability of teachers differs by cohort.[26] Given the cross-sectional nature of the data, I cannot disentangle these two components, but the pattern observed in this analysis is more in line with the vast literature reporting diminishing returns to teacher experience.

3.4.3 Heterogeneity – Country Subsamples

The wide heterogeneity of the countries considered is advantageous for the external validity of the results, although it brings additional challenges. If teacher training differs markedly across countries, holding subject-specific qualifications might mean different things. I therefore focus on the sub-group of OECD countries in the sample, for two main reasons. First, teachers in OECD countries report, on average, fewer subject-specific qualifications despite a higher level of education.[27] This likely indicates that subject-specific qualifications represent teachers' main field of study in OECD countries. Second, OECD surveys provides a wealth of information regarding teacher training. This allows me to provide a clearer picture about the framework in which teachers are selected and trained in these countries. According to the TALIS 2018 survey, in the OECD countries included in my sample except for Canada and Ireland, which are not covered in TALIS 2018, 92.7% of teachers report to have received training in the content of some or all subjects taught, 90% have received training in pedagogy of some or all subjects taught, and 92% in general pedagogy. These figures suggest that teachers in OECD countries likely received some training in both pedagogy and the content of the subjects they teach, regardless of their subject-specific qualifications. Further, the educational requirements for entry into initial teacher training differ little across OECD countries, where the minimum requirement is usually an upper secondary qualification (OECD 2022).

[26] For example, Nagler, Piopiunik, and West (2020) show that teachers who enter the profession during economic downturns are significantly more effective in raising student test scores.

[27] 34% of students in non-OECD countries are taught by teachers who report two or more subject-specific qualifications, while only 26% of students in OECD countries are taught by such teachers.

I report the main results for this subgroup of countries in Table 3.4 with the same specifications used for Table 3.2. All estimated coefficients are positive and statistically significant, although they decrease as I include more controls in the model. Interestingly, the R^2 in Column 1 is much smaller than the R^2 in the same specification in Table 3.2, which indicates that this group of countries is much more homogenous. In the preferred specification of Column 4, the magnitude of the coefficient is 2.8% SD, which is slightly smaller than the coefficient estimated in Table 3.2 for the full sample, although not statistically significantly different from it, as I show in Table 3.5. This implies that, even in the context of OECD countries where teachers likely received extensive training, students perform better in those subjects where their teachers hold subject-specific qualifications. To test whether the impact of teacher-subject specific qualifications varies by country subsamples, I include interactions between teacher subject-specific qualifications and a series of country indicators[28] in Eq. (3.2) and report the results in Table 3.5. First, I explore whether the effect of teacher subject-specific qualifications varies in countries that belong to the OECD (Column 1) or are developed countries[29] (Column 2). A priori, it is unclear if teacher subject-specific qualifications could be more effective in OECD (developed) or non-OECD (developing) countries. This ultimately depends on a variety of factors, such as the already mentioned teacher preparation, the attractiveness of the teaching career and so on. While the interaction term in Column 1 points to the negative area, it does not reach any conventional level of statistical significance. However, the interaction term in Column 2 suggests that teachers with subject-specific qualifications are more effective in developing countries.

The effect of teacher subject-specific qualifications might also depend on countries' average science achievement. A priori it is unclear whether students in countries with high average achievement could benefit more from having teachers with subject-specific qualifications. I therefore split the sample in countries that perform above and below the median science test score in my sample. Results show that teachers with subject-specific qualifications are more effective in countries with average science performance below the median (Column 3). A further distinction between countries that are above and below the median GNI per capita does not show significant

[28] For the list of all countries and the country indicators, see Table A3.3 in the Appendix.

[29] For the developed vs. developing countries classification, I used the WESP classification (United Nations 2014). This classification includes a further category of countries "in transition". However, none of these countries is in the sample I analyze. Being the OECD a club of mostly rich countries, the developed countries group is a subset of the OECD group.

heterogeneities between relatively rich and poor countries (Column 4). A possible explanation for results from this table is that the counterfactual teacher effectiveness, i.e., the effectiveness of teachers in science subjects in which they do *not* have a major, is lower in developing or lower-performing countries. As previously argued, teachers in OECD countries seemingly received pedagogical and content training in the subjects that they teach. While the data at hand do not allow to make similar claims for developing and lower-performing countries, it is possible that teachers in these countries received, on average, less training. For this reason, subject-specific qualifications might have larger value-added for teachers in these countries.

3.4.4 Robustness Checks

As discussed in Section 3.3, the main threat to the identification strategy comes from unobserved subject-specific confounders, while subject-invariant confounders are accounted for by student and teacher fixed effects. I therefore perform a series of robustness checks to ensure that any remaining bias due to subject-specific heterogeneities should not invalidate my estimates. A possible concern comes from different instruction time devoted to science subjects. If schools or countries that emphasize one science subject over the others are also more likely to appoint teachers with subject-specific qualifications in that subject and devote more instruction time to the same subject, estimates might be upward biased.[30] To mitigate this concern, I replicate my main analysis using TIMSS 2011, which allows me to control for the share of instruction time that teachers report to dedicate to each science subject. Results are reported in Table A3.4. First, it is reassuring to see that I can essentially replicate the main result of the paper also using TIMSS 2011. The within-student within-teacher across-subjects specification in Column 4 is positive and statistically significant, albeit slightly smaller in magnitude than the main specification in Column 4, Table 3.2. Second, the results are robust when I control for instruction time in Column 3 and 5, although the coefficient in the preferred within-student within-teacher across-subjects specification in Column 5 slightly decreases. Following Bietenbeck, Piopiunik, and Wiederhold (2018), I address the issue of the remaining subject-specific student and teacher sorting by restricting the sample of my main analysis with TIMSS 2015 to students living in areas with less than 30 thousands, 15 thousands people or in rural areas. Students in these areas likely have little choice between different schools, which makes the issue of sorting less worrying. I report the

[30] However, instruction time can also be an outcome of teacher subject-specific qualifications if teachers systematically devote more instruction time to the subjects in which they have a major. In this case, controlling for instruction time would be problematic.

results from this analysis in Table A3.5. Results are robust to these specifications and, if anything, they are larger in magnitude. Finally, I conduct an analysis of unobservable selection and coefficient stability following Oster (2019). I compare the coefficient estimated through the within-student within-teacher across-subjects specification (Column 4 of Table 3.2) to the specification including only country and subject fixed effects (Column 1 of Table 3.2) and setting $R_{max} = 1$ and $\delta = 1$.[31] Results, reported in Table A3.6, indicate that the estimated bias-adjusted treatment effect β^* is .035, which is identical to the preferred estimate. The value of δ for which $\beta = 0$ is 19.51, which far exceeds the standard cutoff of 1 and implies that the selection on unobservable characteristics needs to be almost 20 times larger than the selection on observables characteristics to drive the effect of teacher subject-specific qualifications to zero.

To ensure that results are not driven by a specific subject where teachers might benefit particularly from holding a subject-specific qualification, I replicate the main result by excluding one subject at a time. Results in Table A3.7 show that the effects are robust to the exclusion of each science subject. These results also address a concern raised in Section 3.3 about the potential bias induced by heterogenous effects of teacher subject-specific qualifications and confirm that the results are rather homogeneous across different science subjects.

Given the heterogeneity of the countries considered in my analysis, it is possible that the effect of teacher subject-specific qualifications is driven by some countries where teachers with such qualifications are particularly effective in raising student test scores. I address this concern by replicating the main result excluding one country at a time. Results from the leave-one-country-out exercise in Table A3.8 are robust to the exclusion of each country in the sample. It seems therefore unlikely that results are driven by some outliers in the sample of countries considered. The effect of teacher subject-specific qualifications varies between 2.3% and 3.9% of a SD, with the lower and upper bound obtained when Egypt and Japan are excluded, respectively. Japan and Egypt lie at the opposite extremes of the distribution of science performance, with Japan being among the highest performing countries and Egypt among the lowest performing countries in the sample. This finding corroborates the evidence that the

[31] These values denote the R^2 from a hypothetical regression of the outcome on the treatment and both observed and unobserved controls, and the relative degree of selection on observed and unobserved variables (Oster 2019), respectively. In practice, Oster (2019) recommends an $R_{max} = 1.3\tilde{R}$, where $\tilde{R}$ denotes the R^2 obtained in the regression with all controls, which in my case is .94 (see column 4 of Table 3.2). I therefore set $R_{max} = 1$ since setting $R_{max} = 1.3\tilde{R}$ would imply an implausible $R_{max} > 1$.

effect of teacher subject-specific qualifications is stronger in lower-performing countries.

A further issue concerns the weight that each country has in the analysis. Due to different sample sizes across countries, different countries carry different weights in the analysis. Instead of using the sampling weights provided by TIMSS, I replicate the results using rescaled weights so that each country carries the same weight ("senate weights"). Results, shown in Column 2 in Table A3.9, are robust to this specification, although slightly smaller in magnitude.[32]

I also address issues related to the complex design of international assessment in Table A3.10. First, to minimize manipulation of the test scores, I replicate the main results using the raw (i.e., non-standardized) first plausible value for each science subject as outcome (Column 2). I find that the impact of being taught by a specialized teacher is equivalent to 4.37 points, which corresponds to 3.7% SD,[33] in line with the coefficient estimated in the main specification (3.5% SD). Second, to account for the uncertainty about the process through which student test scores are computed, I use all five plausible values for each science subject.[34] The results (Column 3) show that the effect of teacher subject-specific qualifications is robust to using all five plausible values and virtually identical to those obtained using only the first plausible value, and the standard error is roughly 10% larger. Finally, I address the issue of sampling variance typical of large-scale assessment such as TIMSS. To estimate standard errors that consider its multistage cluster sampling design, TIMSS suggests using the Jackknife Repeated Replication (JRR) technique.[35] Again, results in Column 4 are robust to this specification, with the JRR technique inflating the standard errors by a

[32] Some studies using international assessments (Lavy 2015; Rivkin and Schiman 2015; Cattaneo, Oggenfuss, and Wolter 2017; Bietenbeck, Piopiunik, and Wiederhold 2018) do not apply weights. I also check that my results are robust to this specification (in Table A3.9, Column 3) and similar to those obtained using "senate weights".

[33] This coefficient is obtained dividing the coefficient in Column 2 (4.37) by the SD of the first plausible values of all science subjects (118.56).

[34] I touched upon this point in Section 3.2. It has been generally acknowledged that the use of single plausible values does not make a substantial difference in large samples (Jerrim et al. 2017). However, my study slightly deviates from the cases discussed in the literature as the test scores for each science subject that I use are based on a limited number of questions (between 12 and 18), thus making the issue potentially relevant.

[35] Interested readers may find more detail about this technique and its application to the TIMSS data in Mullis and Martin (2013). In a nutshell, the JRR technique consists of subdividing the sample into clusters of sampling units (e.g., schools) and repeatedly replicating the statistics of interest by modifying the weight given to the sampling units within the cluster.

further 10% with respect to Column 3. I also replicate the main results clustering standard errors at different levels, namely at the school, student, or teacher level. Results (not shown) are robust to these specifications.

Last, I check the robustness of the results by dropping all observations for which teacher subject-specific qualifications is missing (11.8% of the sample). Results are also robust to this specification and virtually identical to those obtained in the main specification (teacher subject-specific qualifications coefficient = .034, *p*-value < .01).

3.5 Mediation Analysis

Having shown that teacher subject-specific qualifications increase student science test scores, I now explore a possible mediator through which this effect materializes. I focus on the share of topics that teachers feel confident to teach described in Section 3.2. Thanks to the increased subject knowledge that teachers acquire through a subject-specific qualification, teachers might feel more confident to teach topics in subjects in which they hold such qualification. A more confident teacher could be more effective in teaching a certain subject. Thus, the increased confidence in teaching certain topics is a possible channel through which teacher subject-specific qualifications affect student test scores. To substantiate this hypothesis, I perform a mediation analysis in the spirit of Heckman, Pinto, and Savelyev (2013) and Heckman and Pinto (2015), following recent empirical implementations (Kosse et al. 2020; Resnjanskij et al. 2021; Hermes et al. 2021).

Variables must satisfy two conditions to act as mediators: they must be significantly affected by the independent variable of interest (specifically, teacher subject-specific qualifications) and be related to the outcome (student test scores). To test the first condition, I estimate the model described in Eq. (3.2) with the mediator as the dependent variable instead of student test scores. Results in Table A3.11 (Panel B) suggest that teachers with subject-specific qualifications are significantly more confident to teach topics that belong to the subject in which they hold a major. The result confirms that the mediator is significantly affected by teacher subject-specific qualifications. Looking at the magnitude of the coefficient, teacher subject-specific qualifications seem to have a large impact on the share of topics that teacher feel confident to teach, equivalent to 14.2 percentage points (or 39% SD).

To test the second condition, I include the mediator on the right-hand side of the baseline model of Eq. (3.2). Results are reported in Table A3.11 (Panel A). First, I report the impact of teacher subject-specific qualifications excluding the mediator (Column

1) and then with the mediator (Column 2). The mediator is significantly related to the outcome. As expected, the magnitude of the impact of subject-specific qualifications on student test scores decreases when the mediator is included, as the mediator captures part of the impact.

Finally, I compute the share of the effect of teacher subject-specific qualifications that can be attributed to the mediator.[36] As graphically shown in Figure 3.2, 20% of the effect of teacher subject-specific qualifications on student test scores is explained by teachers being more confident to teach topics that belong to subject in which they hold a major, while the remaining part is due to unobserved factors. Such factors might be, for example, increased subject or pedagogical knowledge acquired through teacher subject-specific qualifications.

3.6 Conclusion

In this paper, I explore the effect of teacher subject-specific qualifications on student science test scores. I find that teachers with subject-specific qualifications raise student science test scores in the subjects in which teachers hold a major by 3.5% SD. The effect is robust to a variety of specifications and across different groups. The effect is larger for female students, especially when they are taught by female teachers, and for students from more disadvantaged backgrounds. Further, I find that the effect of teacher subject-specific qualifications is stronger in lower-performing countries. The mediation analysis reveals that 20% of the effect can be explained by the fact that teachers with subject-specific qualifications feel more confident to teach topics that belong to the subject in which they hold a major.

These findings are important for three reasons. First, I provide evidence of the importance of teacher subject-specific qualifications for student test scores in a broad set of countries. This finding adds to the existing literature on teacher subject-specific qualifications, which has focused almost exclusively on the US. Second, I shed light on an understudied yet important subject, science, for which existing evidence is mixed. Third, I exploit the richness and international nature of the data to provide further

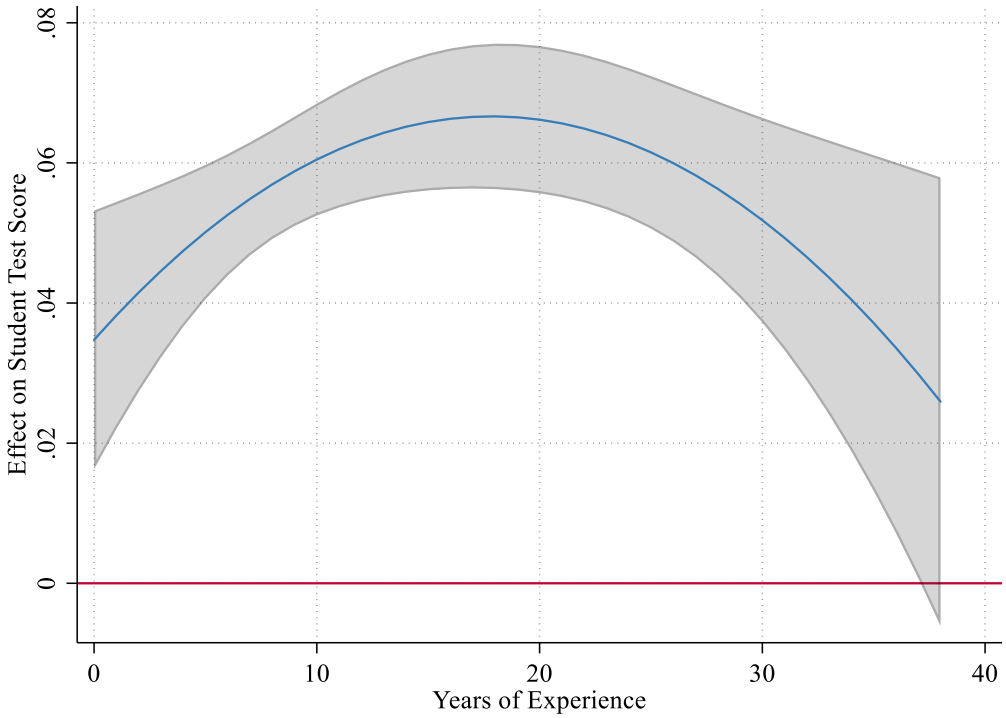
[36] The share is obtained by multiplying the coefficient of the coefficient of the impact of the independent variable on the mediator (.142, reported in Table A3.11, Panel B) by the association between the mediator and the outcome of interest (.05, report in Table A3.11, Column 2, Panel A) and dividing by the impact of the independent variable on the outcome (.035, reported in Table A3.11, Column 1, Panel A).

insights into the contexts and countries where subject-specific qualifications may have the greatest impact.

In terms of policy implications, countries should promote the acquisition of subject-specific qualifications, especially for science teachers in secondary schools. For example, countries could raise the standards required to become science teachers. This appears to be even more important for female students, for disadvantaged students and for lower-performing countries. Such policies could therefore increase both equity and efficiency in education systems worldwide. It is unclear whether students would also benefit from a further division of labor where teachers would only teach subjects in which they hold a major. Previous findings on such division of labor in elementary schools for math and reading are not encouraging (Fryer 2018), although findings for science are more promising (Bastian and Fortner 2020), thus calling for more research on this topic.

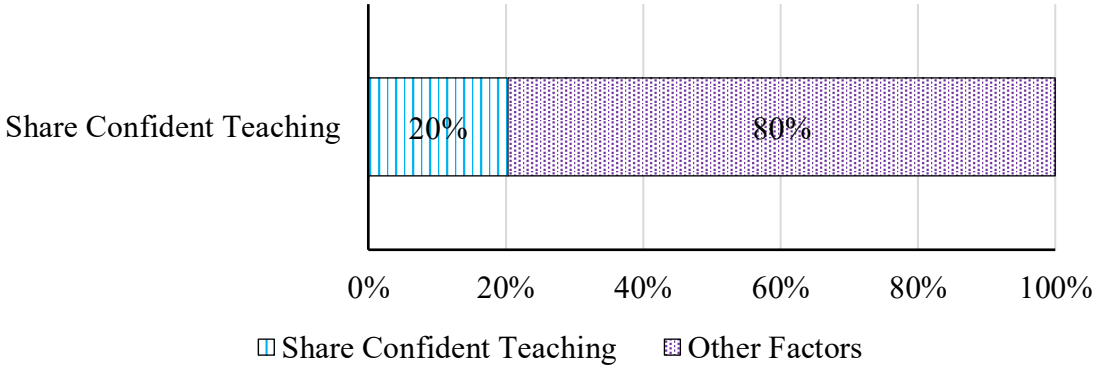
Figures and Tables

Figure 3.1: Effect of Teacher Subject-Specific Qualifications - Interaction with Teacher Experience



Note: The figure depicts the marginal effect of teacher subject-specific qualifications on student test scores along the domain of teacher experience with 95% confidence intervals. Estimates have been obtained by interacting teacher subject-specific qualifications with teacher experience in Eq. (3.2) and are reported in Table 3.3 in Column 6.

Figure 3.2: Share of the Effect of Teacher Subject-Specific Qualifications Attributed to the Mediator



Note: The figure depicts the share of the effect of teacher subject-specific qualifications on student test scores that can be attributed to the mediator. The estimates to compute such share can be found in Table A3.11.

Table 3.1: Descriptive Statistics

	Mean (1)	SD (2)	Min-Max (3)
Teacher Subject-Specific Qualifications			
Biology	0.42	(0.47)	0.0-1.0
Chemistry	0.36	(0.46)	0.0-1.0
Physics	0.31	(0.44)	0.0-1.0
Earth Science	0.20	(0.37)	0.0-1.0
Teacher Variables			
N. of Subject-Specific Qualifications	1.24	(1.13)	0.0-4.0
At Least One Subject-Specific Qualification	0.73	(0.44)	0.0-1.0
Bachelors' Teachers	0.62	(0.49)	0.0-1.0
Masters' Teachers	0.22	(0.40)	0.0-1.0
Experience (y)	14.54	(9.26)	0.0-38.0
Any Major in Education	0.61	(0.47)	0.0-1.0
Female Teacher	0.58	(0.48)	0.0-1.0
Teaching time per week (hours)	5.65	(1.00)	3.0-10.0
Share Topics Confident to Teach	0.54	(0.37)	0.0-1.0
Student Variables			
Female Student	0.50	(0.50)	0.0-1.0
Student SES Indicator	10.04	(1.93)	4.2-13.9
Speak Language of Test at Home	0.79	(0.41)	0.0-1.0
Born in Country	0.95	(0.21)	0.0-1.0
# Observations		897,760	
# Students		224,454	
# Teachers		11,243	
# Countries		30	

Note: The table reports weighted descriptive statistics for the main variables of interest. The unit of observation is the student-subject combination. In the Teacher Subject-Specific Qualifications panel, I report the average number of students taught by teachers with a subject-specific qualification, separately for each science subject. In the Teacher Variables panel, I report the average number of subject-specific qualifications that teachers have and the share of students taught by teachers who hold at least one subject-specific qualifications (i.e., at least one major in either biology, chemistry, physics, or earth science). I also report the share of students taught by teachers who hold a Bachelors' degree, a Masters' degree, the years of experience of teachers, the share of teachers who hold any major in education (i.e., either in education, education-science or education math). The teaching time per week is the overall weekly instruction time in science reported by the teachers. The share of topics that teachers feel confident to teach is calculated within each subject as the share of topics that teachers feel very confident to teach. In the Student Variables panel, I report the student gender, the student SES indicator provided by TIMSS, which is a comprehensive measure of the socioeconomic status of the students, and it is based on questions regarding parents' education, number of books at home and number of home study supports available for students (such as an own room or internet connection). Speak language of test at home is a dummy variable that takes value "one" if a student speaks the language of the test always or almost always at home and "zero" otherwise. Born in country is a dummy variable that takes value "one" if a student is born in the country where the test is administered. I also report the total number of observations, the number of distinct students, teachers, and countries. As each student is observed four times (one for each subject), the total number of observations is equal to the number of distinct students multiplied by four.

Table 3.2: Effect of Teacher Subject-Specific Qualifications on Student Test Scores

	(1)	(2)	(3)	(4)
Teacher Subject-Specific Qualifications	0.033**	0.036***	0.035***	0.035***
	(0.016)	*(0.011)*	*(0.011)*	*(0.004)*
Subject FE	YES	YES	YES	YES
Country FE	YES	YES	YES	NO
Student, School Controls	NO	YES	YES	NO
Teacher Controls	NO	NO	YES	NO
Student, Teacher FE	NO	NO	NO	YES
Observations	897,760	897,760	897,760	897,760
R^2	0.33	0.48	0.48	0.94

Note: The table reports OLS estimation using a set of controls (Column 1,2,3) and student and teacher fixed effects (Column 4). The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights, subject fixed effects, and an imputation dummy for teacher subject-specific qualifications. Student controls include: student SES, gender, language spoken at home, mother's immigrant status, father's immigrant status, student's immigrant status, student's education expectations. School and class controls include class size, share of students with language difficulties, share of economically disadvantaged students, indicator for shortage of resources for science instruction, school discipline problems, school location, school emphasis on academic success. Teacher controls include teacher experience, gender, level of education, major in education. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table 3.3: Heterogenous Effect of Teacher Subject-Specific Qualifications on Student Test Scores – Student and Teacher Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Teacher Subject-Specific Qualifications	0.005	0.031***	0.078***	0.034***	0.025***	0.013
	(0.005)	*(0.006)*	*(0.017)*	*(0.005)*	*(0.005)*	*(0.010)*
× F. Student	0.059***					
	(0.006)					
× F. Teacher		0.006				
		(0.008)				
× Student SES			-0.004**			
			(0.002)			
× Teacher holds Masters' Degree				0.004		
				(0.008)		
× Teacher holds Major in Ed.					0.020***	
					(0.008)	
× Teacher Experience						0.004***
						(0.001)
× Teacher Experience^2 (× 100)						-0.010***
						(0.004)
Subject, Student, Teacher FE	YES	YES	YES	YES	YES	YES
Observations	897,760	897,760	897,760	897,760	897,760	897,760

Note: The table reports OLS estimation using subject, student and teacher fixed effects. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights and an imputation dummy for teacher subject-specific qualifications. I include an interaction between teacher subject-specific qualifications and student gender in Column 1, and teacher gender in Column 2. In Column 3 I include an interaction with the student SES indicator. In Column 4 and 5 I include an interaction for whether the teacher holds a Masters' degree or major in education, respectively. In Column 6, I include an interaction with teacher years of experience and years of experience squared multiplied by 100. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table 3.4: Effect of Teacher Subject-Specific Qualifications on Student Test Scores – OECD Countries

	(1)	(2)	(3)	(4)
Teacher Subject-Specific Qualifications	0.052***	0.044***	0.043***	0.028***
	(0.020)	*(0.012)*	*(0.012)*	*(0.004)*
Subject FE	YES	YES	YES	YES
Country FE	YES	YES	YES	NO
Student, School Controls	NO	YES	YES	NO
Teacher Controls	NO	NO	YES	NO
Student, Teacher FE	NO	NO	NO	YES
Observations	349,244	349,244	349,244	349,244
R^2	0.09	0.31	0.31	0.92

Note: The table reports OLS estimation using a set of controls (Column 1,2,3) and student and teacher fixed effects (Column 4) for OECD countries only (for the list of OECD countries, see Table A3.3). The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights, subject fixed effects, and an imputation dummy for teacher subject-specific qualifications. Student controls include: student SES, gender, language spoken at home, mother's immigrant status, father's immigrant status, student's immigrant status, student's education expectations. School and class controls include class size, share of students with language difficulties, share of economically disadvantaged students, indicator for shortage of resources for science instruction, school discipline problems, school location, school emphasis on academic success. Teacher controls include teacher experience, gender, level of education, major in education. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table 3.5: Heterogenous Effect of Teacher Subject-Specific Qualifications on Student Test Scores – Country Characteristics

	(1)	(2)	(3)	(4)
Teacher Subject-Specific Qualifications	0.039***	0.041***	0.048***	0.033***
	(0.008)	*(0.006)*	*(0.008)*	*(0.005)*
× OECD Country	-0.007			
	(0.009)			
× Developed Country		-0.015*		
		(0.008)		
× High-Performing Country			-0.023***	
			(0.009)	
× High-GNI p.p. Country				0.007
				(0.008)
Subject, Student, Teacher FE	YES	YES	YES	YES
Observations	897,760	897,760	897,760	897,760

Note: The table reports OLS estimation using subject, student and teacher fixed effects. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights and an imputation dummy for teacher subject-specific qualifications. I include an interaction between teacher subject-specific qualifications and an indicator for whether a country belongs to the OECD (Column 1), whether a country is a developed country according to the WESP classification (Column 2), whether a country average science score is above the median of the science test scores in the sample (Column 3) and whether a country GNI per capita in 2015 is above the median GNI per capita of the countries in the sample (Column 4). Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Appendix

Table A3.1: List of Science Topics Covered in TIMSS 2015

Panel A: Topics
Biology
a) Differences among major taxonomic groups of organisms (plants, animals, fungi, mammals, birds, reptiles, fish, amphibians) b) Major organs and organ systems in humans and other organisms (structure/function, life processes that maintain stable bodily conditions) c) Cells, their structure and functions, including respiration and photosynthesis as cellular processes d) Life cycles, sexual reproduction, and heredity (passing on of traits, inherited versus acquired/learned characteristics) e) Role of variation and adaptation in survival/extinction of species in a changing environment (including fossil evidence for changes in life on Earth over time) f) Interdependence of populations of organisms in an ecosystem (e.g., energy flow, food webs, competition, predation) and factors affecting population size in an ecosystem g) Human health (causes of infectious diseases, methods of infection, prevention, immunity) and the importance of diet and exercise in maintaining health
Chemistry
a) Classification, composition, and particulate structure of matter (elements, compounds, mixtures, molecules, atoms, protons, neutrons, electrons) b) Physical and chemical properties of matter c) Mixtures and solutions (solvent, solute, concentration/dilution, effect of temperature on solubility) d) Properties and uses of common acids and bases e) Chemical change (transformation of reactants, evidence of chemical change, conservation of matter, common oxidation reactions – combustion, rusting, tarnishing) f) The role of electrons in chemical bonds
Physics
a) Physical states and changes in matter (explanations of properties in terms of movement and distance between particles; phase change, thermal expansion, and changes in volume and/or b) Energy forms, transformations, heat, and temperature c) Basic properties/behaviors of light (reflection, refraction, light and color, simple ray diagrams) and sound (transmission through media, loudness, pitch, amplitude, frequency) d) Electric circuits (flow of current; types of circuits - parallel/series) and properties and uses of permanent magnets and electromagnets e) Forces and motion (types of forces, basic description of motion, effects of density and pressure)
Earth Science
a) Earth's structure and physical features (Earth's crust, mantle, and core; composition and relative distribution of water, and composition of air) b) Earth's processes, cycles, and history (rock cycle; water cycle; weather versus climate; major geological events; formation of fossils and fossil fuels) c) Earth's resources, their use and conservation (e.g., renewable/nonrenewable resources, human use of land/soil, water resources) d) Earth in the solar system and the universe (phenomena on Earth - day/night, tides, phases of moon, eclipses, seasons; physical features of Earth compared to other bodies)

(continues)

Table A3.1

Panel B: Answer choices for each topic
Choose the response that best describes when the students in this class have been taught each topic
Mostly taught before this year
Mostly taught this year
Not yet taught or just introduced
How well prepared do you feel you are to teach the following science topics?
Not applicable
Very well prepared
Somewhat prepared
Not well prepared

Note: The list of topics comes from the TIMSS 2015 8th-grade science teacher questionnaire and comprises 7 topics in Biology, 6 in chemistry, 5 in physics and 4 in earth science (Panel A). For each topic, teachers are asked when students have been taught a topic and how well they feel prepared to teach it (Panel B).

Table A3.2: Descriptive Statistics – Number of Subject-Specific Qualifications by Major in Education

N. of Subject-Specific Qualifications	Any Major in Education		Total
	No	Yes	
	(1)	(2)	(3)
Zero	4.8	24.9	29.7
One	25.2	15.1	40.2
Two	6	8.4	14.4
Three	2.1	5.8	7.9
Four	1.1	6.7	7.8
Total	39.2	60.8	100

Note: The table reports the weighted share of students taught by teachers who hold zero, one, two, three or four subject-specific qualifications by whether they also hold any major in education (i.e., either major in education, education-science or education-mathematics).

Table A3.3: Descriptive Statistics by Country

	N. of Subject-Specific Qualifications	At Least One Subject-Specific Qualification	Within-Teacher Variation	OECD	Developed	High Perf.	High GNI	Science	# Observations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Australia	1.60	0.86	0.82	Yes	Yes	Yes	Yes	511.6	39,404
Bahrain	1.78	0.95	0.86	No	No	No	Yes	460.3	18,512
Botswana	1.01	0.67	0.67	No	No	No	No	390.4	23,232
Canada	0.76	0.53	0.39	Yes	Yes	Yes	Yes	526.2	35,008
Canada (Ontario)	0.40	0.41	0.26	Yes	Yes	Yes	Yes	524.1	18,080
Canada (Quebec)	1.12	0.70	0.67	Yes	Yes	Yes	Yes	529.5	15,800
Chile	1.19	0.66	0.61	Yes	No	No	No	451.5	17,972
Chinese Taipei	1.17	0.93	0.90	No	No	Yes	No	567.4	21,832
Egypt	1.74	0.77	0.61	No	No	No	No	370.2	31,288
England	1.87	0.97	0.92	Yes	Yes	Yes	No	531.4	14,776
Hong Kong SAR	0.92	0.79	0.79	No	No	Yes	Yes	544.4	16,352
Iran	0.73	0.30	0.18	No	No	No	No	456.4	24,520
Ireland	1.54	0.94	0.93	Yes	Yes	Yes	Yes	529.4	18,808
Israel	2.14	0.92	0.82	Yes	No	Yes	No	505.0	16,716
Italy	1.95	0.95	0.78	Yes	Yes	Yes	No	498.1	17,924
Japan	1.15	0.85	0.78	Yes	Yes	Yes	No	567.6	16,240
Jordan	1.12	0.83	0.77	No	No	No	No	426.1	31,460
Kuwait	1.79	0.90	0.74	No	No	No	Yes	409.8	18,012
Malaysia	1.04	0.76	0.75	No	No	Yes	No	470.8	38,904
Morocco	1.36	0.97	0.97	No	No	No	No	392.8	51,840
New Zealand	1.39	0.92	0.91	Yes	Yes	Yes	No	512.8	32,568
Norway	0.77	0.48	0.38	Yes	Yes	Yes	Yes	509.3	18,364
Norway (8th Grade)	0.83	0.53	0.45	Yes	Yes	No	Yes	488.8	18,724

(continues)

Table A3.3
(continued)

	N. of Subject-Specific Qualifications	At Least One Subject-Specific Qualification	Within-Teacher Variation	OECD	Developed	High Perf.	High GNI	Science	# Observations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Oman	1.82	0.96	0.83	No	No	No	No	454.1	35,532
Qatar	1.82	0.93	0.81	No	No	No	Yes	448.6	20,548
Saudi Arabia	1.14	0.83	0.77	No	No	No	Yes	396.2	15,036
Singapore	1.66	0.95	0.93	No	No	Yes	Yes	596.8	24,464
South Africa	1.58	0.83	0.80	No	No	No	No	356.6	50,056
South Korea	1.00	0.93	0.91	Yes	No	Yes	No	553.9	15,208
Thailand	0.98	0.61	0.52	No	No	No	No	455.8	25,928
Turkey	1.37	0.57	0.44	Yes	No	No	No	492.9	24,316
United Arab Emirates	1.19	0.86	0.83	No	No	No	Yes	470.4	62,716
United Arab Emirates (Abu Dhabi)	1.19	0.84	0.80	No	No	No	Yes	453.3	18,868
United Arab Emirates (Dubai)	1.31	0.89	0.86	No	No	No	Yes	517.4	19,416
United States	0.95	0.71	0.68	Yes	Yes	Yes	Yes	531.6	29,336
All Countries	1.24	0.73	0.66	16	12	17	17	478.3	897,760

Note: The table reports weighted statistics and indicators for each national entity included in the sample. In Column 1, I report the average number of teachers subject-specific qualifications. In Column 2, the share of students taught by teachers who hold at least one subject-specific qualification is reported (i.e., at least one major in either biology, chemistry, physics, or earth science). In Column 3, I report the share of students whose teachers differ in their subject-specific qualifications across subjects (i.e., students who are taught by teachers who have one, two or three subject-specific qualifications). In Columns 4-7, I report country indicators for whether a country belongs to the OECD (Column 4), is a developed country according to the WESP classification (Column 5), is above the median science test score of the countries in the sample (Column 6), or is above the median GNI in 2015 of the countries in the sample (Column 7). The average science test score is reported in Column 8 and the number of observations in Column 9. In the last row, the weighted average of Column 1, 2, 3, and 8 is reported, while the sum of the indicators for Column 4-7 and 9 is reported.

Table A3.4: TIMSS 2011 with Instruction Time

	(1)	(2)	(3)	(4)	(5)
Teacher Subject-Specific Qualifications	0.045***	0.045***	0.043***	0.026***	0.022***
	(0.014)	*(0.011)*	*(0.011)*	*(0.004)*	*(0.004)*
Subject FE	YES	YES	YES	YES	YES
Country FE	YES	YES	YES	NO	NO
Student, School Controls	NO	YES	YES	NO	NO
Teacher Controls	NO	NO	YES	NO	NO
Instruction Time	NO	NO	YES	NO	YES
Student and Teacher FE	NO	NO	NO	YES	YES
Observations	867,012	867,012	867,012	867,012	867,012
R^2	0.37	0.49	0.49	0.94	0.94

Note: The table reports OLS estimation using a set of controls (Column 1,2,3) and student and teacher fixed effects (Column 4 and 5) using TIMSS 2011 data. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights, subject fixed effects, and an imputation dummy for teacher subject-specific qualifications. Student controls include: student SES, gender, language spoken at home, mother's immigrant status, father's immigrant status, student's immigrant status, student's education expectations. School and class controls include class size, share of students with language difficulties, share of economically disadvantaged students, indicator for shortage of resources for science instruction, school discipline problems, school location, school emphasis on academic success. Teacher controls include teacher experience, gender, level of education, major in education. I include instruction time as a control in Column 3 and 5. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A3.5: Sample of Schools Located in Scarcely Populated Areas

	< 30k	< 15k	Small Town/Village
	(1)	(2)	(3)
Teacher Subject-Specific Qualifications	0.043***	0.055***	0.038***
	(0.008)	*(0.011)*	*(0.009)*
Subject, Student, Teacher FE	YES	YES	YES
Observations	320,556	210,072	227,956
R^2	0.94	0.94	0.94

Note: The table reports OLS estimation using subject, student and teacher fixed effects. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights and an imputation dummy for teacher subject-specific qualifications. In Column 1, I report the result for schools located in areas with less than 30,000 inhabitants, in Column 2 in areas with less than 15,000 inhabitants, and in Column 3 for schools located in small towns, villages or rural areas. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A3.6: Analysis of Unobservable Selection and Coefficient Stability following Oster (2019)

	(1)	(2)
Teacher Subject-Specific Qualifications	0.033**	0.035***
	(0.016)	*(0.004)*
Subject FE	YES	YES
Country FE	YES	NO
Student and Teacher FE	NO	YES
Observations	897,760	897,760
R^2	0.33	0.94
Oster (2019) diagnostics		
Bound β^* for $\delta = 1$	0.035	
δ to match $\beta = 0$	19.51	

Note: The table reports OLS estimation using a country (Column 1) and student and teacher fixed effects (Column 2). The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights, subject fixed effects, and an imputation dummy for teacher subject-specific qualifications. The table also reports Oster (2019) diagnostics computed with $R_{max} = 1$ and $\delta = 1$ using TIMSS 2015. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A3.7: Leave One Subject Out

	Full Sample	Excluding Biology	Excluding Physics	Excluding Chemistry	Excluding Earth Science
	(1)	(2)	(3)	(4)	(5)
Teacher Subject-Specific Qualifications	0.035***	0.034***	0.038***	0.036***	0.033***
	(0.004)	*(0.005)*	*(0.005)*	*(0.004)*	*(0.005)*
Subject, Student, Teacher FE	YES	YES	YES	YES	YES
Observations	897,760	673,286	673,326	673,326	673,286

Note: The table reports OLS estimation using subject, student and teacher fixed effects. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights and an imputation dummy for teacher subject-specific qualifications. In Column 1, I report the result for the entire sample. I then replicate the results by excluding one science subject at a time, namely biology (Column 1), physics (Column 2), chemistry (Column 3) and earth science (Column 4). Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A3.8: Leave One Country Out

Excluded Country	Teacher Subject-Specific Qualifications	Std. Error	Observations
	(1)	(2)	(3)
Australia	0.035***	*(0.004)*	858,356
Bahrain	0.035***	*(0.004)*	879,248
Botswana	0.035***	*(0.004)*	874,528
Canada	0.035***	*(0.004)*	862,752
Canada (Ontario)	0.035***	*(0.004)*	879,680
Canada (Quebec)	0.035***	*(0.004)*	881,960
Chile	0.035***	*(0.004)*	879,788
Chinese Taipei	0.035***	*(0.004)*	875,928
Egypt	0.023***	*(0.003)*	866,472
England	0.037***	*(0.004)*	882,984
Hong Kong SAR	0.035***	*(0.004)*	881,408
Iran	0.038***	*(0.004)*	873,240
Ireland	0.035***	*(0.004)*	878,952
Israel	0.035***	*(0.004)*	881,044
Italy	0.035***	*(0.004)*	879,836
Japan	0.039***	*(0.005)*	881,520
Jordan	0.035***	*(0.004)*	866,300
Kuwait	0.035***	*(0.004)*	879,748
Malaysia	0.036***	*(0.004)*	858,856
Morocco	0.037***	*(0.004)*	845,920
New Zealand	0.035***	*(0.004)*	865,192
Norway	0.035***	*(0.004)*	879,396
Norway (8th Grade)	0.035***	*(0.004)*	879,036
Oman	0.035***	*(0.004)*	862,228
Qatar	0.035***	*(0.004)*	877,212
Saudi Arabia	0.036***	*(0.004)*	882,724
Singapore	0.034***	*(0.004)*	873,296
South Africa	0.035***	*(0.005)*	847,704
South Korea	0.037***	*(0.004)*	882,552
Thailand	0.036***	*(0.004)*	871,832
Turkey	0.032***	*(0.004)*	873,444
United Arab Emirates	0.035***	*(0.004)*	835,044
United Arab Emirates (Abu Dhabi)	0.035***	*(0.004)*	878,892
United Arab Emirates (Dubai)	0.035***	*(0.004)*	878,344
United States	0.024***	*(0.004)*	868,424

Note: The table reports OLS estimation using subject, student and teacher fixed effects. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable of interest is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights and an imputation dummy for teacher subject-specific qualifications. In each row, I report the coefficient of teacher subject-specific qualifications obtained estimating Eq. (3.2) by dropping from the estimation sample the country indicated in each row; the corresponding estimated coefficient is reported in Column 1, the standard error of the estimate in Column 2 and the number of observations in Column 3. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A3.9: Different Weights

	Sampling Weights (1)	Senate Weights (2)	Without Weights (3)
Teacher Subject-Specific Qualifications	0.035***	0.022***	0.020***
	(0.004)	*(0.002)*	*(0.002)*
Subject, Student, Teacher FE	YES	YES	YES
Observations	897,760	897,760	897,760

Note: The table reports OLS estimation using subject, student and teacher fixed effects. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include an imputation dummy for teacher subject-specific qualifications. In Column 1, I report the result using the sampling weights. I use "senate weights", i.e., rescaled weights such that each country carries the same weight, in Column 2 and no weights in Column 3. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A3.10: Plausible Values and JRR

	Std. Score	PV1	PV1-PV5	PV1-PV5 & JRR
	(1)	(2)	(3)	(4)
Teacher Subject-Specific Qualifications	0.035***	4.370***	4.343***	4.343***
	(0.004)	*(0.532)*	*(0.597)*	*(0.655)*
Subject, Student, Teacher FE	YES	YES	YES	YES
Observations	897,760	897,760	897,760	897,760

Note: The table reports OLS estimation using subject, student, and teacher fixed effects. The outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score (Column 1), the first subject-specific plausible value (Column 2) and all five subject-specific plausible values (Column 3 and 4). In Column 4, I perform the Jackknife Repeated Replication (JRR) method to account for the sampling variance. The explanatory variable is teacher subject-specific qualifications. An observation corresponds to a student-subject combination. All regressions include weights and an imputation dummy for teacher subject-specific qualifications. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table A3.11: Mediation Analysis

	(1)	(2)
Panel A: Effect of Mediator on Student Test Scores		
Teacher Subject-Specific Qualifications	0.035***	0.028***
	(0.004)	*(0.004)*
Share Topics Confident to Teach		0.050***
		(0.006)
Subject, Student, Teacher FE	YES	YES
Observations	897,760	897,760
R^2	0.94	0.94
Panel B: Effect of Teacher Subject-Specific Qualifications on Mediator		
Teacher Subject-Specific Qualifications	0.142***	
	(0.009)	
Mean (SD) of Dep. Variables	0.54 (0.37)	
Subject, Student, Teacher FE	YES	
Observations	897,760	
R^2	0.64	

Note: The table reports OLS estimation using subject, student and teacher fixed effects. In Panel A, the outcome of interest is the standardized subject-specific (biology, chemistry, physics, and earth science) test score. Test scores have been standardized within each subject. The explanatory variable is teacher subject-specific qualifications. In Column 1, I report the effect of teacher subject-specific qualifications on student test scores. I then include the mediator, the share of topics a teacher feels confident to teach, in Column 2. In Panel B, the outcome of interest is the subject-specific (biology, chemistry, physics, and earth science) share of topics that a teacher feels confident to teach. The explanatory variable is teacher subject-specific qualifications in a subject. In all regressions, an observation corresponds to a student-subject combination. All regressions include weights and an imputation dummy for the explanatory variables. Standard errors (in parentheses) have been clustered at the classroom level. *** p<0.01, ** p<0.05, * p<0.1

4 Can Patience Account for Within-Country Differences in Student Achievement? A Regional Analysis of Facebook Interests*

4.1 Introduction

Human capital theory posits that activities that advance people's education can be understood as investments in skills (Becker 1964). An important implication of this intertemporal aspect is that differences in discount rates should affect educational decisions, behaviors, and outcomes. We therefore suggest that differences in people's time preferences – patience – are an important cause of the large differences in student achievement that exist across different regions in many countries. These achievement differences are important for regional income differences; for example, skill differences account for a substantial share of income differences across U.S. states (Hanushek, Ruhose, and Woessmann 2017). However, the deeper sources of this substantial regional variation in achievement are not well understood. Investigations of whether regional differences in discount rates can account for regional variation in schooling outcomes have been stymied by a lack of region-specific measures of time preference parameters. In this paper, we exploit the massive data available from social media – in particular, Facebook interests – with machine-learning algorithms to derive new measures of regional variations in patience that permit direct assessment of the role of patience in accounting for regional differences in student achievement within countries.

Many countries have large differences in student achievement across regions. In the United States, the difference in the average math achievement of eighth-grade students on the National Assessment of Educational Progress (NAEP) between the top- and bottom-performing state is equivalent to the average learning of three school years (Hanushek, Ruhose, and Woessmann 2017). A similar magnitude is found between the top- and bottom-performing region in Italy on the Istituto Nazionale per la Valutazione del Sistema Dell'Istruzione (INVALSI) test in eighth-grade math. When German states took the international test of the Programme for International Student

* This chapter is joint work with Eric A. Hanushek, Lavinia Kinne, and Ludger Woessmann. It is based on the paper "Can Patience Account for Within-Country Differences in Student Achievement? A Regional Analysis of Facebook Interests", *mimeo*.

Assessment (PISA) in 2000, state differences turned out nearly as large as international differences (Woessmann 2010).

Since the earliest analyses of human capital, it has been recognized that discount rates constitute a fundamental determinant of individual investment decisions. But that is just part of the full impact of time preferences. Patience, the relative valuation of present versus future payoffs, appears in many decisions that relate to human capital investments. At the individual level, students must weigh current gratification such as play time with friends against study time that may lead to deferred rewards in later life. At the group level, communities and societies must trade off present against future costs and benefits when deciding how much to invest in schools, how strongly to motivate children to learn, and whether to design institutional structures to incentivize learning. Variations in patience may be relevant for understanding regional differences in educational achievement because of systematic variations of both individuals and groups across regional populations. However, the regional empirical analysis is impeded by the fact that representative survey measures of economic preferences such as patience are generally not readily available for any distinct regions within countries.

The key methodological innovation of our paper is to use social-media data to derive a measure of patience at the regional level. The underlying idea is that social-media data contain important information about people's underlying preferences such as patience. For marketing purposes, Facebook has developed an algorithm to classify the interests of over two billion people based on their observed behavior on Facebook and beyond. Specifically, self-reported interests, clicks and "likes" on Facebook, software downloads, clicks on advertisements that Facebook places on other sites, and additional inference from overall behavior and location suggest that Facebook interests post a fertile ground for investigating fundamental preferences. Following Obradovich et al. (2022), we use dictionary vocabulary to scrape Facebook's marketing application programming interface (API) in order to derive 1,000 Facebook interests with the largest audience sizes worldwide and then use these as raw data for describing key preference differences.

Our derivation of within-country measures of patience builds on recent advances in the international analysis of culture. Expanding the approach developed by Obradovich et al. (2022), we collect data on the prevalence of Facebook interests in each country and region. After reducing the dimensionality of these by fitting a principal component analysis (PCA), we train an international model to predict the

measure of patience contained in the Global Preference Survey (GPS), which developed scientifically validated measures of various preferences of country populations (Falk et al. 2018). We then use the estimated parameters of this cross-country model to predict patience for within-country regions based on their observed Facebook interests.

We validate these measures of patience by performing an international preference analysis using student achievement data from PISA. First, within the sample of GPS countries, the Facebook-derived measure performs just as well as the original GPS measure (previously used in Hanushek et al. 2022) in predicting student achievement on the international PISA test. Second, out-of-sample prediction from the trained model allows us to expand the country sample from 48 to 80 countries. Results in the expanded sample – as well as in the subsample of 32 new countries – are again very consistent in predicting PISA achievement. Third, both validation results are confirmed in a model that uses the subsample of migrant students and assigns them the preference parameters of their countries of origin, thereby allowing to condition on fixed effects for residence countries to shield against bias from unobserved features of students' residence countries.

We apply our method to measure patience at the regional level in two countries, Italy and the United States. In Italy, the large North-South variation across the 20 regions has raised substantial interest in policy and research (e.g., Putnam 1993; Ichino and Maggi 2000; Guiso, Sapienza, and Zingales 2004). As a large federal country, the United States allows for regional analyses for a large sample of 50 U.S. states. Both countries show substantial regional variation in the Facebook-derived measure of patience with a noteworthy North-South gradient.

We employ the newly derived regional measure of patience in analyses of regional student achievement in the two countries. For Italian regions, we use achievement data from over 200,000 students on the national INVALSI test. For the United States, we use regional achievement data on the national NAEP test. By studying achievement differences for regions within individual countries, the estimation is less prone to confounding from unobserved national traits such as languages, constitutions, and institutional factors that may hamper prior cross-country analyses.

Results indicate that the Facebook-derived measure of patience is strongly associated with student achievement both across Italian regions and U.S. states. In both countries, students in regions with higher levels of patience score significantly higher on the respective achievement tests. In Italy, a one standard deviation (SD) increase

in regional patience is associated with a 1.5 SD increase in eighth-grade math achievement, which is only slightly smaller than the estimate obtained in the cross-country analysis. In the United States, the equivalent estimate is only about one quarter in size.

In both countries, regional differences in patience account for substantial parts of the subnational variation in student achievement. The models account for over two thirds of the variation in test scores across Italian regions and for over one third across U.S. states. The smaller role in the United States may reflect that the substantial internal mobility of the U.S. population across states might introduce attenuation bias in the regional measurement of intergenerationally transmitted cultural traits.

Consistent with skill development as a cumulative process, the estimated association of patience with student achievement increases across grade levels. In the Italian INVALSI tests, estimates grow steadily across the four testing occasions from second to tenth grade. Similarly, estimates for the U.S. NAEP are smaller in fourth than in eighth grade.

Results are stable in a series of robustness analyses such as using reading achievement or the regionally representative participation of Italy in PISA 2012. Throughout, our analysis conditions on regional variation in risk-taking, another preference parameter that can partly capture intertemporal aspects. However, the machine-learning model to predict risk-taking from Facebook interests does not perform very well at the regional level. As patience and risk-taking tend to be positively associated and prior work suggests a negative association of risk-taking with student achievement, the poor measurement of risk-taking may imply that the estimates of patience reflect lower bounds.

Our analysis contributes to two strands of literature. First, we contribute to the analysis of the role of time preferences in human capital investment. Our regional analysis adds a new perspective to the literature that has studied the role of patience for educational outcomes at the individual level (Sutter et al. 2013; Golsteyn, Grönqvist, and Lindahl 2014; Castillo, Jordan, and Petrie 2019) and at the international level (Figlio et al. 2019; Hanushek et al. 2022). Additionally, cross-country work has shown the importance of patience for long-run comparative economic development (Galor and Özak 2016; Sunde et al. 2022). In deriving the regional patience measure, our approach also contributes to the literature that uses Facebook data to measure various concepts of culture and social networks (e.g., Obradovich et al. 2022; Chetty et al. 2022; Bailey et al. 2022), as well as to the literature on culture

and economic outcomes more broadly (e.g., Guiso, Sapienza, and Zingales 2006; Alesina and Giuliano 2015). Second, the consideration of patience contributes a new perspective of deeper causes to the study of regional differences in student achievement. While there are a few studies on proximate causes such as family background, school spending, and institutional settings (e.g., Hanushek and Raymond 2005; Woessmann 2010; Dee and Jacob 2011), most stop at just noting the magnitudes of regional differences without providing convincing explanations of them (e.g. Hanushek 2016).

The remainder of the paper is structured as follows. Section 4.2 describes our method to derive regional measures of patience from data on Facebook interests and includes a validation exercise at the cross-country level. Section 4.3 describes the regional student achievement data. Section 4.4 presents our results. Section 4.5 concludes.

4.2 Methods: Deriving Regional Patience Measure from Facebook Interests

We use social-media data to measure patience at the regional level. Section 4.2.1 introduces the Facebook interest data. Section 4.2.2 validates the suitability of these interests to predict international differences in patience. Section 4.2.3 describes our method to derive regional measures of patience from the Facebook interests.

4.2.1 Facebook Interests

With 2.9 billion monthly active users, Facebook is the world's largest social network.[1] Facebook's core business consists of selling advertising space on its social media platform. In 2021, 97.5 percent of Facebook's revenues came from advertisements.[2] Hence, Facebook's business model depends primarily on its ability to keep users engaged on the platform while advertisers promote their products and services to users who may find them relevant. To this purpose, Facebook puts considerable effort into inferring users' interests (Thorson et al. 2021).

[1] Source: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (last accessed 23 February 2023).

[2] Figures about Facebook's users and revenues are reported by Meta, Facebook's parent company, drawing on the third-quarter 2022 results (https://s21.q4cdn.com/399680738/files/doc_financials/2022/q3/Meta-09.30.2022-Exhibit-99.1-FINAL.pdf, last accessed 2 January 2023) and the 2021 annual report (https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/14039b47-2e2f-4054-9dc5-71bcc7cf01ce.pdf, page 58, last accessed 2 January 2023).

Facebook determines users' interests using a variety of sources, both inside the Facebook platform as well as on external websites (Cabañas, Cuevas, and Cuevas 2018; Obradovich et al. 2022). Inside the Facebook platform, these sources include personal information that users share on Facebook as well as users' activity on Facebook, such as page likes, group membership, and content users engage with. Outside the platform, Facebook tracks users' visited websites, installed apps, and purchasing behavior.[3] Facebook uses these data to deliver content and recommendations based on users' interests and to allow advertisers to target users whose interests are relevant for the products or services that they want to sell.[4]

The hundreds of thousands of interests classified by Facebook are organized in nine main categories: business and industry, entertainment, family and relationships, fitness and wellness, food and drink, hobbies and activities, shopping and fashion, sports and outdoors, and technology. Interests can be very broad, such as "Entertainment" or "Music", or very narrow, such as "Caribbean Stud Poker", a casino table game. Figure 4.1 shows the 1,000 Facebook interests with the largest worldwide audience, where larger font sizes correspond to larger audience sizes. Interests often relate to leisure activities such as sports and beauty, but also to broader categories such as education and politics.

Following Obradovich et al. (2022), we proceed in two steps to retrieve data on the Facebook interests for countries and subnational entities. First, we obtain a comprehensive list of Facebook interests by querying the Facebook Marketing API, the interface that allows advertisers to configure their advertisement campaigns. For any given text input (query), a tool within the API returns a collection of the respective closely related Facebook interests together with their estimated worldwide audience and a unique identifier, which makes them language-independent. We iteratively feed this function with all 25,322 terms of an English dictionary[5] and 2,000 randomly

[3] While official figures on Facebook's off-platform data collection are not available, Aguiar et al. (2022) estimate that, for a representative sample of 5,000 U.S. internet users in 2016, Facebook can track 55 percent of websites visited by Facebook users, which amounts to 41 percent of browsing time. For more information on this practice, see also Facebook's official press release on data collection outside of Facebook at https://about.fb.com/news/2018/04/data-off-facebook/ (last accessed 2 January 2023).

[4] Facebook users can access the interests that Facebook assigns to them. According to a recent report, 59 percent of Facebook users in the US say that these Facebook interests reflect their real-life interests (https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/, last accessed 23 February 2023).

[5] We use a dictionary of popular English words available at https://github.com/dolph/dictionary/blob/master/popular.txt (last accessed 3 January 2023).

selected titles of Wikipedia articles, each of which can yield several Facebook interests. After removing duplicates, we obtain a collection of 41,513 unique interests from this procedure.

Second, we select the 1,000 interests with the largest worldwide audience obtained in the previous step, which ensures cross-country and within-country comparability. For each of these 1,000 interests, we again use the tool from Facebook's Marketing API to separately obtain the estimated audience size for each country in which Facebook has a presence, as well as for each state in the U.S. and region in Italy. For each geographical entity, this process yields a vector of size 1,000 with the estimated audience for all of the 1,000 largest interests by worldwide audience. Finally, we divide the estimated audience by the 2020 population size in each geographical entity to obtain the share of individuals holding each interest.

4.2.2 Using Facebook Interests to Measure Patience: A Cross-Country Validation Exercise

To assess the suitability of the Facebook interest data to measure patience, we perform a cross-country validation exercise which proceeds in four steps. First, we reduce the dimensionality of the Facebook data. Second, we study how well the reduced-dimensionality Facebook data predicts an external measure of patience available at the country level in the Global Preference Survey (GPS). Third, after training the prediction model within the sample of GPS countries, we perform out-of-sample predictions to expand the country sample to countries that are not part of the GPS. Fourth, we use the international PISA test data to validate whether the Facebook-derived measure of patience is associated with student achievement across countries both within and outside the sample of countries participating in the GPS.

We start by reducing the dimensionality of the country-level Facebook interests by a principal component analysis (PCA) fitted on the international sample of all 216 countries and geographical entities featured by Facebook. On top of reducing the dimensionality of the variables that we use to later train the machine-learning model, this step also avoids collinearity problems because the resulting principal components are uncorrelated by construction. The first 10 principal components (PCs) capture 70 percent of the total cross-country variance contained in the Facebook interests, the first 20 PCs capture 80 percent, and the first 48 PCs capture 90 percent.[6] While the additional variance captured by any PC beyond the 10th PC is quite

[6] Details are provided in Appendix Figure A4.1.

small, this still suggests that many PCs are required to capture the full variance in Facebook interests across countries (see also Obradovich et al. (2022)).

Next, we train a machine-learning model to learn the relationship between the country-level PCs of the Facebook interests and an external measure of the countries' patience. As an external measure, we use the measure of patience contained in the GPS, which collected survey-based measures of several preference parameters from representative samples in 76 countries (Falk et al. 2018). The measure of patience combines a qualitative survey item and a hypothetical choice scenario that were chosen based on their predictive capacity for incentivized choices in an ex-ante laboratory setting. Our training sample includes 74 countries, namely all the countries that participated in the GPS survey except for Iran and Russia, for which Facebook data are currently not available.[7] We use a 10-fold cross-validated least absolute shrinkage and selection operator (LASSO) model for the cross-country training. The performance of the model is quite satisfactory: Independent of whether 10, 20, 30, 40, 50, or even 100 PCs are used, the R^2 of the in-sample prediction of patience by the reduced-dimensionality Facebook interests is quite stable between 0.65 and 0.70.[8]

We use the parameter estimates of the machine-learning model to make out-of-sample predictions of patience for all countries that participated in at least one PISA wave and for which Facebook interests can be retrieved. Given the limited size of the sample used to train the machine-learning model, we prefer the most parsimonious specification with 10 PCs for the out-of-sample predictions to avoid overfitting.[9] The resulting sample for which we have both Facebook-derived patience measures as well as student test scores consists of 80 countries.[10]

We perform the same training and prediction models for risk-taking, another intertemporal preference contained in the GPS that has been used to study international student achievement. The R^2 of the in-sample prediction for risk-taking

[7] A list of the countries is shown in Column 4 of Appendix Table A4.1.

[8] Details are provided in Appendix Figure A4.2.

[9] Less parsimonious models tend to obtain better in-sample performance (although this is hardly the case for patience, see Appendix Figure A4.2) but can lead to worse out-of-sample performance especially with small samples.

[10] The countries are reported in columns 1 and 3 of Appendix Table A4.1.

is somewhat lower than for patience,[11] which suggests that risk-taking is harder to predict from Facebook interests compared to patience in the cross-country setting.

To validate our Facebook-derived measures of patience and risk-taking, we estimate their relationship with student achievement across countries. The model setup for the validation follows Hanushek et al. (2022), using math achievement on the PISA test over all seven available waves 2000-2018 to estimate the following OLS model:

$$T_{ict} = \beta_1 Patience_c + \beta_2 Risk_c + \alpha_1 B_{ict} + \mu_t + \varepsilon_{ict} \quad (4.1)$$

where T, the standardized PISA test score of student i in country c in year t, is a function of the country-level measures of patience and risk-taking of country c, a vector of control variables B (student gender, age, and migration status), and an error term ε_{ict}. Fixed effects for test waves μ_t account for time trends and idiosyncrasies of the individual tests. The coefficients of interest are β_1 and β_2 which characterize the relationship of patience and risk-taking with student achievement. Regressions are weighted by students' sampling probability, giving equal weight to each country. Standard errors are clustered at the country level.

The Facebook-derived measures of patience and risk-taking perform very well in the cross-country validation exercise. As a baseline, the first column of Panel A of Table 4.1 shows that patience has a strong and significant positive relationship with student achievement when using the original GPS measure, whereas risk-taking has a strong and significant negative relationship.[12] Column 2 substitutes the GPS measures of patience and risk-taking with our Facebook-derived measures, using the same sample of countries.[13] The results are very much in line with those obtained using the original GPS measures, which corroborates the validity of the Facebook-derived measures. Point estimates are in fact slightly larger (in absolute terms) than the original estimates.[14] The out-of-sample predictions allow us to extend the analysis of the Facebook-derived measures of patience and risk-taking from a sample of 48 to 80

[11] See Appendix Figure A4.2.

[12] This model replicates the main estimates of Hanushek et al. (2022) after dropping Russia (which has no Facebook data), with estimates hardly changed (see column 3 of their Table 1).

[13] The measures are obtained with 10 PCs of Facebook interest. Appendix A.1 shows that results are very similar when using additional (20-50) PCs to derive the measures.

[14] The coefficient on patience in column 2 of Table 4.1 is significantly larger than in column 1 in the cross-country analysis, whereas all other differences between columns 1 and 2 are statistically insignificant.

countries – all countries that participated in PISA and have Facebook data – encompassing over 2.6 million student observations. Results generalize very well to the extended sample, with increased precision and without significantly different estimates (Column 3). Even in the sample of 32 countries that were not part of the original GPS analysis, results are qualitatively the same and statistically highly significant (Column 4).

In the international analysis, we can also perform a migrant analysis that aims to account for unobserved differences across residence countries. The analysis restricts the sample to students with a migrant background and assigns them the values of patience and risk-taking of their home countries (see Figlio et al. (2019); Hanushek et al. (2022)). By observing migrant students from different countries of origin who are schooled in the same residence country, this setup allows to take out fixed effects of the residence countries (as well as their full interaction with wave fixed effects), thereby excluding the possibility that the relationships are driven by other factors of the country of schooling.

The migrant analysis further validates the informational content of the Facebook-derived measures. Results in Panel B of Table 4.1 show that again, the positive patience relationship and the negative risk-taking relationship again replicate very well when using the Facebook-derived rather than the original GPS measures.[15] The risk-taking coefficient is somewhat less precisely estimated but actually increases in (absolute) size. Estimates become quite imprecise (and larger) when restricting the sample to non-GPS countries (Column 4), indicating limited power of the migrant analysis in the smaller sample.

Overall, the cross-country validation exercise shows that the measures of patience and risk-taking predicted using the Facebook data follow very closely the patterns from externally validated survey measures of these preferences. This implies that the information contained in the Facebook interests and their underlying principal components are suitable to infer such measures for geographical entities that do not have representative measures from surveys.

[15] With the Facebook data, we expand the countries of origin considered in the migrant analysis from 56 to 93 (see Appendix Table A4.2). The destination countries increase only from 46 to 50 because other PISA countries do not report students' and parents' country of birth required to determine migrants' country-of-origin preferences.

4.2.3 Predicting Regional Patience from Reduced-Dimensionality Facebook Interests

Our method to derive measures of patience for subnational regions from the Facebook interests, which extends the method developed by Obradovich et al. (2022) to our regional analysis, proceeds in three steps. First, we again reduce the dimensionality of the Facebook interests using a PCA, but this time fitting the PCA across the regions *within* a given country. Second, we use the PC loadings obtained from the within-country PCA to reduce the dimensionality of the country-level Facebook interests in the international sample. This allows us to train a machine-learning model that learns the relationship between these country-level PCs and the survey-based measure of patience contained in the GPS. Third, we use the parameter estimates from the internationally trained machine-learning model with the PC loadings derived from fitting the PCA at the regional level to make out-of-sample predictions of patience for the subnational regions based on their Facebook interests.

We fit the PCA to reduce the dimensionality of the Facebook interests separately within the two countries we study, i.e., for Italian regions and for U.S. states. Fitting the PCA at the regional level ensures that the PCs capture variance in dimensions of Facebook interests that are relevant at the regional level within the specific country. For the Italian regions, the first 10 PCs already capture 90 percent of the within-country variance in Facebook interests.[16] For the U.S. states, the same portion of variance is captured by the first 15 PCs. In both cases, each subsequent PC only captures a small portion of variance.

To train a prediction model of the country-level patience measures, we first apply the respective within-country PCA to the international sample. That is, we use the PC loadings obtained in the previous step for dimensionality reduction of the country-level Facebook interests. Because these PC loadings capture the contribution of the regional-level Facebook interests to the PCs, the resulting country-level PCs will preserve the respective variance that can be found in Facebook interests across Italian regions or U.S. states. We then use these PCs to train a 10-fold cross-validated LASSO model to learn the relationship between the PCs and the GPS measure of patience

[16] Details are provided in Appendix Figure A4.3 and Figure A4.4 for Italy and the United States, respectively.

across countries.[17] Since the country-level PCs are now constructed to resemble the regional-level variance in Facebook interests, the model should be capable of generalizing the estimated relationship between country-level PCs and countries' GPS measures to Italian regions and U.S. states.

The in-sample performance of the model in predicting the GPS measure of patience is relatively good, both when the PC loadings are derived from fitting the PCA on the Facebook interests of Italian regions and of U.S. states. Few PCs already capture a considerable portion of the variation in Facebook interests within countries: with 10 PCs, the R^2 of the in-sample prediction reaches 0.5 in the case of Italian regions and over 0.6 in the case of U.S. states.[18] In both cases, increasing the number of PCs and, hence, the amount of variance used, is accompanied by an increase in the in-sample performance of the model, but we again prefer more parsimonious models for the out-of-sample predictions to avoid overfitting.

We then derive regional measures of patience by using the parameter estimates from the internationally trained model to predict patience from the Facebook interests observed in Italian regions and U.S. states, respectively.

Figure 4.2 contains maps that show the regional variation of the Facebook-derived measure of patience in Italy and the United States.[19] In Italy, the regions in the lowest deciles of patience are Sicily and Campania in the South. The region with the highest level of patience is Trentino-Alto-Adige, located in the North-East. Interestingly, parts of Trentino-Alto-Adige belonged to Austria and the former Austro-Hungarian empire for long periods of time, and large parts of the population in the region speak German as their first language. According to the country-level GPS measures, Austria has a much higher level of patience than Italy.[20] The fact that this region exhibits the largest level of patience thus bodes well for the Facebook-derived measure. In the United

[17] The GPS measure is standardized to have mean zero and standard deviation one across individuals in the 76 countries participating in the GPS, so that estimates in our subsequent analysis can be interpreted in terms of standard deviations.

[18] Details are provided in Appendix Figure A4.5 and Figure A4.6 for Italy and the United States, respectively.

[19] The figure shows values obtained with 4 PCs; patience measures obtained with different numbers of PCs yield the same graphical representation.

[20] The country-level GPS measure of patience for Austria (0.61) is half a standard deviation higher than for Italy (0.11). A similar argument can be made for the Aosta Valley region in the North-West of Italy, whose culture is deeply intertwined with neighboring France. France's GPS measure of patience is a quarter of a standard deviation higher than Italy's.

States, the states that exhibit the highest level of patience are Vermont and Maine in the North-East. Both countries tend to show a North-South gradient in the Facebook-derived measure of patience.

When performing the same prediction analysis for risk-taking, the performance of the prediction model is substantially worse. Both for Italian regions and for U.S. states, the R^2 of the in-sample prediction is well below 0.2 for all models with up to 10 PCs and well below 0.4 even for a model with 20 PCs.[21] We include the measure of risk-taking as a control variable in our regional analysis throughout.[22] However, its poor measurement when PC loadings are fitted at the regional level means that the estimates on patience are likely lower bounds because patience and risk-taking are positively associated and risk-taking has the opposite sign from patience in the cross-country analysis (Hanushek et al. 2022).

4.3 Data on Regional Student Achievement

To estimate the association of patience with student achievement for subnational regions, we use data on the largest student assessments for Italy and the United States, respectively, that are both representative at the regional level: INVALSI (Section 4.3.1) and NAEP (Section 4.3.2).

4.3.1 Italy: INVALSI

Since 2007, the Istituto Nazionale per la Valutazione del Sistema Dell'Istruzione (INVALSI) assesses a random sample of Italian students in math and Italian every year. Furthermore, INVALSI administers student, teacher, and principal questionnaires to collect background information about the educational environment. We use data on math achievement in the school years 2017-2018 and 2018-2019, the last years before the COVID-19 pandemic. In our main analysis, we focus on eighth-grade students since they are closest in age to the students in PISA and NAEP. The sample of eighth-graders consists of 59,034 students. In additional analyses, we also use data for students in grades 2, 5, and 10, with an entire sample size of 235,661 students.

[21] See Appendix Figure A4.5 and Figure A4.6. The performance with 20 PCs is a spike that likely reflects overfitting of the data in this case.

[22] See Appendix Figure A4.7 for maps depicting the regional distributions of risk-taking in Italy and the United States, but these should be interpreted with care because of the poor performance of the prediction model.

The random sample of students is drawn following a two-step procedure, where a varying number of classes is randomly selected within a random sample of schools stratified at the regional level. Crucially for our analysis, the sample is representative at the regional level for 19 of the 20 regions in Italy (Falorsi, Ricci, and Falzetti 2019). The exception is Trentino-Alto-Adige, where only students in the autonomous municipalities of Bolzano and Trento are tested. The difference between the lowest and highest performing region in Italy in 8th-grade math amounts to roughly three quarters of a standard deviation, equivalent to the average learning of almost three school years.

In robustness checks, we complement the INVALSI analysis using Italian data from PISA 2012 where Italy oversampled students in each region to obtain a representative sample of students.

4.3.2 United States: NAEP

We use data from the National Assessment of Educational Progress (NAEP), the largest nationally representative assessment of students in the United States. In our main analysis, we focus on NAEP mathematics test scores in grade eight. We use mathematics test scores for each state using data from the last three waves of NAEP before the COVID-19 pandemic, namely NAEP 2015, 2017 and 2019. The resulting dataset consists of state-level test scores for the 50 U.S. states and the federal district of Washington, D.C. Approximately 140,000 students take part in a typical NAEP assessment.[23] In additional analyses, we also use data on fourth-grade students. Also in the United States, the difference between the lowest and highest performing state in 8th-grade math is equivalent to roughly three years of schooling.

We divide both INVALSI and NAEP test scores by the student-level standard deviation in the respective country, so that regression coefficients can be interpreted in terms of standard deviations.

4.4 Results

We use our regional measure of patience derived from Facebook interests to study whether differences in patience can account for the substantial differences in student

[23] Source: https://nces.ed.gov/nationsreportcard/guides/statsig.aspx#:~:text=A%20NAEP%20national%20assessment%20typically,samples%20of%20approximately%20140%2C000%20students (last accessed 23 February 2023).

achievement that exist across Italian regions and U.S. states. The estimated models are versions of equation (4.1) applied to the regional rather than the country level.[24] Compared to the cross-country analysis, the within-country analysis is less prone to bias that may arise from national factors such as languages, laws, and institutional settings. In this section, we report our results for Italy (Section 4.4.1) and the United States (Section 4.2), followed by robustness analyses (Section 4.4.3).

4.4.1 Italy

Italy represents an interesting case study for the regional analysis because of its well-known North-South divide in many dimensions, including student test scores. This regional divide is surprising given the relatively centralized structure of the country: the schooling system is regulated mostly at the country level, having the same structure across regions.[25] Hence, the within-country association between patience and student test scores is unlikely to be severely biased by institutional factors.

The Facebook-derived regional measure of patience is strongly and significantly associated with student achievement across Italian regions. Panel A of Table 4.2 shows results of student-level analyses of math achievement in eighth grade using patience measures obtained with 4, 7, and 10 PCs of Facebook interest, which showed good in-sample performance in Section 4.2.3. Irrespective of the number of PCs used to derive the patience measure, the coefficient estimates are highly significant and indicate that a one standard deviation (SD) increase in patience is associated with an increase in math test scores by 1.35-1.51 SD. The Italian regional estimates are only slightly smaller than the cross-country estimates reported in Table 4.1.

When estimated at the regional level, results suggest that regional differences in patience can account for at least two thirds of the variation in student achievement across Italian regions. Using student test scores aggregated to the regional level in Panel B of Table 4.2, point estimates are very similar, albeit slightly smaller than in the student-level analysis. The R^2 indicates that the model accounts for 0.68-0.80 of the

[24] The model specification is very parsimonious as we think of patience as a deep determinant of student achievement. Proximate inputs often included in education production functions such as parental education or school resources would be bad controls in this setting as they are endogenous to regions' patience.

[25] The matters in which the state has exclusive legislation are listed in Article 117 of the Italian Constitution (https://www.governo.it/it/costituzione-italiana/parte-seconda-ordinamento-della-repubblica/titolo-v-le-regionile-province-e-i; last accessed 30 January 2023).

region-level variation, indicating that patience accounts for a large portion of the differences in student achievement across Italian regions.

Interestingly, the association of patience with student achievement increases strongly with increasing grade levels. Panels A and B of Table 4.3 show results for all four grade levels available in INVALSI for the patience measure obtained with 4 PCs of Facebook interests.[26] Column 3 replicates our main results from the previous table that refer to students in grade 8. The other columns show results for students in grades 2, 5, and 10, respectively. Coefficient estimates increase continuously from an insignificant 0.29 SD in grade 2 to a highly significant 1.77 in grade 10 when estimated at the student level. Region-level estimates are again quite similar. These results suggest that as educational investments are cumulative, the role of patience keeps adding up across grades.

4.4.2 United States

As a large federal country, the United States provide a large regional sample of 50 states plus Washington, D.C that feature large differences in student outcomes.[27] With data accessible only at the state level, Panel C of Table 4.2 reports the results of our state-level regressions. The analysis again refers to math achievement in 8th grade and uses Facebook-derived measures of patience obtained with 4,7, and 10 PCs.

Also in the United States, patience is significantly associated with higher student achievement at the regional level. A one SD increase in the Facebook-derived measure of patience is associated with an increase of 0.17-0.29 SD in test scores across U.S. states. The point estimates are only about a quarter of the ones estimated for Italian regions. The model accounts for slightly more than one third of the variation in test scores across U.S. states.

While patience plays an important role in accounting for cross-state differences in student test scores in the United States, the role is less prominent than in Italy. A possible explanation is that the population in the United States is substantially more mobile and mixed. According to census estimates, 42 percent of the U.S. population lives in a state different from their state of birth.[28] Because cultural traits such as

[26] Results are very similar when using 7 or 10 PCs (not shown).

[27] Results are similar when excluding Washington, D.C. from the analysis (not shown).

[28] Own calculations based on the ACS 2019 table of state of residence by place of birth available at https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-of-residence-place-of-birth-acs.html (last accessed 25 February 2023).

patience are mostly transmitted across generations (e.g., Bisin and Verdier 2011; Alesina and Giuliano 2014), such an extent of internal migration makes cultural traits harder to measure at the state level. This might induce measurement error in the estimates of patience and cause attenuation bias in the regressions.

Consistent with the Italian evidence, the association between patience and student achievement is smaller in lower grades also in the United States. While also statistically significant, the coefficient estimate in 4th grade is only about half the size as in 8th grade (Panel C of Table 4.3), corroborating that the role of patience adds up as educational efforts accumulate.

4.4.3 Robustness Analysis

Results prove stable in a series of robustness analyses. Both in Italy and the United States, we find similar results for reading achievement, with slightly smaller point estimates. Results are also robust in the separate waves available in both countries. They show similarly for girls and boys, with no significant gender difference.

The availability of individual-level data for Italy allows for additional in-depth analyses. Consistent with a leading role of culture, estimates are larger for native students than for migrant students. Results are robust to excluding Trentino-Alto-Adige whose sample is not representative for the entire region and whose German-language population might limit comparability. Results are also robust in an Oster (2019) analysis of unobservable selection and coefficient stability. Furthermore, results are remarkably similar when using Italian regional performance on the PISA 2012 test. Appendix A provides the details of these robustness analyses, together with the respective estimation tables.

4.5 Conclusion

Regional differences in student achievement are poorly understood and understudied. In this paper, we deploy social-media-derived measures of time preferences to provide evidence that patience can account for large portions of such differences. We first show that our Facebook-derived measures perform just as well as scientifically validated survey measures of patience and risk-taking when studying cross-country differences in student achievement. We leverage the broader coverage of our new measures to show that patience and risk-taking are strongly associated with student test scores in a much larger sample of countries than previously studied.

In our regional analysis of Italy and the United States, we test the extent to which patience can account for differences in student achievement across regions. We find that even within countries, where schooling systems and educational inputs tend to be more homogenous than between countries, patience is strongly positively associated with student test scores. The model can account for over two thirds of the regional variation in student achievement in Italy and over one third in the United States.

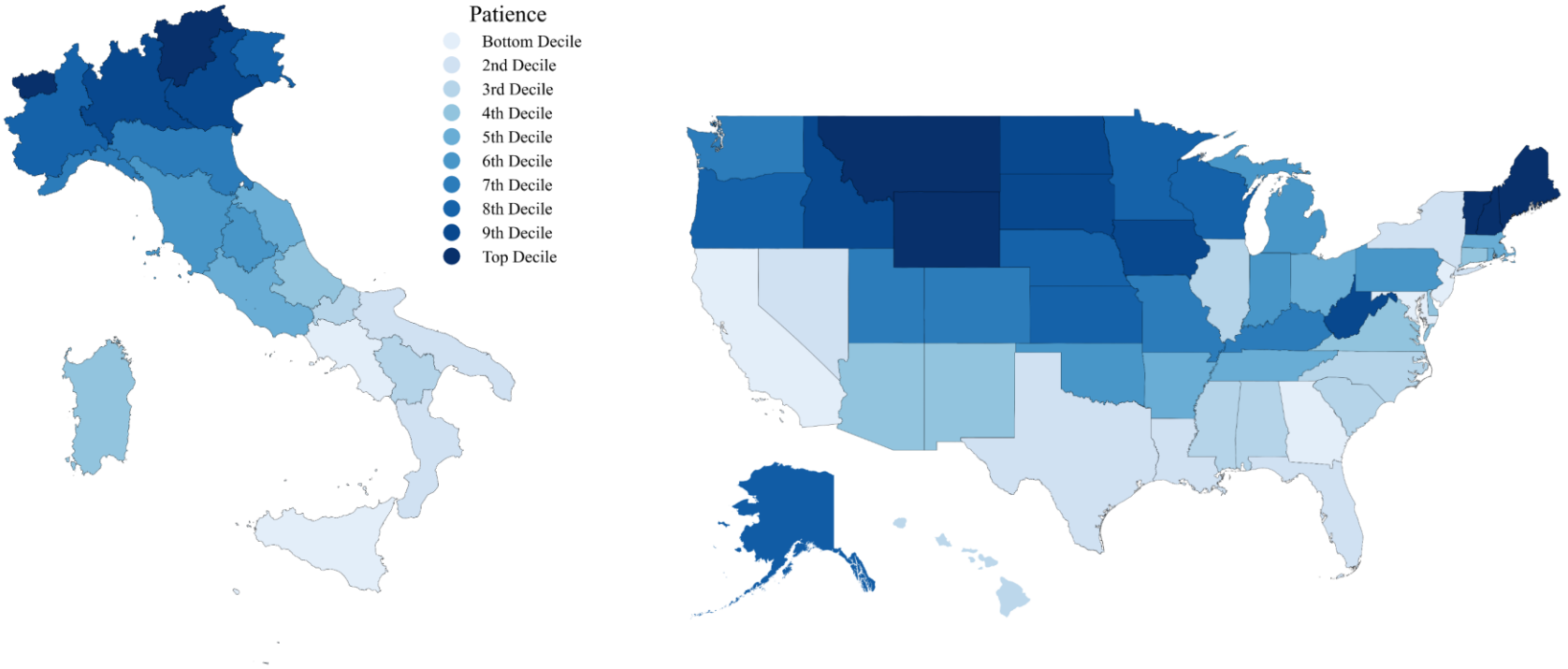
Our findings imply that due to differences in patience, similar educational inputs can lead to substantially different outcomes. When addressing within-country differences in student achievement, policymakers should therefore take possible differences in patience into account. While cultural traits are considered hard to change (e.g., Guiso, Sapienza, and Zingales 2006; Bisin and Verdier 2011), recent evidence shows that traits such as patience are malleable, especially at a young age, and can be improved through specific interventions (e.g., Bird 2001; Alan and Ertac 2018; Jung, Bharati, and Chin 2021). Hence, policies aimed at increasing patience seem a promising avenue to address regional deficits in student outcomes.

Figures and Tables

Figure 4.1: Word Cloud of Facebook Interests

Notes: Word cloud of 1,000 Facebook interests with largest worldwide audience. Larger font sizes indicate larger audience

Figure 4.2: Measure of Patience Derived from Facebook Interests for Italian Regions and U.S. States



A. Italian Regions **B. U.S. States**

Notes: The figures show maps of the Facebook-derived measure of patience obtained with 4 PCs for Italian regions (Panel A) and U.S. states (Panel B), respectively. Each color corresponds to a decile of the distribution of patience within each country. Darker colors denote higher levels of patience.

Table 4.1: Patience, Risk-taking, and Student Achievement: Cross-Country Validation Exercise

	GPS measure	Facebook measure (10 PCs)		
		Original sample	Extended sample	Non-GPS sample
	(1)	(2)	(3)	(4)
A. Cross-country analysis				
Patience	1.225***	1.673***	1.712***	1.761***
	(0.132)	*(0.134)*	*(0.118)*	*(0.209)*
Risk-taking	-1.229***	-1.336***	-1.507***	-1.625***
	(0.188)	*(0.304)*	*(0.249)*	*(0.378)*
Control variables	Yes	Yes	Yes	Yes
Observations	1,954,840	1,954,840	2,660,408	705,568
Residence countries	48	48	80	32
R^2	0.200	0.210	0.220	0.241
B. Migrant analysis				
Patience	0.957***	0.805***	0.902***	1.766***
	(0.115)	*(0.182)*	*(0.205)*	*(0.481)*
Risk-taking	-0.315**	-0.677**	-1.221***	-3.531***
	(0.124)	*(0.278)*	*(0.350)*	*(0.549)*
Control variables	Yes	Yes	Yes	Yes
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes
Observations	78,403	78,403	90,983	12,580
Countries of origin	56	56	93	37
Residence countries	46	46	50	34
R^2	0.280	0.272	0.298	0.310

Notes: Dependent variable: PISA math test score. Least squares regressions. Panel A: all PISA waves 2000-2018; weighted by students' sampling probability. Panel B: waves 2003-2018; students with both parents not born in the country where the student attends school; including 180 fixed effects for each residence-country by wave cell. Control variables: Panel A: student gender, age, and migration status; imputation dummies; and wave fixed effects; Panel B: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level (migrant analysis: country of origin) in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; Falk et al. (2018); own elaboration of Facebook data.

Table 4.2: Patience and Student Achievement: Regional Analysis for Italy and the United States

	4 PCs (1)	7 PCs (2)	10 PCs (3)
A. Italy (individual level)			
Patience	1.505***	1.350***	1.437***
	(0.197)	*(0.114)*	*(0.117)*
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	59,034	59,034	59,034
Regions	20	20	20
R^2	0.092	0.099	0.099
B. Italy (regional level)			
Patience	1.246***	1.134***	1.207***
	(0.193)	*(0.095)*	*(0.099)*
Wave fixed effects	Yes	Yes	Yes
Observations	42	42	42
Regions	20	20	20
R^2	0.679	0.790	0.795
C. United States (state level)			
Patience	0.293***	0.172*	0.285**
	(0.089)	*(0.096)*	*(0.132)*
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
Regions	51	51	51
R^2	0.360	0.348	0.360

Notes: Dependent variable: Panels A and B: INVALSI 8th-grade math test score in waves 2018 and 2019; Panel C: NAEP 8th-grade math test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination; Panel C: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional (state) level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

Table 4.3: Patience and Student Achievement at Different Grade Levels

	Grade 2	Grade 4/5	Grade 8	Grade 10
	(1)	(2)	(3)	(4)
A. Italy (individual level)				
Patience	0.291	0.534*	1.505***	1.767***
	(0.193)	*(0.286)*	*(0.197)*	*(0.236)*
Control variables	Yes	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes	Yes
Observations	48,812	50,608	59,034	77,207
Regions	20	20	20	20
R^2	0.028	0.032	0.092	0.151
B. Italy (regional level)				
Patience	0.182	0.365	1.246***	1.466***
	(0.202)	*(0.237)*	*(0.193)*	*(0.247)*
Wave fixed effects	Yes	Yes	Yes	Yes
Observations	42	42	42	42
Regions	20	20	20	20
R^2	0.044	0.075	0.680	0.678
C. United States (state level)				
Patience	–	0.156**	0.293***	–
		(0.064)	*(0.089)*	
Wave fixed effects		Yes	Yes	
Observations		153	153	
Regions		51	51	
R^2		0.158	0.360	

Notes: Dependent variable: Panels A and B: INVALSI math test score in waves 2018 and 2019; Panel C: NAEP math test score in all NAEP waves 2015-2019. Grade level indicated in column headers (col. 2 refers to 5th grade in Italy and 4th grade in the United States). Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination; Panel C: state-wave combination. Patience measure computed with 4 principal components (PCs). Regressions include the risk-taking measure computed with 4 PCs. Controls cariables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional (state) level in parentheses. Significance level: ***1 percent, **5 percent, *10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data

Appendix

Appendix A: Robustness Analysis

This appendix reports a series of robustness checks for the cross-country validation exercise (Appendix A.1), for the analysis of Italian regions (Appendix A.2), and for the analysis of the U.S. states (Appendix A.3). The analysis of the cross country-validation exercise shows that results do not depend on the specific procedure used to derive the measures of patience and risk-taking. For Italy and the United States, the analysis shows that results are robust to different student outcomes and across various subsamples. The availability of individual-level data for Italy allows a more in-depth analysis than for the United States, where the analysis is constrained by the regional-level data.

A.1 Cross-Country Validation Exercise

To make sure that the results of the validation exercise in Section 4.2.2 do not depend on the specific way of predicting patience and risk-taking from the Facebook data, we present results for alternative predictions that vary the number of PCs used in the LASSO that predict patience and risk-taking from the Facebook interests. Table 4.1 in the main text shows results using the first 10 PCs resulting from the PCA performed on the international sample of Facebook interests. Here, we report variations of up to the first 50 PCs.

Table A4.3 shows the results from alternative predictions of patience and risk-taking for the cross-country analysis. Columns 1-4 report results when using the first 20, 30, 40, and 50 PCs, respectively, when predicting the two traits in the international sample. Panel A performs the analyses for the sample of 48 countries that participated in the GPS. Panel B shows the same analyses for the extended sample of 80 countries. Results are qualitatively and quantitatively very similar to the respective results in Table 4.1, which implies that the relationship between the Facebook interests and the two cultural traits is very stable in the international sample.

Table A4.4 shows the equivalent results for the same variation in PCs in the migrant analysis. The results for patience are stable across the different numbers of PCs. By contrast, the significantly negative estimate on risk-taking also shows with 20 PCs, but not beyond. This is in line with the observation from the regional analysis that risk-taking seems to be harder to predict from the Facebook data.

A.2 Italy

The first additional analysis for Italian regions shows that the significant positive association of patience with student achievement also holds for reading. Our main analysis in Section 4.4.1 focuses on math achievement, which is generally considered the most comparable subject across countries. Conversely, student reading outcomes are inherently language-specific, which makes them less suitable for cross-country analysis. We exploit the within-country nature and the richness of the INVALSI data to replicate our analysis using reading outcomes. Results in Table A4.5 show that a one SD increase in patience is associated with a 0.99-1.22 SD increase in student reading achievement in the individual-level sample. At the regional level, a one SD increase in patience is associated with an increase of 0.71-0.91 SD in reading scores. The magnitude of the coefficients in reading is slightly smaller than in math but results clearly show in both subjects.

Results are also very robust across subsamples of waves and gender. The first two columns of Table A4.6 show that results do not depend on the year in which the assessment was conducted. This suggests that our estimates are not driven by the timing of the observation of the achievement data. Results also hold similarly for girls and boys, and the gender difference is not statistically significant (Columns 3-4).[92]

In line with a leading role of cultural traits as a deep determinant of student achievement, results are stronger for native students than for migrant students. Results in Table A4.7 show that a one SD increase in patience is associated with a 1.42-1.58 SD increase in achievement for native students, a 0.75-0.91 SD increase in achievement for students with a second-generation migrant background, and a 0.56-0.89 SD increase in achievement for students with a first-generation migrant background. This pattern would be expected if it is indeed patience as a cultural trait that drives the achievement results, as the culture of the residence region is presumably less important for migrant students who have been less exposed to the regional culture.[93]

An additional robustness check ensures that results are not driven by student achievement in Trentino-Alto-Adige. In the INVALSI test of this region, only students in the autonomous municipalities of Bolzano and Trento are tested (see Section 4.3.1).

[92] Reported results are based on Facebook-derived measures obtained with 4 PCs, but results are qualitatively the same with 7 and 10 PCs (not shown).

[93] Hanushek et al. (2022) find a similar pattern in their analysis of international student achievement.

This sampling in municipal areas only may bias our estimates, not least because Trentino-Alto-Adige is the Italian region with the highest estimated level of patience (see Section 4.2.3). Furthermore, we want to be sure that results are not driven by the Austrian history and the partially German-speaking population of the region. When omitting these municipalities from the analysis in Table A4.8, results are qualitatively the same and, if anything, slightly larger in magnitude.

We also perform an analysis of unobservable selection and coefficient stability proposed by Oster (2019). We compare our baseline models in Panel A of Table A4.9 to a restricted model without control variables. We follow the standard procedure and set $R_{max} = 1.3\tilde{R}$. The results in Table A4.9 imply that assuming an equal degree of selection between observables and unobservables, $\delta = 1$, the estimated bias-adjusted coefficient $\beta^*$for patience is between 1.487 and 1.705. In all cases, the bias-adjusted coefficient β^* is larger than our main estimates. The values δ for which $\beta = 0$ lie between -2.680 and -4.117. In all cases, these values are much larger than the standard cutoff $\delta = 1$. These results imply that the selection on unobservables would need to be more than 2.6 times larger than the selection on observables to push the coefficient of patience to 0.

Finally, we make use of the fact that Italy participated with a regionally representative sample in the international PISA test in 2012 to show that results hold equally well in this alternative achievement test. Intriguingly, the PISA results shown in Table A4.10 are very similar to the INVALSI results shown in Panel A of Table 4.2, indicating that a one SD increase in patience is associated with a 1.47-1.57 SD increase in the PISA math score.

A.3 United States

For the U.S. states, we first replicate the main results of the analysis in Section 4.4.2 using reading outcomes. The results reported in Table A4.11 closely mirror the findings for Italy: the magnitude of the coefficient of patience is slightly smaller compared the analysis of math achievement. A unit increase in patience is associated with an increase of 0.14-0.23 SD in reading achievement. Again, this analysis confirms that results do not depend on a particular subject.

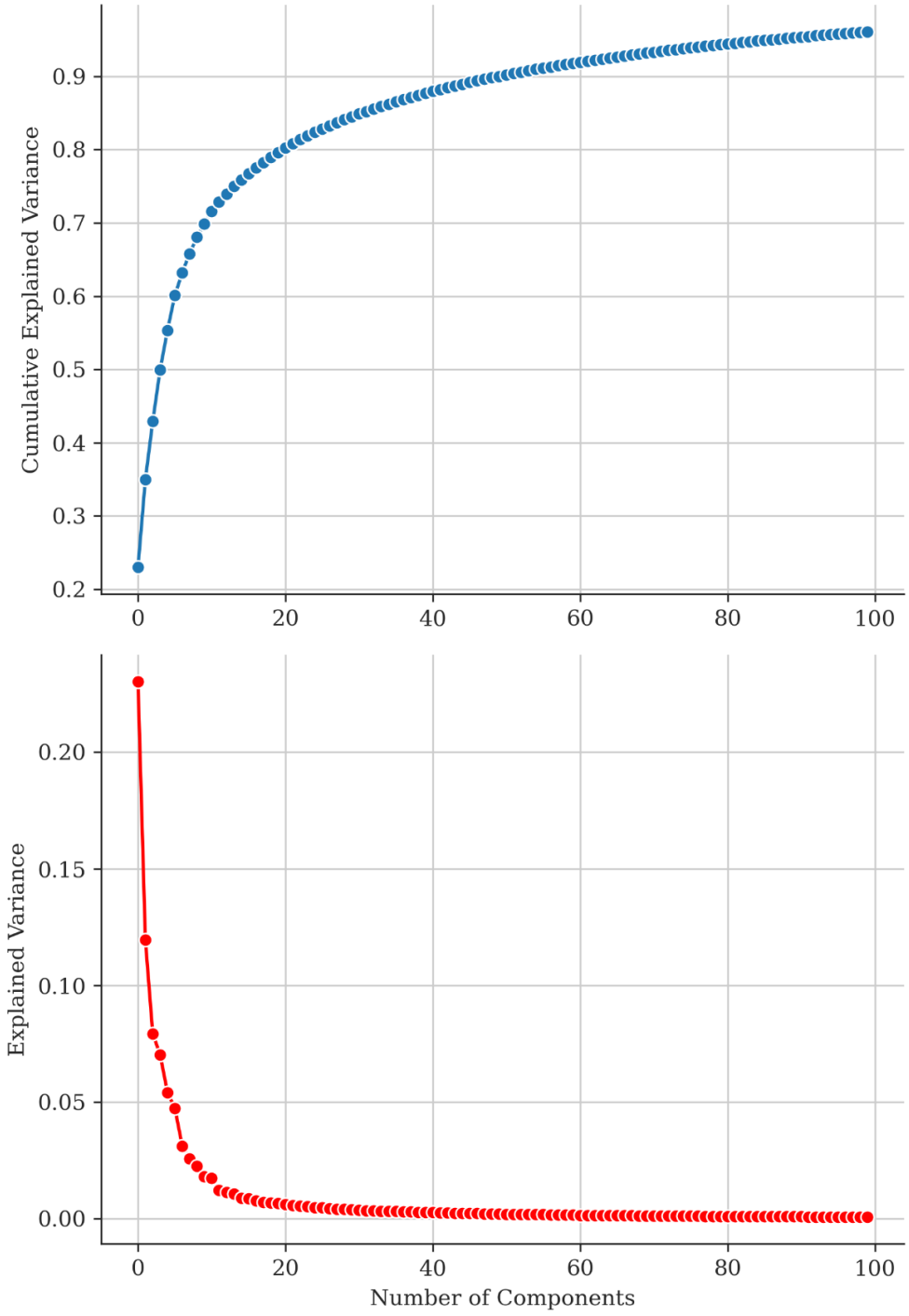
We also check that results do not depend on the specific year in which student achievement is observed. Table A4.12 reports results using each wave of NAEP data – 2015, 2017, and 2019 – separately. Results are qualitatively the same for all analyzed waves. The magnitude of the patience coefficient tends to be smaller in the most

recent wave, although not statistically significantly so. Overall, these results suggest that the findings do not depend on the specific year in which student test scores are observed.

Finally, the U.S. results are also similar across genders. Results in Table A4.13 show that patience is significantly positively associated with student achievement of both boys and girls. The coefficient estimates are somewhat larger for boys than for girls, but not significantly so, suggesting that results are qualitatively similar with respect to student gender.

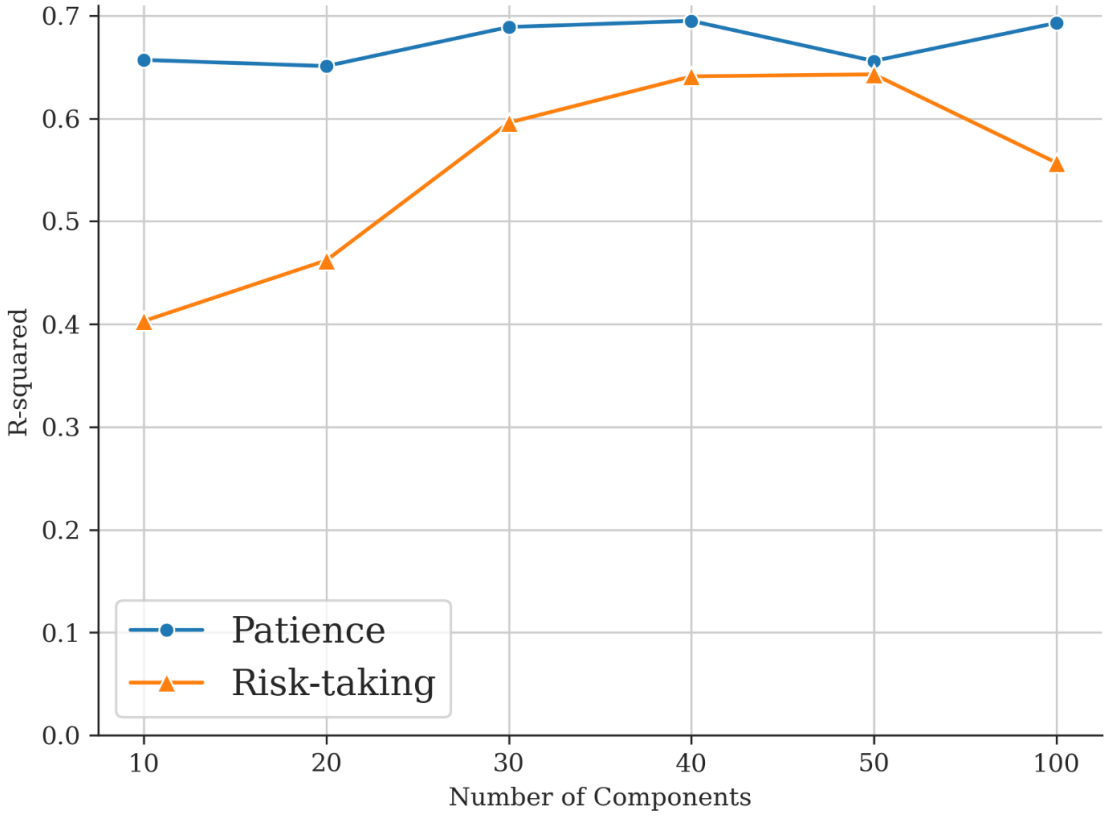
Appendix Figures and Tables

Figure A4.1: Variance in Facebook Interests Captured by PCs: International Sample



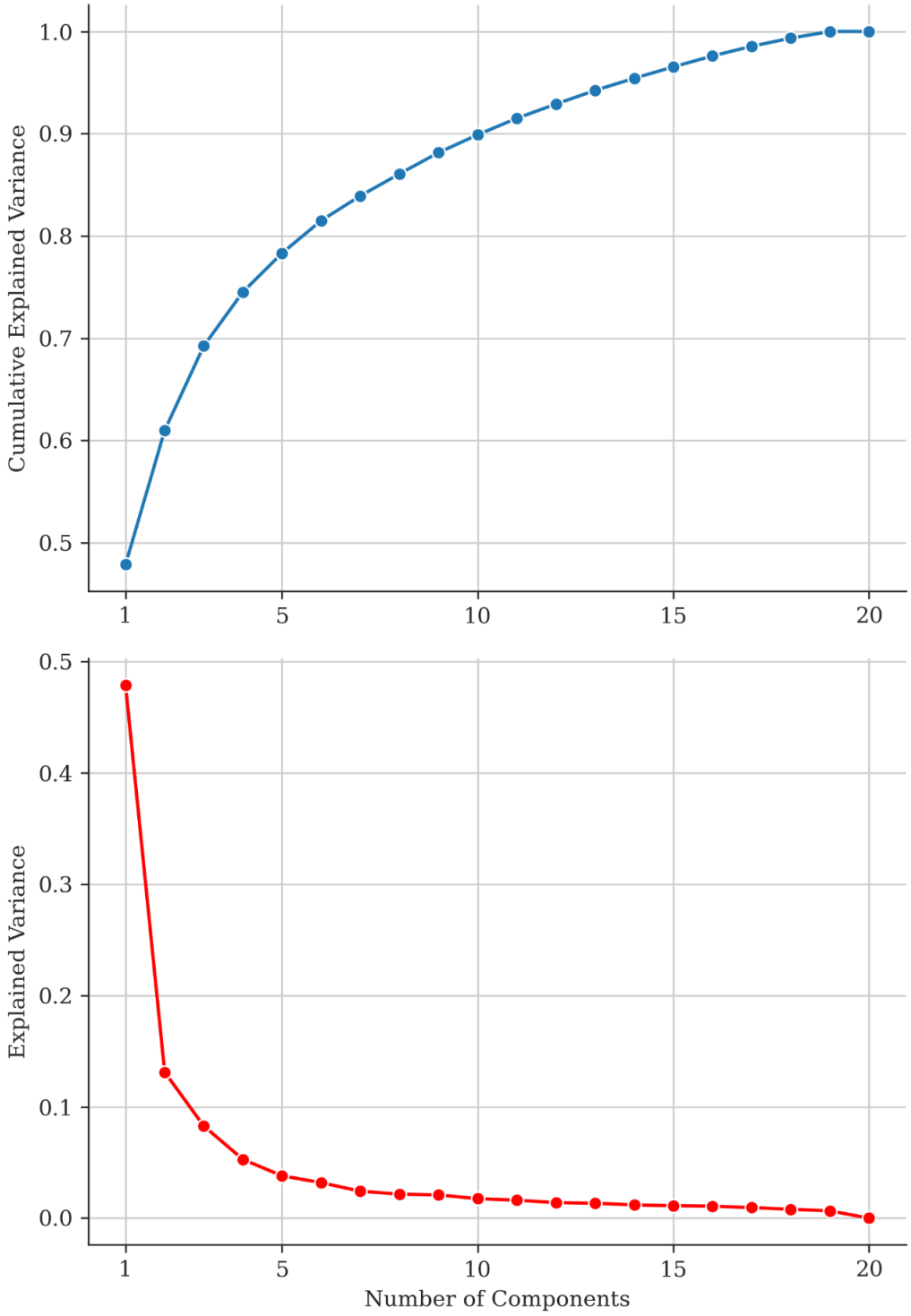
Notes: The top figure shows the cumulative variance in Facebook interests captured by the PCs of the Facebook interests in the international sample, the bottom figure shows the variance captured by each component.

Figure A4.2: Performance of GPS Prediction with Facebook Interests: International Sample



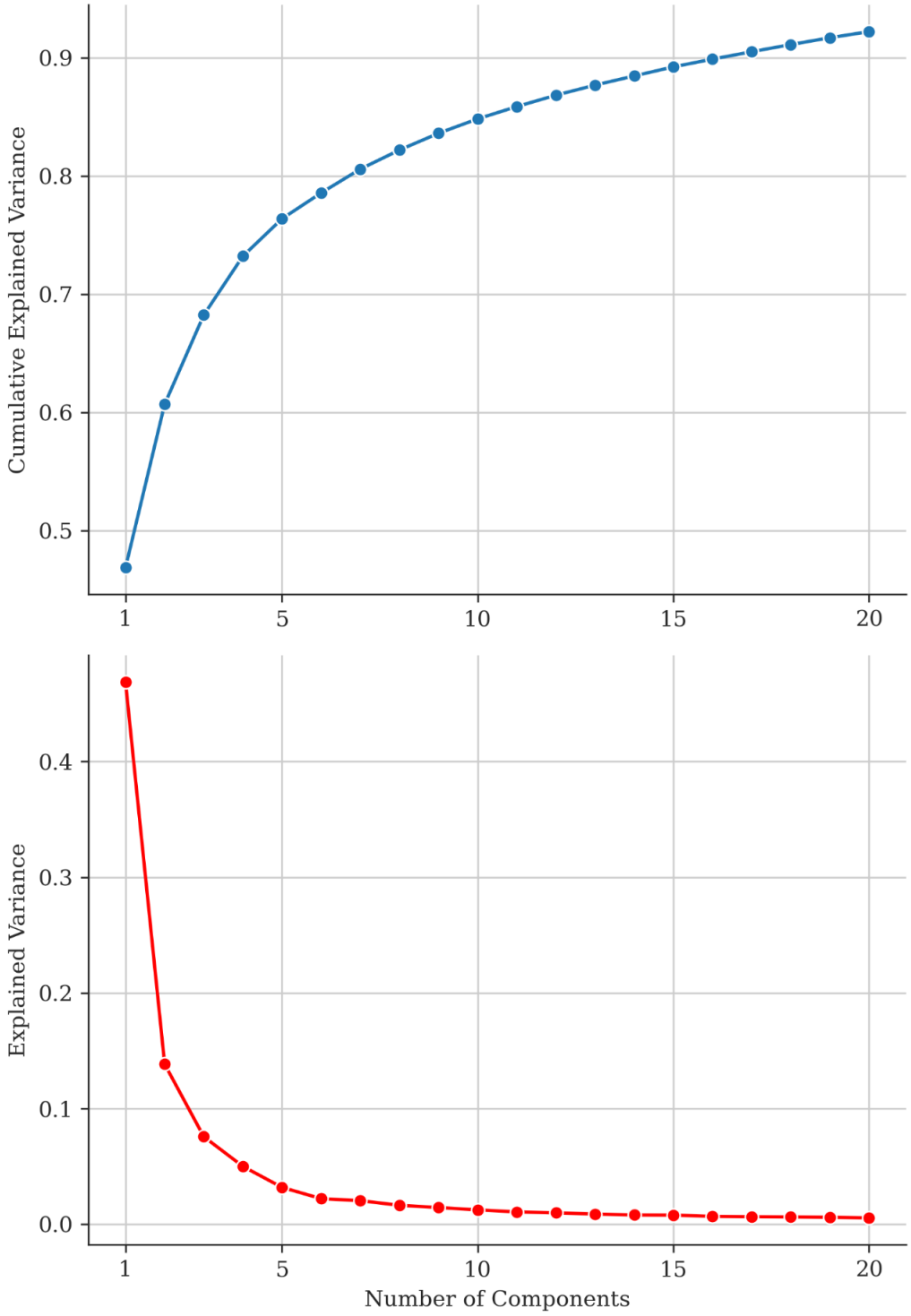
Notes: The figure shows the R^2 of regressions of the GPS measures of patience and risk-taking, respectively, on the PCs of Facebook interests (obtained with PC loadings of country-level Facebook interests) for different numbers of PCs used in the regression. 10-fold cross-validated LASSO model. Sample: all 74 countries for which GPS and Facebook data are available.

Figure A4.3: Variance in Facebook Interests Captured by PCs: Italian Regions



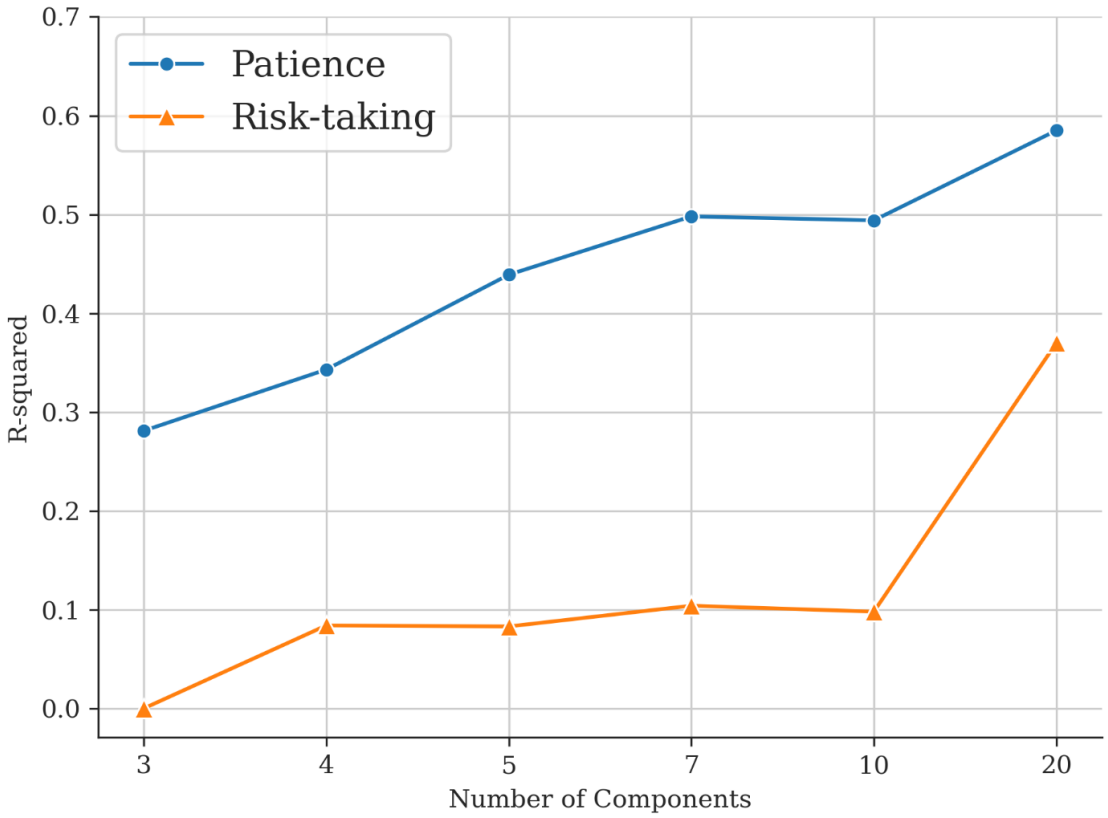
Notes: The top figure shows the cumulative variance in Facebook interests captured by the PCs of the Facebook interests in the Italian regions, the bottom figure shows the variance captured by each component.

Figure A4.4: Variance in Facebook Interests Captured by PCs: U.S. States



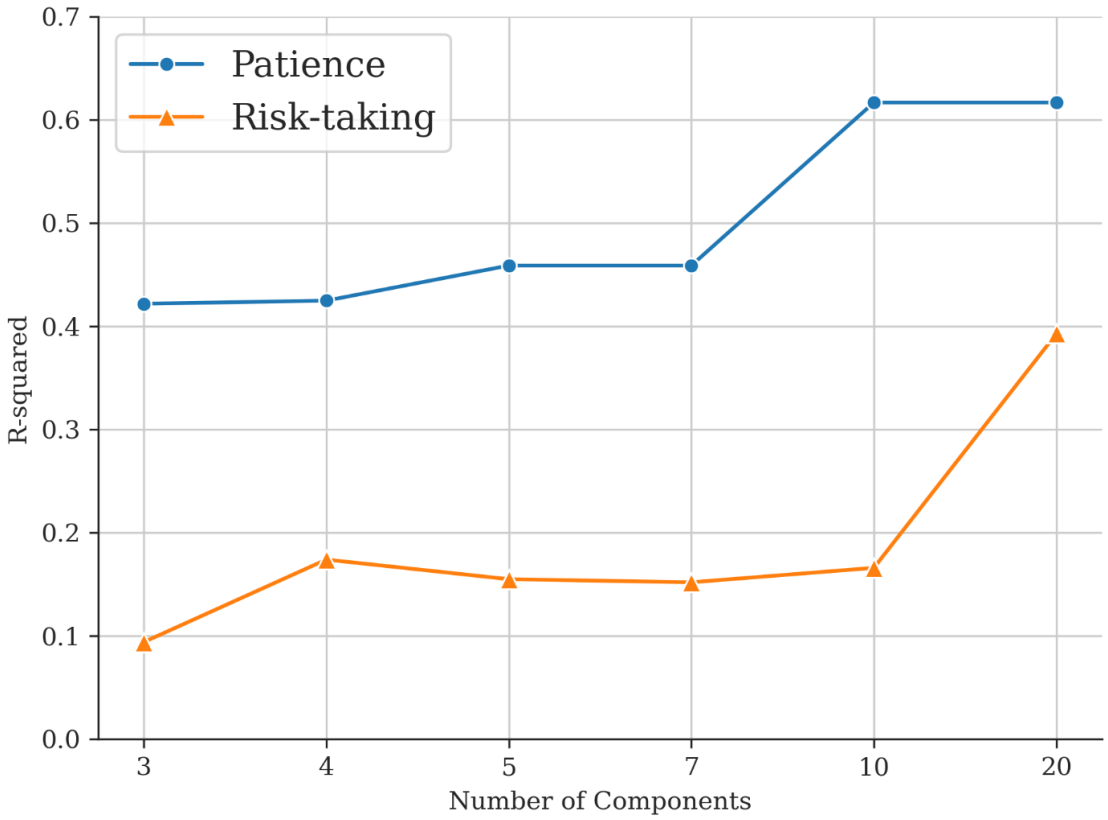
Notes: The top figure shows the cumulative variance in Facebook interests captured by the PCs of the Facebook interests in the U.S. states, the bottom figure shows the variance captured by each component.

Figure A4.5: Performance of GPS Prediction with Facebook Interests: PC Loadings from Italian Regions



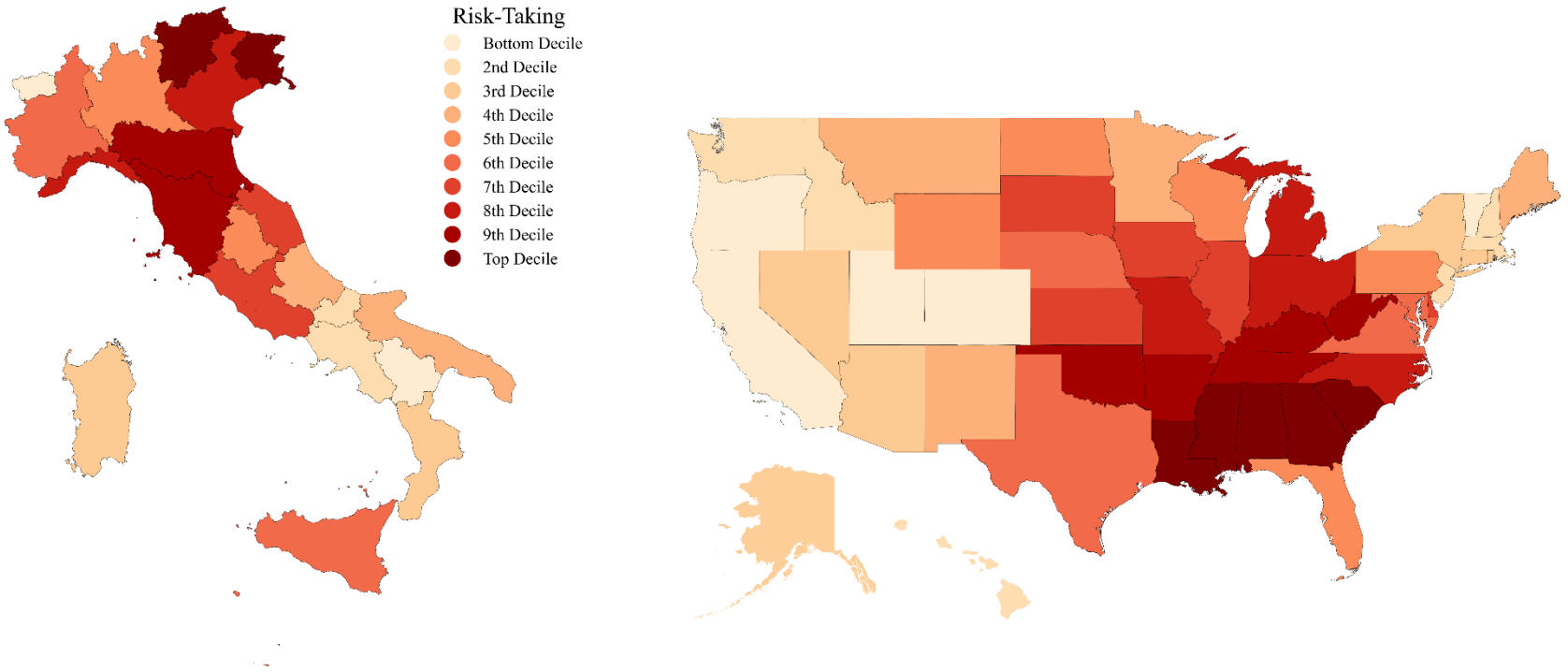
Notes: The figure shows the R^2 of regressions of the GPS measures of patience and risk-taking, respectively, on the PCs of Facebook interests (obtained with the PC loadings of Italian-region-level Facebook interests) for different numbers of PCs used in the regression. 10-fold cross-validated LASSO model. Sample: all 74 countries for which GPS and Facebook data are available.

Figure A4.6: Performance of GPS Prediction with Facebook Interests: PC Loadings from U.S. States



Notes: The figure shows the R^2 of regressions of the GPS measures of patience and risk-taking, respectively, on the PCs of Facebook interests (obtained with PC loadings of U.S. state-level Facebook interests) for different numbers of PCs used in the regression. 10-fold cross-validated LASSO model. Sample: all 74 countries for which GPS and Facebook data are available.

Figure A4.7: Measure of Risk-Taking Derived from Facebook Interests for Italian Regions and U.S. States



A. Italian Regions **B. U.S. States**

Notes: The figures show maps of the Facebook-derived measure of risk-taking obtained with 4 PCs for Italian regions (Panel A) and U.S. states (Panel B), respectively. Each color corresponds to a decile of the distribution of risk-taking within each country. Darker colors denote higher levels of risk-taking.

Table A4.1: Countries in the Cross-country Validation Exercise

	PISA countries			Training sample
	Only Facebook (1)	Only GPS (2)	Facebook and GPS (3)	Facebook and GPS (4)
Afghanistan				x
Albania	x			
Algeria			x	x
Argentina			x	x
Australia			x	x
Austria			x	x
Azerbaijan	x			
Bangladesh				x
Belarus	x			
Belgium	x			
Bolivia				x
Bosnia and Herzegovina			x	x
Botswana				x
Brazil			x	x
Brunei Darussalam	x			
Bulgaria	x			
Cambodia				x
Cameroon				x
Canada			x	x
Chile			x	x
China				x
Colombia			x	x
Costa Rica			x	x
Croatia			x	x
Czech Republic			x	x
Denmark	x			
Dominican Republic	x			
Egypt				x
Estonia			x	x
Finland			x	x
France			x	x
Georgia			x	x
Germany			x	x
Ghana				x
Greece			x	x
Guatemala				x
Haiti				x
Hong Kong	x			
Hungary			x	x
Iceland	x			
India				x
Indonesia			x	x
Iraq				x
Ireland	x			
Israel			x	x
Italy			x	x
Japan			x	x
Jordan			x	x

(continued on next page)

Table A4.1 (continued)

	PISA countries			Training sample
	Only Facebook (1)	Only GPS (2)	Facebook and GPS (3)	Facebook and GPS (4)
Kazakhstan			x	x
Kenya				x
Korea			x	x
Kyrgyzstan	x			
Latvia	x			
Lebanon	x			
Liechtenstein	x			
Lithuania			x	x
Luxembourg	x			
Macao	x			
Malawi				x
Malaysia	x			
Malta	x			
Mauritius	x			
Mexico			x	x
Moldova			x	x
Montenegro	x			
Morocco			x	x
Netherlands			x	x
New Zealand	x			
Nicaragua				x
Nigeria				x
North Macedonia	x			
Norway	x			
Pakistan				x
Panama	x			
Peru			x	x
Philippines			x	x
Poland			x	x
Portugal			x	x
Qatar	x			
Romania			x	x
Russia		x		
Rwanda				x
Saudi Arabia			x	x
Serbia			x	x
Singapore	x			
Slovakia	x			
Slovenia	x			
South Africa				x
Spain			x	x
Sri Lanka				x
Suriname				x
Sweden			x	x
Switzerland			x	x
Tanzania				x
Thailand			x	x
Trinidad and Tobago	x			

(continued on next page)

Table A4.1 (continued)

	PISA countries			Training sample
	Only Facebook	Only GPS	Facebook and GPS	Facebook and GPS
	(1)	(2)	(3)	(4)
Tunisia	x			
Turkey			x	x
Uganda				x
Ukraine			x	x
United Arab Emirates			x	x
United Kingdom			x	x
United States			x	x
Uruguay	x			
Venezuela				x
Vietnam			x	x
Zimbabwe				x
Total: 107 countries	32	1	48	74

Notes: Sample of countries: Col. 1-3: countries included in the cross-country validation exercise (Panel A of Table 4.1). Col. 4: countries included in training the machine learning model. Country names are as reported in PISA codebooks or Facebook/GPS data and do not represent any political views of the authors.

Table A4.2: Countries in the Migrant Analysis

	GPS/Facebook country of origin			PISA destination country	
	Only GPS	Only Facebook	Both	GPS analysis	Facebook analysis
	(1)	(2)	(3)	(4)	(5)
Afghanistan			x		
Albania		x			
Algeria					
Argentina			x	x	x
Armenia		x			
Australia			x	x	x
Austria			x	x	x
Azerbaijan		x			
Bangladesh			x		
Belarus		x		x	x
Belgium		x		x	x
Bolivia			x		
Bosnia and Herzegovina			x	x	x
Brazil			x		
Brunei Darussalam				x	x
Bulgaria		x			
Cape Verde		x			
Canada			x	x	x
Chile			x		
China			x		
Colombia			x		
Costa Rica				x	x
Croatia			x	x	x
Czech Republic			x	x	x
Denmark		x		x	x
Dominican Republic		x		x	x
Egypt			x		
Estonia			x		
Ethiopia		x			
Fiji		x			
Finland			x	x	x
France			x		
Georgia			x		x
Germany			x	x	x
Greece			x		x
Haiti			x		
Hong Kong				x	x
Hungary			x		
Iceland		x			
India			x		
Indonesia			x	x	x
Iran	x				
Iraq			x		
Ireland		x		x	x
Israel				x	x
Italy			x		
Japan					
Jordan			x	x	x

(continued on next page)

Table A4.2 (continued)

	GPS/Facebook country of origin			PISA destination country	
	Only GPS (1)	Only Facebook (2)	Both (3)	GPS analysis (4)	Facebook analysis (5)
Kazakhstan			x		
Kuwait		x			
Kyrgyzstan					x
Latvia				x	x
Lebanon		x			
Libya		x			
Liechtenstein		x		x	x
Lithuania			x		
Luxembourg				x	x
Macao		x		x	x
Malaysia		x			
Mauritius				x	x
Mexico				x	x
Moldova			x	x	x
Montenegro		x		x	x
Morocco			x	x	x
Netherlands			x	x	x
New Zealand		x		x	x
Nicaragua			x		
Nigeria			x		
North Macedonia				x	x
Norway		x		x	x
Pakistan			x		
Palestine		x			
Panama		x		x	x
Paraguay		x			
Peru					
Philippines			x	x	x
Poland			x		
Portugal			x	x	x
Qatar		x		x	x
Romania			x		
Russia	x				
Samoa		x			
Saudi Arabia			x	x	x
Serbia			x		x
Singapore		x			
Slovakia		x		x	x
Slovenia		x		x	x
Somalia		x			
South Africa			x		
South Korea			x	x	x
Spain			x		
Suriname			x		
Sweden			x		
Switzerland			x	x	x
Tajikistan		x			
Thailand			x		

(continued on next page)

Table A4.2 (continued)

	GPS/Facebook country of origin			PISA destination country	
	Only GPS	Only Facebook	Both	GPS analysis	Facebook analysis
	(1)	(2)	(3)	(4)	(5)
Tonga		x			
Turkey			x	x	x
Ukraine			x	x	x
United Arab Emirates			x		
United Kingdom			x	x	x
United States			x		
Uruguay		x		x	x
Uzbekistan		x			
Venezuela			x		
Vietnam			x		
Yemen		x			
Zambia		x			
Total: 108 countries	2	37	56	46	50

Notes: Sample of countries that serve as countries of origin (col. 1-3) or destination countries (col. 4-5) in the migrant analysis (Panel B of Table 4.1). Country names are as reported in PISA codebooks or Facebook/GPS data and do not represent any political views of the authors.

Table A4.3: Validation of Cross-Country Analysis: Different Numbers of Principal Components (PCs)

	20 PCs	30 PCs	40 PCs	50 PCs
	(1)	(2)	(3)	(4)
A. Original country sample (GPS countries)				
Patience	1.598***	1.588***	1.601***	1.610***
	(0.132)	*(0.140)*	*(0.139)*	*(0.140)*
Risk-taking	-1.598***	-0.883***	-0.898***	-1.004***
	(0.452)	*(0.316)*	*(0.308)*	*(0.276)*
Control variables	Yes	Yes	Yes	Yes
Observations	1,954,840	1,954,840	1,954,840	1,954,840
Residence countries	48	48	48	48
R^2	0.207	0.195	0.197	0.202
B. Extended country sample (all Facebook countries)				
Patience	1.641***	1.598***	1.607***	1.597***
	(0.121)	*(0.126)*	*(0.129)*	*(0.130)*
Risk-taking	-1.640***	-1.265***	-1.160***	-1.126***
	(0.336)	*(0.285)*	*(0.263)*	*(0.229)*
Control variables	Yes	Yes	Yes	Yes
Observations	2,660,408	2,660,408	2,660,408	2,660,408
Residence countries	80	80	80	80
R^2	0.205	0.203	0.200	0.199

Notes: Dependent variable: PISA math test score in all PISA waves 2000-2018. Least squares regressions weighted by students' sampling probability. Control variables: student gender, age, and migration status; imputation dummies; and wave fixed effects. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2000-2018; own elaboration of Facebook data.

Table A4.4: Validation of Migrant Analysis: Different Numbers of Principal Components (PCs)

	20 PCs (1)	30 PCs (2)	40 PCs (3)	50 PCs (4)
A. Original sample (GPS countries of origin)				
Patience	0.783***	0.876***	0.885***	0.875***
	(0.193)	*(0.197)*	*(0.192)*	*(0.216)*
Risk-taking	-0.676**	0.008	0.087	0.156
	(0.306)	*(0.367)*	*(0.322)*	*(0.371)*
Control variables	Yes	Yes	Yes	Yes
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes
Observations	78,403	78,403	78,403	78,403
Countries of origin	56	56	56	56
Residence countries	46	46	46	46
R^2	0.271	0.271	0.272	0.270
B. Extended sample (all Facebook countries of origin)				
Patience	0.838***	1.027***	1.033***	0.995***
	(0.211)	*(0.198)*	*(0.191)*	*(0.211)*
Risk-taking	-1.155***	-0.067	0.064	0.154
	(0.422)	*(0.357)*	*(0.297)*	*(0.341)*
Control variables	Yes	Yes	Yes	Yes
Residence-country by wave fixed effects	Yes	Yes	Yes	Yes
Observations	90,983	90,983	90,983	90,983
Countries of origin	93	93	93	93
Residence countries	50	50	50	50
R^2	0.295	0.294	0.294	0.291

Notes: Dependent variable: PISA math test score, waves 2003-2018. Least squares regressions, including 180 fixed effects for each residence-country by wave cell. Sample: students with both parents not born in the country where the student attends school. Control variables: student gender, age, dummy for OECD country of origin, imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA international student achievement test, 2003-2018; own elaboration of Facebook data.

Table A4.5: Patience and Reading Achievement: Analysis of Italian Regions

	4 PCs (1)	7 PCs (2)	10 PCs (3)
A. Individual level			
Patience	1.218***	0.986***	1.050***
	(0.201)	*(0.123)*	*(0.128)*
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	59,441	59,441	59,441
Regions	20	20	20
R^2	0.105	0.110	0.110
B. Regional level			
Patience	0.905***	0.716***	0.762***
	(0.177)	*(0.094)*	*(0.098)*
Wave fixed effects	Yes	Yes	Yes
Observations	42	42	42
Regions	20	20	20
R^2	0.496	0.617	0.625

Notes: Dependent variable: INVALSI 8th-grade reading test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI reading achievement test, 2017-2019; own elaboration of Facebook data.

Table A4.6: Patience and Math Achievement: Analysis of Italian Regions by Subgroups

	2018	2019	Males	Females
	(1)	(2)	(3)	(4)
A. Individual level				
Patience (4 PCs)	1.588***	1.422***	1.579***	1.427***
	(0.191)	*(0.217)*	*(0.211)*	*(0.198)*
Control variables	Yes	Yes	Yes	Yes
Wave fixed effects	No	No	Yes	Yes
Observations	29,359	29,675	30,530	28,504
Regions	20	20	20	20
R^2	0.095	0.089	0.097	0.082
B. Regional level				
Patience (4 PCs)	1.331***	1.161***	1.305***	1.185***
	(0.221)	*(0.241)*	*(0.226)*	*(0.227)*
Wave fixed effects	No	No	Yes	Yes
Observations	21	21	42	42
Regions	20	20	20	20
R^2	0.693	0.668	0.682	0.657

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination. Patience measure computed with 4 principal components (PCs). Regressions include the risk-taking measure computed with 4 PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI reading achievement test, 2017-2019; own elaboration of Facebook data.

Table A4.7: Patience and Math Achievement: Analysis of Italian Regions by Migrant Status

	4 PCs (1)	7 PCs (2)	10 PCs (3)
A. Native students			
Patience	1.581***	1.423***	1.514***
	(0.188)	*(0.115)*	*(0.118)*
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	51,691	51,691	51,691
Regions	20	20	20
R^2	0.084	0.091	0.091
B. Second-generation migrant students			
Patience	0.909***	0.748***	0.820***
	(0.237)	*(0.215)*	*(0.220)*
Wave fixed effects	Yes	Yes	Yes
Observations	3,572	3,572	3,572
Regions	20	20	20
R^2	0.033	0.035	0.035
C. First-generation migrant students			
Patience	0.565**	0.842***	0.893***
	(0.235)	*(0.112)*	*(0.124)*
Wave fixed effects	Yes	Yes	Yes
Observations	1,719	1,719	1,719
Regions	20	20	20
R^2	0.079	0.083	0.083

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: student. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables: student gender and age; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; own elaboration of Facebook data.

Table A4.8: Patience and Math Achievement: Analysis of Italian Regions Excluding Trentino-Alto-Adige

	4 PCs (1)	7 PCs (2)	10 PCs (3)
A. Individual level			
Patience	1.717***	1.412***	1.520***
	(0.158)	*(0.122)*	*(0.124)*
Control variables	Yes	Yes	Yes
Wave fixed effects	Yes	Yes	Yes
Observations	55,437	55,437	55,437
Regions	19	19	19
R^2	0.095	0.098	0.098
B. Regional level			
Patience	1.462***	1.220***	1.314***
	(0.171)	*(0.094)*	*(0.097)*
Wave fixed effects	Yes	Yes	Yes
Observations	38	38	38
Regions	19	19	19
R^2	0.783	0.835	0.846

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: Panel A: student; Panel B: region-wave combination. Students in the autonomous municipalities of Trento and Bolzano are dropped from the estimation sample. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Controls variables (Panel A): student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; own elaboration of Facebook data.

Table A4.9: Analysis of Unobservable Selection and Coefficient Stability following Oster (2019): Analysis of Italian Regions

	4 PCs		7 PCs		10 PCs	
	Restr.ed	Ext.ed	Restr.ed	Ext.ed	Restr.ed	Ext.ed
	(1)	(2)	(3)	(4)	(5)	(6)
Patience	1.252***	1.505***	1.136***	1.350***	1.208***	1.437***
	(0.210)	*(0.197)*	*(0.122)*	*(0.114)*	*(0.129)*	*(0.117)*
Control variables	No	Yes	No	Yes	No	Yes
Wave fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	59,034	59,034	59,034	59,034	59,034	59,034
Regions	20	20	20	20	20	20
R^2	0.043	0.092	0.049	0.099	0.050	0.099
Oster (2019) diagnostics						
Bound β^* for $\delta = 1$	1.705		1.487		1.581	
δ to match $\beta = 0$	-4.117		-2.687		-2.680	

Notes: Dependent variable: INVALSI 8th-grade math test score in waves 2018 and 2019. Least squares regressions with wave fixed effects. Unit of observation: student. Students in the autonomous municipalities of Trento and Bolzano are dropped from the estimation sample. Patience measure computed with number of principal components (PCs) indicated in column header. Regressions include the risk-taking measure computed with the equivalent number of PCs. Odd columns: restricted model with wave fixed effects. Even columns: baseline models with wave fixed effects, student gender, age, and migration status; imputation dummies. Oster statistics computed using $R_{max} = 1.3\tilde{R}$, where $\tilde{R}$ denotes the R^2 reported in even columns. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: INVALSI mathematics achievement test, 2017-2019; own elaboration of Facebook data.

Table A4.10: Patience and Math Achievement: Analysis of Italian Regions using PISA 2012 Data

	4 PCs (1)	7 PCs (2)	10 PCs (3)
Patience	1.484***	1.473***	1.570***
	(0.264)	*(0.132)*	*(0.138)*
Control variables	Yes	Yes	Yes
Observations	31,073	31,073	31,073
Regions	20	20	20
R^2	0.106	0.113	0.113

Notes: Dependent variable: PISA 2012 math test score. Least squares regressions. Unit of observation: student. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Control variables: student gender, age, and migration status; imputation dummies. Robust standard errors adjusted for clustering at the regional level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: PISA student achievement test, 2012; own elaboration of Facebook data.

Table A4.11: Patience and Reading Achievement: Analysis of U.S. States

	4 PCs	7 PCs	10 PCs
	(1)	(2)	(3)
Patience	0.228***	0.141*	0.227**
	(0.074)	*(0.077)*	*(0.103)*
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
States	51	51	51
R^2	0.385	0.375	0.396

Notes: Dependent variable: NAEP 8th-grade reading test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Robust standard errors adjusted for clustering at the state level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

Table A4.12: Patience and Math Achievement: Analysis of U.S. States by Wave

	4 PCs (1)	7 PCs (2)	10 PCs (3)
A. 2015			
Patience	0.335***	0.194**	0.346***
	(0.081)	*(0.082)*	*(0.119)*
States	51	51	51
R^2	0.426	0.410	0.430
B. 2017			
Patience	0.309***	0.179**	0.290**
	(0.084)	*(0.085)*	*(0.125)*
States	51	51	51
R^2	0.373	0.360	0.372
C. 2019			
Patience	0.235***	0.142*	0.228*
	(0.077)	*(0.077)*	*(0.114)*
States	51	51	51
R^2	0.277	0.267	0.278

Notes: Dependent variable: NAEP 8th-grade math test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Robust standard errors adjusted for clustering at the state level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

Table A4.13: Patience and Math Achievement: Analysis of U.S. States by Gender

	4 PCs (1)	7 PCs (2)	10 PCs (3)
A. Males			
Patience	0.322***	0.194*	0.305**
	(0.101)	*(0.108)*	*(0.147)*
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
States	51	51	51
R^2	0.388	0.377	0.385
B. Females			
Patience	0.263***	0.147*	0.258**
	(0.079)	*(0.086)*	*(0.119)*
Wave fixed effects	Yes	Yes	Yes
Observations	153	153	153
States	51	51	51
R^2	0.319	0.304	0.321

Notes: Dependent variable: NAEP 8th-grade math test score in all NAEP waves 2015-2019. Least squares regressions with wave fixed effects. Unit of observation: state-wave combination. Col. 1-3 use the patience measure computed with 4, 7, and 10 principal components (PCs), respectively. Regressions include the risk-taking measure computed with the equivalent number of PCs. Robust standard errors adjusted for clustering at the state level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent. Data sources: NAEP mathematics achievement test, 2015-2019; own elaboration of Facebook data.

5 Topic Salience and Political Polarization: Evidence from the German "PISA Shock"*

5.1 Introduction

Little is known about the relationship between the salience of a topic and the polarization in related political debates. Understanding this relationship is crucial since the salience of a topic, namely the amount of attention that it receives, can be manipulated. Traditional and digital media, for example, are prone to presenting reported events in a sensationalized way (Ryu 1982; Soroka et al. 2018; Bleich and van der Veen 2021; Kayser and Peress 2021; Berger 2022). Social media can exacerbate this phenomenon through the "echo chambers" they tend to create (Sunstein 2018; Settle 2018), thus contributing to an increase in the perceived salience of various issues. At the same time, there is mounting evidence of a historically high ideological divide observed in the United States (Bonica 2013; McCarty, Poole, and Rosenthal 2016; Gentzkow, Shapiro, and Taddy 2019; Iyengar et al. 2019) as well as in other countries (Boxell, Gentzkow, and Shapiro 2022). Particularly in parliamentary debates, the phenomenon of polarization has received considerable attention in recent years (Peterson and Spirling 2018; Gentzkow, Shapiro, and Taddy 2019; Goet 2019; Salla 2020; Fiva, Nedregård, and Øien 2022; Lewandowsky et al. 2022). This literature has mostly provided descriptive evidence on its evolution in different countries, but it has been surprisingly silent on *why* it occurs. The aim of this paper is to provide causal evidence on the salience of a topic as a potential determinant of polarization in related parliamentary debates.

Theoretically, it is an open question in which direction topic salience might affect polarization of parliamentary debates. If the salience of a topic increases, parties might pursue a median voter strategy to appeal to more centrist voters, thus resulting in less polarized debates. The theoretical foundation for this argument follows Downs' (1957) seminal work on the *median voter theorem*. Conversely, parties might exploit the increased salience of a topic to amplify their ideological distinctiveness, which would lead to an increase in polarization. Such behavior would be consistent with the *cleavage theory* framework, which dates back to Lipset and Rokkan (1967). Empirically, it is hard to establish whether topic salience affects polarization as, for

* This chapter is based on the paper "Topic Salience and Political Polarization: Evidence from the German 'PISA Shock'", *mimeo*.

example, politicians are known to focus on divisive issues (Ash, Morelli, and van Weelden 2017), which would lead to reverse causation.

To test whether topic salience affects the polarization of parliamentary debates, I leverage a natural experiment that led to an increase in the salience of a specific topic: education. I exploit the release of the results of the first *Programme for International Assessment* (PISA) study in December 2001 in the context of German state parliaments. Due to the unexpectedly low performance of German students and the media attention that this event received, this event was soon renamed the "PISA shock". I focus on the parliamentary debates of all German state parliaments for the period 2000-2008, which I have collected and digitized for this project. These debates constitute a novel data source, and the German context provides an ideal setting for my analysis. Germany is a federal country, where each of its sixteen states has its own parliament with exclusive legislative authority on a set of topics, including education. Hence, state-level parliamentary debates about education are policy-relevant and abundant in this context.

Empirically, I combine machine-learning algorithms and text analysis techniques to classify the topic of each speech in the parliamentary debates and compute topic-specific polarization measures. I use a supervised machine-learning model to classify speeches about the main topic of interest: education. I then classify the topics of all the other speeches with an unsupervised machine-learning algorithm, the correlated topic modelling (CTM) (Blei and Lafferty 2007). Using a measure of text similarity, the cosine similarity, I compute topic-specific measures of polarization, which is defined as the extent to which opinions on an issue are opposed across parties. Assuming that expressing different opinions requires people to use different words, more polarized speeches will be less similar. My main measure of polarization is therefore the dissimilarity between speeches from a benchmark party and speeches from other parties on the same topic.

Identifying the impact of salience on the polarization of education debate is challenging because polarization evolves over time. I therefore conduct a difference-in-differences analysis, where the debates on topics other than education act as the counterfactual group. This approach enables me to control for fluctuations in the general level of polarization in parliamentary debates due to time trends or other unrelated factors, such as upcoming elections or the idiosyncratic compositions of the parliaments. I find that topic salience induced by the PISA shock had a substantial impact on parliamentary debates. First, I find a 22% increase in the share of speeches

about education following the PISA shock. Second, I find a sizable increase in polarization of parliamentary debates about education equivalent to 8.8% of a standard deviation (SD). The impact corresponds to about 18% of the average polarization between the main center-right (CDU/CSU) and center-left (SPD) parties in the German political landscape. Using an event-study specification, I show that the shock also had a long-lasting impact. It took roughly six years for polarization in education debates to go back to its pre-shock level.

The interaction between members of parliament's (MPs) party affiliation and the treatment status reveals that the increase in polarization is driven by a cleavage between the main center-right (CDU/CSU) and center-left (SPD) parties. Overall, this result aligns well with a *cleavage-theory* framework, where the main parties drift away from each other in their rhetoric over a subject matter.

While the salience of education undoubtedly increased because of the PISA shock in Germany, it is also possible that the increase in polarization was driven by the information revealed by the release of the PISA results. I address this issue by exploiting an additional feature of this setting: the release of state-specific PISA results in June 2002. This event showed large heterogeneities in performance across German states, with the best performing states in Germany placing themselves among the top performing countries. Nonetheless, I do not find significant heterogeneities in the impact of the shock on polarization with respect to the performance of each state. Further, state-specific results were not released for two states, Berlin and Hamburg, and I also do not find any heterogeneities for these states. These findings seemingly suggest that the salience of the topic, rather than the actual performance of the students, affected the polarization of parliamentary debates. Further, I find that the PISA shock also had a positive impact on the number of proposed bills about education, and the impact is driven by rejected bills.

I also provide suggestive evidence on the issues that likely caused an increase in the polarization about education debates. I develop a polarization score to capture terms that are disproportionally used by MPs of one party. Terms that refer to prominent issues at the time of PISA shock, such as developing a monitoring system of student achievement, "all-day schools", and the tracking system, feature among the most polarized terms. This suggests that debates about such issues contributed to the increase in polarization in education.

This study contributes to two strands of the literature. First, I contribute to the growing literature investigating political polarization. Most studies in this field have

focused on the determinants of polarization among voters. This strand of research has shown a relationship between the rise in political polarization and rising import competition (Autor et al. 2020), intensified media partisanship (DellaVigna and Kaplan 2007; Levendusky 2013; Prior 2013), and financial crises (Mian, Sufi, and Trebbi 2014; Funke, Schularick, and Trebesch 2016). A polarized electorate can lead to more polarization in parliamentary debates, but this link is far from being established in the literature. In fact, causal evidence on the determinants of polarization in the context of parliamentary debates is largely absent. This is surprising given the outburst of studies documenting polarization in parliamentary debates observed in the last years, with evidence from the US (Jensen et al. 2012; Lauderdale and Herzog 2016; Gentzkow, Shapiro, and Taddy 2019), the UK (Peterson and Spirling 2018; Goet 2019), Germany (Lewandowsky et al. 2022), Norway (Fiva, Nedregård, and Øien 2022), and Finland (Salla 2020).[1] I therefore contribute to this literature by providing causal evidence of the effect of topic salience on polarization in parliamentary debates.

Second, I contribute to the political economy of education literature. I show that international standardized assessments, such as PISA, can influence the political discourse about education. Other studies have highlighted the role of interest groups, unions (McDonnell and Weatherford 2013; Galey-Horn et al. 2020), and teacher strikes (Lyon and Kraft 2021) in shaping education policymaking. Public opinion and interest groups are often considered to have a greater role in shaping education policy than insights drawn from empirical data (West and Woessmann 2021). I challenge this notion by providing evidence on the far-reaching consequences of the introduction of an international standardized assessment on the policy-making debate about education. A likely reason behind the impact of the PISA shock is that PISA introduced accountability for policymakers in education. Accountability has been often cited as a key factor to improve the quality of education systems (Woessmann et al. 2009; Figlio and Loeb 2011; Global Education Monitoring Report Team 2017; Bergbauer, Hanushek, and Woessmann 2021). In fact, the lack of comparable student assessments in many countries prevented policymakers from being held accountable for students' performance. This dramatically changed after PISA, as the strong reaction of German policy makers clearly illustrates. The influence of PISA, and the PISA shock, for policymaking in education in various countries has been widely

[1] Using US congressional vote choices rather than parliamentary debates, Canen, Kendall, and Trebbi (2020a, 2020b) highlight the role of party discipline as a driver of political polarization.

acknowledged in the literature.[2] To the best of my knowledge, no study has attempted to establish a causal relationship between PISA results and the political debate about education. I therefore fill the gap in this literature by providing causal evidence on how the international standardized assessment can shape education policymaking.

The remainder of this paper is structured as follows. In Section 5.2, I provide details about the PISA shock, the concept of topic salience, and the German political system. In Section 5.3, I present the data and methods used to compute the polarization measures as well as descriptive statistics. In Section 5.4, I present the empirical strategy. In Section 5.5, I report the main results and robustness checks. I provide evidence on the polarizing issues in Section 5.6. Section 5.7 concludes.

5.2 Institutional Background

5.2.1 The PISA Shock

The publication of the results of the first PISA study on the 4th of December 2001 was a watershed in the discourse on education in Germany. The poor and largely unequal performance of German students in PISA sparked heated public debates, with newspaper headlines such as "Catastrophic Results for German Students" (FAZ 2001), "A Disaster in Almost Every Respect" (TAZ 2001), or "Are German students stupid?" (Der Spiegel 2001) populating German newspapers for months. In the two months after the publication of the PISA results, the OECD calculated that daily and weekly newspapers published 774 pages of printed article about this event in Germany, compared to 8 in Finland, the "PISA champion country", 32 in France, whose placement was well above Germany in the PISA ranking, and 16 in Italy, whose performance was akin to Germany (Hopmann, Brinek, and Retzl 2007). The "tsunami-like" impact of this event in Germany (Gruber 2006) was so great that it was soon dubbed the PISA shock and its consequences shaped the public and political debate about education in the following years. In June 2002, roughly six months after the PISA shock, results for German federal states were published and revealed large differences in achievement between the states.[3] Although there were already some indications of

[2] A vast literature has discussed the implication of PISA for education policy in various countries (Rinne, Kallo, and Hokka 2004; Grek 2009; Bieber and Martens 2011; Breakspear 2012; Martens and Niemann 2013), among others). Several studies have also investigated the consequences of the PISA shock in Germany (Tillmann 2004; Ertl 2006; Waldow 2009; Neumann, Fischer, and Kauertz 2010; Davoli and Entorf 2018, among others).

[3] Results were published for all states but Berlin and Hamburg, which did not meet the required criteria for overall reporting (Artelt et al. 2002) State-specific results are reported in Table A5.1.

such heterogeneities (Ebenrett, Hansen, and Puzicha 2003), this event further fueled the already heated debate about education.

Several reasons lie behind the stir caused by the publication of the first PISA results. First, PISA contradicted the public's perception of the German education system, an assessment that was characterized by self-confidence and belief in its efficiency, which reflected the strong country's economy (Sloane and Dilger 2005; Davoli and Entorf 2018). Second, it represented a threat to a major exporting economy that relies on human capital and skills for its competitive advantage. Third, PISA, and the *International Mathematics and Science Study* (TIMSS) before it, ended a long phase of German abstention from international large-scale assessments (Waldow 2009). In fact, Germany's participation and low performance in the first TIMSS study in 1995 was the first wake-up call for the German education system, but this event, unlike PISA, was largely ignored by the German media (OECD 2011). Germany's decade-long abstention from international assessments was in line with educators' mainstream paradigm that "what is important about education cannot be measured" (Bos and Postlethwaite 2002). PISA abruptly ended this phase, and Germany committed itself to participating in international assessments for years to come.

The PISA shock provided a formidable impetus for reforms in the German education systems. While an exhaustive exposition of such reforms is outside the scope of this paper,[4] they mostly revolved around three areas: developing a monitoring system with common educational standards and central examination, expanding "all-day school" offers, and reforming the tracking system.

5.2.2 Topic Salience

In this section, I clarify the concept of salience, which plays a crucial role in my analysis. I adhere to the concept of salience defined in a recent review of the literature that studies the role of salience in economic choice by Bordalo, Gennaioli, and Shleifer (2022). The authors describe salience as "the property of a stimulus that draws attention bottom up" (p. 524). Psychologists differentiate between top-down and bottom-up attention as the two methods through which human minds select what to focus on. Top-down attention is voluntary and is the result of an active cognitive process, whereas bottom-up attention is involuntary and occurs automatically.

[4] Interested readers may find detailed accounts of these in Ertl (2006), Gruber (2006), Waldow (2009), OECD (2011), and Davoli and Entorf (2018), among others.

Bordalo, Gennaioli, and Shleifer (2022) identify three factors that make a stimulus salient: contrast with surroundings (contrasting), surprise, and prominence.

It is easy to reconcile this definition of salience with the PISA shock. First, the PISA shock can be identified as a stimulus that drew public attention toward education bottom up, as it came as a reaction to the information made available by the PISA study. Second, the three factors that make a stimulus salient accurately depict the PISA shock: contrast with surroundings, surprise and prominence. An important feature that emerged from the first PISA results was that Germany was a country below the OECD average in terms of student test scores. This element of comparison with other countries—*contrast with surroundings*—contributed to the prominence that the publication of the first PISA results received. As argued in the previous section, PISA revealed a picture of the German education system that was largely unexpected and, therefore, *surprising*. Further, the PISA shock was very *prominent* due to its wide coverage on the media.

5.2.3 The German Political System

Germany is a federal country and comprises 16 states (*Länder)*.[5] Each state *(Land)* has its own constitution, elects its own parliament and creates its own government. Matters of national importance, such as foreign affairs, defense, or citizenship, are competence of the federal parliament *(Bundestag)* and government, while each state parliament *(Landtag)* has full autonomy on various subject matters, such as education, culture, police, or the press.[6] Elections in federal states occur at different times and with different electoral laws. A typical legislative period lasts five years.[7] Parliamentary debates in each state parliaments occur regularly, and, on average, 1.9 parliamentary sessions take place each month in each state.

The main political forces in the German political systems in the period analyzed in this paper, 2000-2008, consist of a left-leaning social democratic bloc, represented by the Social Democrats (SPD) and the Green Party (GRÜNE), and a right-leaning

[5] An exhaustive description of the German political system is outside the scope of this paper. In this section, I highlight only the features that are most relevant for the scope of this paper.

[6] A further category, which includes subjects such as environment, nature protection or land use, are jointly regulated by the federal and state parliaments. Interested readers may find the complete list of competences in https://www.bpb.de/medien/189018/Foederalismus.pdf.

[7] Except for Bremen, where legislatures last four years.

conservative bloc, represented by the Christian-Democratic Union (CDU) with its sister Bavarian denomination (CSU), and the Liberal Party (FDP).

5.3 Measuring Polarization in Parliamentary Debates: Data, Methods, and Descriptive Statistics

In this section, I describe the data sources, the methods used to compute the polarization measure, and report descriptive statistics of the main data sources.

5.3.1 Parliamentary Debates of the German States

The main source of data for this paper consists of parliamentary debates of the 16 German states for the period January 2000 – August 2008. Data limitations, discussed in more detail in Appendix B, prevent me from using data before the year 2000. The financial crisis that began in September 2008 serves as the cutoff point for my analysis, as it may have influenced the salience of numerous topics. Parliamentary debates constitute the preferred data source to measure the polarization for a variety of reasons. First, they convey timely and abundant information about MPs' opinions as opposed to voting patterns of member of parliaments, an alternative approach that has often been used to measure polarization in the US.[8] Second, parliamentary debates are a crucial way through which politicians obtain visibility in the media (Maltzman and Sigelman 1996; Tresch 2009; Salmond 2014) and express their views (Proksch and Slapin 2012), thus making them relevant for the policymaking process. As some scholars have argued, MPs use parliamentary speeches mainly as an act of position-taking rather than to persuade opponents or win political arguments (Proksch and Slapin 2015). Parliamentary debates are therefore particularly suited to study the extent to which MPs' policy positions evolve over time and across parties.

The federal structure of Germany also provides the ideal setting for this study. First, with respect to other studies using national parliamentary debates (Peterson and Spirling 2018; Goet 2019; Salla 2020; Fiva, Nedregård, and Øien 2022), this setting yields a much higher density of parliamentary debates, which is crucial to overcome

[8] Ideological positions measured with roll call-based approaches tend not to be informative in parliamentary systems such as Germany (Spirling and McLean 2007; Peterson and Spirling 2018). Further drawbacks of roll-call analyses include the selection of votes subject to roll call and their ability to capture only high levels of inter-party disagreement (Proksch and Slapin 2015). As noted in Slagter and Loewenberg (2007), roll-call votes occurred frequently in the German Bundestag in the period between 1949 and 1957, a period characterized by considerable party differences, whereas their frequency plunged between 1957 and 1983, which reflected an inter-party consensus on many issues and a desire to avoid public scrutiny.

the high-dimensionality issue inherent to text data (Gentzkow, Shapiro, and Taddy 2019). Second, state elections do not occur at the same time, which ensures that my results are not driven by the idiosyncratic distance from upcoming elections or political leanings.

I obtained the entire population of parliamentary debates for the period of interest of each German state as PDF documents by scraping each state's official website.[9] I then created a dataset that includes all speeches from the 16 German states for the period 2000-2008. This process involved several steps to extract the data from the gathered documents in order to record all the relevant information contained in the documents, such as the speeches, name and role of the speaker, party affiliation, interruptions, state, and date in which the debates occurred.[10] I complemented this dataset with information about the date of the latest and next election and with the shares obtained by the two major German parties, the CDU/CSU and SPD, in the latest election in each state.[11]

The unit of analysis is a speech as recorded in the parliamentary debates. I consider a speech the continuous utterance issued by the same person. During a speech, speakers are often interrupted by remarks of other speakers, applauses etc. Such interruptions are excluded from the speeches.

5.3.2 Topic Classification of Parliamentary Debates

I classify the topic of each speech in the parliamentary debates. This step enables me to compute topic-specific measures of polarization, which are crucial for both my identification strategy and to overcome issues inherent to measuring polarization in parliamentary debates that I describe in the next subsection.

I use a combination of supervised and unsupervised machine learning algorithms.[12] First, I used a supervised machine learning method to classify speeches in a binary way: whether they are about education or not. This approach requires a subset of

[9] Parliamentary debates of Saarland are not available in the official website for the period considered in this analysis. Nevertheless, these debates were made available for my research upon request.

[10] Interested readers can find detailed information about the process of gathering the necessary documents, extracting text from the documents and creating a unified corpus of parliamentary debates in Appendix B.

[11] I retrieved these data from Metawahl, an open-source project that collects data of all German elections (last accessed 7th November 2022).

[12] Interest readers may find a detailed description of this classification task in Appendix C. In this section, I describe only the most relevant aspects.

manually-labelled speeches which are used to train the model. For this purpose, I obtained a set of 3,346 manually-labelled speeches with which I trained a supervised machine learning model. I then used the best-performing model, a Logistic Classifier, to make the out-of-sample predictions for the entire corpus. I report the in-sample performance of the classifier in Table C5.1 and the result of a validation exercise in Table C5.2. Both tables suggest a reliable classification. The share of speeches classified as being about education is 8.9%, or 18,703 speeches.

I then used an unsupervised machine learning model, namely the correlated topic model (CTM), to classify the topic of all the speeches that were classified as *not* being about education in the previous step. The key hyperparameter to tune the CTM is the number of topics. A CTM with 30 topics provided good results in terms of interpretability of the topics. I then aggregated the estimated topics into 11 topics of similar size as the education topic classified in the previous step. I report the estimated topics, most representative words and the assigned label in Table C5.3.

5.3.3 Measuring Polarization in Parliamentary Debates

Measuring polarization in parliamentary debates is challenging. A fundamental problem is that the words used in legislative speeches are a function of both the topic of the debate and the position of the speaker (Lauderdale and Herzog 2016).[13] Hence, the use of different words across MPs from different parties might be mistakenly attributed to polarization when in fact it might be due to MPs discussing different topics. Previous work has dealt with this issue by, for example, limiting the analysis to a single legislative act (Herzog and Benoit 2015), by comparing speeches only within a specific debate (Lauderdale and Herzog 2016) , or, conversely, combining speeches over many debates for each legislator or party, assuming that the resulting documents contain the same mixture of topics (e.g., Giannetti and Laver 2005; Proksch and Slapin 2010). I tackle this issue in a novel way. I first classify the topic of each speech in the parliamentary debates, as explained in the previous subsection. I then compute polarization within each topic, which allows me to isolate the different words used by MPs due to polarization from the different words used due to MPs talking about different topics.

A further issue concerns the finite-sample bias that arises because the pool of words a speaker can choose from is large relative to the total amount of speech we observe

[13] In fact, there are even more sources of variation in word usage. In descending order of importance, these are: language, style, topic, and position (Lauderdale and Herzog 2016). Given the context, it is safe to assume that language and style are reasonably homogeneous within parliamentary debates.

(Gentzkow, Shapiro, and Taddy 2019). This implies that many words are used only by MPs of one party just by chance, and naïve estimators might interpret such differences as evidence of polarization. I tackle this issue by excluding words that are mentioned in less than 10 speeches within a topic from the computation of the polarization measure.[14] This ensures that rare words, which are more likely to be uttered only by MPs of one party just by chance, do not drive my measure of polarization.[15] Germany's 16 state parliaments provide a substantial amount of parliamentary debates in each topic, enabling this approach.

To compute polarization, I first perform standard preprocessing steps such as removal of stopwords, punctuation and numbers. I then transform each speech *d* about topic *s* into an adjusted term-frequency vector according to the following topic-specific term-frequency inverse-document frequency ($tf-idf$) formula:

$$tf-idf_{ds} \equiv \frac{c_{dw}}{\sum_{k\in d} c_{dk}} \times \ln\left(\frac{D^S}{\sum_{n\in D^S} \mathbb{I}(c_{nw}>0)}\right), \tag{5.1}$$

where the relative term frequency of each term *w* in speech *d* ($c_{dw}/\sum_{k\in d} c_{dk}$) is weighted by the natural logarithm of the inverse frequency of the term *w* in all the speeches *D* in topic $s \in S$ ($\ln(D^S/\sum_{n\in D^S} \mathbb{I}(c_{nw} > 0))$).

Compared to the standard $tf-idf$ transformation of a document, the topic-specific $tf-idf$ that I use also upweights words that occur frequently in a document, but downweights words that appear often in many documents *about the same topic.* Hence, words that are mentioned often only in a specific topic will receive less weight, thus alleviating the risk of attributing the use of different words to polarization when in fact it is due to speakers discussing different topics. Further, I also drop rare words, which mitigates the finite-sample bias mentioned previously.[16]

I define polarization as the extent to which opinions on a topic are opposed. Assuming that politicians use different words to express different opinions, the more polarized the speeches, the less similar they are. I therefore use a straightforward measure of

[14] In Section 5.5, I show that results are robust to different thresholds.

[15] I elaborate more on this intuition and formalize it in footnote 29 in Section 5.6. To account for the finite-sample bias, Gentzkow, Shapiro, and Taddy (2019) specify a multinomial model of speech that they estimate through a penalized Lasso model to compute an accurate measure of polarization. Their approach, however, does not account for the different topics MPs address in their speeches.

[16] Formally, for a threshold τ, only words *w* for which $\sum_{n\in D^S} \mathbb{I}(c_{nw} > 0) > \tau$ are kept. In Section 5.5, I show that results are robust to different thresholds used at this stage.

text (dis)similarity: the opposite of the cosine similarity between the vector representation of each speech *d* in topic *s* and state-legislative period cell *l* and the vector representation of all the speeches from a benchmark party *r*, the CDU/CSU,[17] in topic *s* and state-legislative period cell *l*. Formally the polarization of a speech is computed as follows:

$$polarization_{dsl} \equiv -\frac{\sum_i A_{idsl}\bar{B}_{isl}}{\sqrt{\sum_i A_{idsl}^2}\sqrt{\sum_i \bar{B}_{isl}^2}}, \tag{5.2}$$

where A_{idsl} is the $tf-idf$ vector representation of speech *d* in topic *s* and state-legislative period cell *l,* and $\bar{B}_{isl}$ is the average of the vector representation of all the speeches by MPs that belong to benchmark party *r* in topic *s* and state-legislative period cell *l*:

$$\bar{B}_{isl} \equiv \frac{\sum_{p\in r} B_{ipsl}}{\sum_{p\in r} \mathbb{I}(B_{psl})} \tag{5.3}$$

Thus, $\bar{B}_{isl}$ captures the "average" speech of a benchmark party *r* in a specific topic, state, and legislative period. The less similar a speech is to $\bar{B}_{isl}$, the larger the polarization measure. In the next subsection, I provide evidence to validate the polarization measure by showing that it captures differences across parties in word use.

5.3.4 Descriptive Statistics

The entire dataset consists of 622,946 speeches. I drop all the speeches by the President of each state parliament, 327,498 speeches, as these are strictly procedural and not informative of the political debates. I also drop all speeches with less than 100 words, namely 100,816 speeches, as these are too short to be reliably classified among different topics. The resulting sample consists of 210,006 speeches, and descriptive statistics of the dataset are reported in Table 5.1. The average length of a speech is 663.6 words. The share of speeches by ministers of each state parliament is 24%. The share of speeches issued by members of the main center-right party, CDU/CSU, is 34%, while the share for main center-left party, SPD, is 27%. These parties represent the main political forces in Germany and are the only parties that have been part of each German state parliament in the entire period considered. The second tier of political forces in the German landscape in this period is represented by the Green party and the FDP, the liberal party, with a share of speeches of 14% and 11%, respectively.

[17] I show in Section 5.5 that the results are robust to using different benchmark parties or factions.

Speeches from these four parties make up 86% of the entire corpus of parliamentary debates. The remaining 14% of speeches are uttered by member of minor parties, none of which reaches the threshold of 10% of all the speeches in the corpus.[18]

The PISA shock had a substantial impact on the public debate about education. In Figure 5.1, I report the share of respondents from a representative survey of the German population that indicate education as the most or second most important problem in Germany. Such share increased dramatically after the PISA shock. In the two years prior to the PISA shock, only 2.6% of respondents indicated education as the most or second most important problem in Germany on average. This share more than doubled after the PISA shock: on average, 5.7% of respondents indicated education as the most or second most important problem in Germany in the seven years after the PISA shock. The release of the results of the subsequent PISA study, three years later, had a similar impact on the public opinion. It is also interesting to note that the PISA shock triggered an upward trend in the importance of education, as it never reverted to its pre-shock level in the seven years after the shock.

A similar pattern emerges when looking at parliamentary debates. I report the share of speeches about education and the number of times that "PISA" was mentioned in parliamentary debates in Figure 5.2. This figure clearly depicts the "tsunami-like" impact of the release of the first PISA results on the political debate about education. The share of speeches about education increased by 1.8 percentage points after the PISA shock. This effect translates into a 22% increase with respect to the pre-shock share of 7.3% and is statistically significant (see Table A5.3). In the first six months after the PISA shock, the term "PISA" was mentioned more than 2,000 times in parliamentary debates. Overall, "PISA" was mentioned almost 11,000 times after the PISA shock. These figures substantiate the claim that the salience of education increased dramatically because of the PISA shock. I will analyze the impact of this exogenously induced increase in salience of education on the polarization of political debates in Section 5.5.

I report the estimated topics and size in Figure A5.1. With roughly 9% of the speeches, education is a mid-sized topic in the corpus, whereas the largest topic concern

[18] Among these minor parties, the most relevant is the Left party, with its various denomination over time and states (DIE LINKE, the current one or, previously, Linksfraktion, Linkspartei.PDS, PDS, REGENBOGEN), whose share of speeches is 8.8%. Other minor parties include a series of extreme right parties (DVU, DVU-FL, FDVP, NPD, PRO, REP, Ronald-Schill-Fraktion), whose combined share of speeches in the corpus is 2.9%. The remaining 2.1% of speeches are uttered by MPs of local parties (0.96%), MPs whose party could not be identified (0.8%), or without a political affiliation (0.37%).

economic issues and the lawmaking process. For about 5% of the speeches no clear topic could be identified, and I therefore assigned the label "Other" to this topic. In Figure A5.2, I report the speeches' topic size by state. No major difference in the distribution of topics across states can be observed. The education topic, in green, appears to be quite homogenous across states.

Finally, I report evidence to validate the polarization measure in Figure A5.3. As expected, the polarization measure aggregated at the party level is much lower for the CDU/CSU when the CDU/CSU is used as the benchmark party in the left panel. By this measure, the average speech from a member of the SPD is 0.48 SD more polarized than the average speech from a member of the CDU/CSU party. Similarly, the average polarization measure for members of the CDU/CSU is much larger when the SPD is used as the benchmark party. This suggests that the polarization measure captures meaningful differences in word use across MPs of different parties.

5.3.5 Additional Data Sources: State-Specific PISA Results and Bills

The performance in the PISA 2000 reading test of each German state is reported in Table A5.1. State-specific results were released on the 25th of June 2002, almost seven months after the PISA shock. There is a large heterogeneity in the performance. The average score of the best performing German state, Bayern, is 62% of a standard deviation higher than the lowest performing state, Bremen. Such difference corresponds to the distance between the best performing state in the reading test of PISA 2000, Finland, and Germany, whose performance was well below the OECD average. It is also important to note that the state-specific results of Berlin and Hamburg were not released due to low participation rates.

I also use data from the "Pattern of Lawmaking in the German Länder" dataset (Stecker, Kachel, and Paasch 2021), which comprises all 16,610 bills that have been initiated in the 16 German state parliaments between 1990 and 2020. The dataset contains a wealth of information regarding the bills. For the purpose of my analysis, the main variables of interest are the initial date on which the bill was initiated, the status of each bill—whether the bill was adopted, rejected or other—, the topic of each bill, which has been manually coded, and a German state identifier. For consistency with the rest of the analysis, I use data for the period January 2000 - August 2008 and report their descriptive statistics in Table A5.2. Specifically, I report the total number of initiated bills by each topic as defined in the dataset, the share of bills by each topic, as well as the total number of bills by their status. With 525 initiated bills, education is the largest topic in the dataset and covers 10% of the bills. Reassuringly, this share is

very close to the share of speeches about education in parliamentary debates (roughly 9%, see Figure A5.1). Since topics in the law-making dataset were manually coded, this improves the credibility of the classification task I carried out for this project. The other topics in the law-making dataset are more narrowly defined than those that I estimated for the parliamentary debates, which makes the comparison less meaningful. During the period of interest, 5,356 bills were initiated. Out of all the initiated bills, 4,116 (76.9%) have been adopted, while 821 (15.3%) were rejected. Thus, the large majority of initiated bills have been adopted, which reflects the fact that bills tend to be initiated by governing parties who have the political power to adopt them.[19] The status of the remaining 419 (7.8%) bills, labelled as "Other", includes exceptional cases of bills which have been withdrawn, discontinued, adjourned etc.

5.4 Empirical Strategy

Estimating the causal effect of the salience of a topic on polarization in parliamentary debates requires exogenous variation in the salience of a topic. As argued in the previous sections, the PISA shock in Germany led to an exogenous increase in the salience of the education topic, which rules out issues of reverse causation. It therefore provides an ideal setting to study its impact on the polarization of parliamentary debates.

I exploit the fact that the PISA shock affected a single topic, education, to implement a difference-in-differences strategy.[20] The key idea is that speeches about unaffected topics act as counterfactuals for speeches about education that occurred after the PISA shock, thus accounting for underlying trends in polarization of parliamentary debates and for time-invariant differences among polarization in different topics. I therefore estimate the following equation:

$$y_{islt,r\neq b} = \theta_s + \alpha PostPISA_t + \beta PostPISA \times Ed_{st} + \gamma' X_{ilt,r\neq b} + \sigma_l + \varepsilon_{islt,r\neq b} \quad (5.4)$$

[19] 92% of bills that are eventually adopted have been initiated by governing parties, whereas 99% of rejected bills have been initiated by opposition parties. Hence, there is almost a complete overlap between adopted (rejected) bills and bills initiated by governing (opposition) parties. A minority of bills have been initiated by bipartisan coalitions (3.7%), and they have been adopted in 97% of the cases.

[20] I show in Section 5.5 that the PISA shock did not affect polarization in other topics.

The outcome variable $y_{islt,r\neq b}$ denotes the polarization between speech *i* by member of party *r* and all the speeches of benchmark party *r* = *b* in topic *s* and state-legislative period cell *l* at time *t*. Speeches from the benchmark parties are therefore omitted from the analysis. θ_s denotes topic fixed effects, which account for differences in level of polarization across topics and the dummy variable $PostPISA_t$ accounts for differences before and after the PISA shock, which occurred on the 4th of December 2001. The interaction term, $PostPISA \times Ed_{st}$ takes value one if a speech occurred after the PISA shock and if it is about education. In this setup, the parameter of interest β can be estimated by means of the two-way fixed effects estimator (TWFE), which accounts for time-invariant differences between treated and untreated units.

$X_{ilt,r\neq b}$ is a vector of speech, state, and time specific controls, such as the length of the speech *i*, the shares of the two main parties, CDU/CSU and SPD, at time *t* in state-legislative period cell *l*, whether the speech *i* is given by a member of a governing party, is given by a minister, distance from the next election in state-legislative period cell *l* at time *t,* year and party fixed effects. The length of speech *i* plays an important role as a control, since it is negatively correlated with the polarization measure and including it in the regression causes a substantial increase in the R^2 of the model. However, including it as a control is potentially problematic if the PISA shock also affected the verbosity of the speeches. At the same time, it ensures that the results are not driven by an increase or decrease in the verbosity of the speeches. σ_l denotes state-legislative period fixed effects; that account for differences in the level of polarization across state-legislative period cells. $\varepsilon_{islt,r\neq b}$ is the idiosyncratic error. I standardize the polarization measure to have mean zero and standard deviation one to interpret the estimated coefficients in terms of standard deviation. I cluster standard errors at the state level throughout the paper.

The identification strategy rests on the assumption of parallel trends of the treated and untreated units. In this application, this means that the polarization in education debates would have trended similarly to other topics in the absence of the PISA shock. While this assumption is not directly testable, I exploit the availability of multiple time periods before the shock to show the absence of different pre-trends between education and other topics in Section 5.5.1.

Another identifying assumption is that the effect of the PISA shock affected the polarization of education debates through topic salience. The effect could also be driven by the negative results of German students revealed by the PISA study rather than the salience of the education topic. I tackle this issue in Section 5.5.2, where I

exploit the fact that in June 2002, six months after the PISA shock, the results for all but two German states were published. Despite the large heterogeneities in the performance of German states and the fact that results were not published for two states, I show that the effect of the PISA shock on polarization was homogenous across German states.[21] Further, the low performance of German students in international standardized assessment was already shown by the TIMSS study in 1995, but this event was largely ignored by the German media (see Section 5.2.1). Hence, the results revealed by the PISA study were not completely new to German MPs. This further corroborates the assumption that the effect on polarization was driven by the salience induced by the PISA shock rather than the information revealed by the PISA study.

5.5 Results

5.5.1 Main Results

I report evidence of the validity of the parallel-trends assumption using an event-study design in Figure 5.3, where I interact the dummy variable indicating whether a speech is about education and year fixed effects. The figure does not show diverging trends in the period prior to the PISA shock, and I cannot reject the null hypothesis of pre-event effects being zero, thus suggesting that polarization in political debates about education and other topics were following the same trend before the shock. Conversely, the test of post-event effects being jointly null is largely rejected. It can also be noted that the impact of the PISA shock on polarization seemingly fades out over time and that polarization reverts to its pre-shock level only about six years after the shock.

I provide further evidence of the validity of the parallel trend assumption in Figure A5.4 and Figure A5.5. In Figure A5.4, I report point estimates of the pre-trends by interacting the education dummy with six-month bins instead of yearly bins to increase the number of pre-trend point estimates. Even in this specification, I do not find significantly different pre-trends between education and other topics, although standard errors become substantially larger. In Figure A5.5, I show the dynamic of polarization in all the estimated topics in the period of interest net of the controls and

[21] Note that in this setting heterogenous treatment effects would not bias the TWFE estimator. As the recent literature on difference-in-differences methods noted, heterogenous treatment effects can bias the TWFE estimator if units are treated at different point in times (see Roth et al. 2022 for a review of this literature), which is not the case in this setting.

fixed effects described in Equation (5.4). The polarization of education debates clearly increased after the shock, while no similar patterns can be detected for other topics. In sum, both figures provide evidence in favor of the validity of the parallel trend assumption.

I report the estimates of Equation (5.4) in Table 5.2. The magnitude of the impact varies between 8% of a SD in the most parsimonious specification in Column 1, and 11.1% SD in a specification that also includes state-legislative period, party, and year fixed effects in Column 2. All coefficients are statistically significant. The main difference between Column 2 and Column 3 concerns the inclusion of the length of a speech as a control, which causes a decrease in the estimated coefficient to 8.8% SD. At the same time, including it more than doubles the R^2 of the model. I therefore prefer the most restrictive specification in Column 3, which should be therefore considered as a conservative estimate.[22] An increase of 8.8% SD in polarization is equivalent to 18% of the polarization between the main center-right (CDU/CSU) and center-left (SPD) parties.[23] Overall, these results show that the PISA shock had a substantial and persistent impact on the political debates about education.

5.5.2 State-Specific Heterogeneity

As argued in Section 5.5.4, a possible concern regarding identification strategy is that the impact of the PISA shock on polarization is not due to the increased salience of education. The new information revealed by the PISA study about the low performance of German students might have also caused the increase in polarization. To test this hypothesis, I leverage the fact that the initial PISA shock, which occurred on 4th of December 2001, was followed by a state-specific PISA shock on the 26th of June 2002. On this date, German state-specific results for all but two states were released and revealed large heterogeneities in the performance of German states (reported in Table A5.1).

I therefore investigate whether the impact of the PISA shock differed with respect to the actual performance of each state. To this purpose, I first create an additional

[22] If the verbosity of the speeches was affected by the PISA shock, the inclusion of length of speeches as a control could be problematic. In fact, I find weak evidence that the PISA shock caused speeches in education to become roughly 5.5% shorter by substituting the logarithm of length of speech as the outcome variable in Equation (5.4). Nonetheless, including length of speech as a control ensures that the impact of the PISA shock on polarization occurred above and beyond the verbosity of the speeches.

[23] The share is the absolute value of the estimated coefficient (0.088) divided by the difference between the polarization measure for the CDU/CSU and the SPD (0.48) when the CDU/CSU is used as the benchmark party reported in Figure A5.3.

treatment variable ("PISA shock (State)") to capture whether a speech occurred after the state-specific PISA shock of the 26th of June 2002. I then interact this variable with a "PISA-Published-Score" dummy variable, that takes value one for the states of Berlin and Hamburg for which the PISA state results were not published. Second, I interact the dummy "PISA shock (State)" with a set of dummies that capture whether each state's performance was in the lower, middle, or upper tercile of the distribution of performance of German states. I further explore this hypothesis by interacting the "PISA shock (State)" treatment with the performance of each German state.[24] Results in Table 5.3 show that the impact of the PISA shock was homogenous not only with respect to whether state specific results were published or not (Column 2), but also with respect to the actual performance of each German state (Column 3 and 4).[25]

The lack of sizable heterogeneity across states also emerges in Table A5.3, which shows little differences of the impact of the PISA shock on the share of education speeches. The only marginally significant difference emerges with respect to the states for which the state-specific results were not published, namely Hamburg and Berlin (Column 3). The share of speeches about education increased slightly less after the PISA shock in these states. Overall, these results suggest that the salience of the topic, rather than the actual performance of the students revealed by the PISA study affected the polarization of the debates.

5.5.3 Heterogeneity by Party

I explore which parties contributed the most to the increase in polarization in Table 5.4. It is worth reminding that, since the benchmark party is the CDU/CSU, party interactions capture the polarization of each party with respect to the CDU/CSU. Results show that the increase in polarization is driven by a cleavage between the two main parties, the CDU/CSU and the SPD. In fact, the interaction between treatment dummy and the SPD dummy is positive and reaches a 10% level of statistical significance in Column 5, where all the interactions are included. Conversely, the FDP and the Green Party do not appear to contribute substantially to the increase in polarization. To corroborate these results, I repeat the analysis using the polarization

[24] PISA tests three subjects: math, reading, and science. In each wave, PISA has a special focus on one of the three subjects. Since reading was the focus of PISA in the first wave, I use the performance in reading (reported in column 1, Table A5.1); using math or science performance leads to the same results (table not shown).

[25] Another potentially interesting dimension of heterogeneity concerns former West and East German states. Again, I do not find statistically significant differences in the impact of the PISA shock on the polarization of education debates in former West and East German states (results not shown).

measure between left- and right-wing parties, which allows me to include speeches from all the parties in the regression.[26] I report results from this specification in Table A5.4. Again, the increase in polarization seems to be driven by the CDU/CSU and SPD, whose associated coefficient is positive in Column 1 and 2, and reaches statistical significance when all the interactions are included in Column 5. Results from this section are compatible with a cleavage theory framework, where the main center-right and center-left parties exploit the increased salience of education induced by the PISA shock to amplify their ideologically distinctiveness.

5.5.4 The Impact of the PISA Shock on the Number of Bills

Topic salience might also affect the number of bills discussed in parliaments. Bills are the main output of parliaments and, therefore, they represent a proxy of parliaments' productivity. I investigate whether MPs respond to the salience of a topic by increasing their effort concerning such topic. I use data on law-making in German state parliaments collected by Stecker, Kachel, and Paasch (2021), which allows me to implement essentially the same identification strategy described in Section 5.4, where my treated group consists of bills about education initiated after the PISA shock. The outcome variable is the logarithm of the number of bills in each topic and state in a six-month bin. I assign bills to the six-month bin in which the bill was initiated. Estimated coefficients from this log-linear model can be therefore interpreted as percentage changes in the number of bills. I report results for the overall number of proposed bills, as well as separately for rejected and adopted bills.

Results indicate a 16.2-21.1% increase in the total number of proposed bills about education because of the PISA shock (Column 1-3). This suggests that MPs indeed put more effort into this topic. At a closer look, the effect is driven by the number of rejected bills (Column 5). As discussed in Section 5.3.5, virtually all rejected bills are proposed by the opposition. It is therefore possible that that MPs in the opposition strategically propose more bills in a salient topic to signal to voters their effort in this topic, despite the very low chances of such bills being adopted.

It is interesting to note that in this context both the polarization and the number of bills in education increased. This is surprising given that polarization has often been linked with gridlocks in parliament (Jones 2001; Binder 2004; Lapinski 2008; McCarty,

[26] I show in Section 5.5 (Table 5.6, column 2) that the main results are essentially the same when using this measure of polarization. The advantage of this measure is that for each speech of members of right-(left-)wing parties, all the speeches from the members of the left-(right-)wing parties in the same topic, state, and legislative period are used as benchmark speeches.

Poole, and Rosenthal 2016), which hinders the law-making process. My results suggest that an increase in polarization and vibrant law-making can coexist, although they do not offer a clear interpretation of the relationship between these two concepts, which lies outside the scope of this paper.

5.5.5 Robustness Checks

A first concern about the validity of my main results regards the polarization measure. The choice of a benchmark party for the computation of the polarization measure entails a certain degree of arbitrariness. I therefore compute alternative measures of polarizations by varying the benchmark parties of faction.[27] In Table 5.6, I show that using speeches of different parties or factions as a benchmark does not appreciably alter the main results. I only report the results using the most restrictive specification, which controls for topic, state-legislative period, party and year fixed effects, as well as the controls described in Equation (5.4).

In Column 1, I report the results obtained using speeches of the SPD as the benchmark party. In Column 2 and 3, I do not use a single party as the benchmark to compute the polarization measure. Instead, I compute the cosine similarity between each speech of right (left)-wing parties and all the speeches from the left (right)-wing parties within the same topic, state, and legislative period. I report results for this polarization measure in Column 2. Similar to Column 2, in Column 3 I report the results obtained computing the cosine similarity between each speech from a governing party and all the speeches from parties in the opposition, and vice versa.

Differently from the specification using a single party as the benchmark corpus to compute the polarization measure, these specifications allow me to include all speeches in the regressions, since an appropriate benchmark exists for all speeches. This comes at the cost of using as a benchmark a corpus of speeches which is more heterogenous, as it comprises speeches of different parties. In fact, despite the substantial increases in the number of observations, the standard errors in Column 2 and 3 do not decrease appreciably, possibly due to the heterogeneity of the benchmark corpus.

Regardless of the benchmark party or faction chosen, the results are remarkably robust. The coefficient estimated in the main specification and reported in Table 5.2, Column 3 (0.088), lies between the coefficient obtained when using the SPD as the

[27] The measures differ because of the different speeches used in Equation (5.3) to compute the "average" speech $\bar{B}_{isl}$ against which the polarization measure is computed.

benchmark party in Column 1 (0.90), and the coefficient estimated when using the left/right-wing polarization measure in Column 2 (0.086).

Second, I conduct a robustness check to ensure that the observed effect is due to the PISA shock and not to other events, such as the release of results of subsequent PISA studies, which occurs every three years, or other events that might affect the polarization in the counterfactual topics. To this purpose, I restrict the sample to speeches that occurred two years before and two years after the shock. This specification also ensures a balanced sample size of the pre- and post-shock period. I report estimates of this specification in Table A5.5. Results are very similar to the main results in Table 5.2 and, if anything, larger in magnitude in the preferred estimated in Column 3 (0.096 SD).

Another concern regards the number of topics. As discussed in Section 5.3 and, more in detail, in Appendix C, the number of topics chosen depends on a variety of factors, such as the size of the corpus, previous knowledge of the researcher, and the downstream task one wants to achieve. To balance the interpretability of the topics and ensure that topics were of similar size to the education topics, I estimated a CTM with 30 topics, which I then aggregated into 11 topics. In Table A5.6, I report the results obtained estimating a CTM with the following number of topics: 9, 10, 11, 13, and 15. To maximize the transparency of this exercise, I also do not aggregate the topics as done previously. Results suggest that using a number of topics similar to the number of aggregated topics that I use does not affect the main results substantially. This suggests that neither the number of topics chosen, nor the aggregation step are driving the results in the preferred specification.

I further corroborate my findings by conducting a placebo test, where I test the effect of the PISA shock on the polarization of the other topics. Had the PISA shock also affected the polarization of other topics, estimates might be biased, since the affected topics would not constitute an appropriate counterfactual. I report results in Figure A5.6. In each row I report the coefficient obtained interacting the PISA shock dummy with a dummy for the topic indicated in each row along with 95% confidence intervals. I report only the estimated coefficient obtained using the preferred specification described in Equation (5.4), which includes topic, state-legislative period, party and year fixed effects and controls for the length of each speech and distance from elections.

In the first row, I report the results from the main specification, where education is the treated topic. In the subsequent rows, I report the coefficients from the placebo

exercise. Besides the coefficient for education, only the coefficients for the topic "Local Politics" and "Social Welfare, Healthcare and Equality" reach the 10% threshold of statistical significance, while the other coefficients do not reach any conventional threshold of statistical significance. I cannot entirely rule out that these effects are due to the PISA shock, but other events occurred in the period 2000-2008 might also have affected the polarization in such topics. As results reported in Table A5.7 show, when restricting the placebo exercise to a symmetric time window around the shock (2000-2004), the placebo coefficients in Column 2 and 3 for "Local Politics" and "Social Welfare, Healthcare and Equality", respectively, are not statistically significant anymore, whereas the coefficient for education in Column 1 remains positive and statistically significant. This suggests that the change in polarization in these topics is likely due to other events that occurred after the PISA shock.

To a large extent, results from the placebo exercise alleviate the concern that the effect of the PISA shock on the polarization of education debates is biased by the simultaneous impact of the PISA shock on the polarization of other topics. I further address such concern with a leave-one-topic-out exercise, where I iteratively estimate Equation (5.4) by dropping one of the counterfactual topics at each iteration. This robustness check shows that results are not driven by any topic in the counterfactual group that might have been affected by the PISA shock or other events. I report results in Figure A5.7 with 95% confidence intervals. In the first row, I include all the topics and coefficient is the therefore same as the coefficient reported in Table 5.2, Column 3. In the subsequent rows, I report the estimated coefficient obtained by dropping the topic indicated in each row. The estimated coefficients are relatively stable and remain statistically significant regardless of which topic is excluded from the estimation sample.

Finally, I report results obtained by changing the threshold above which words are kept to compute the polarization measure in Table A5.8. As mentioned in Section 5.3.3, to avoid the sample-finite bias in the polarization measure I only use words that are mentioned in at least ten speeches. This ensures that rare words, which are more likely to be uttered only by MPs of one party just by chance, do not drive the polarization measure. I have therefore computed alternative measures of the main polarization measure with the CDU/CSU as the benchmark party obtained by imposing more restrictive thresholds. In Column 1-3, I report results obtained by using words that are mentioned in at least 20, 30 or 40 speeches within a topic. In Columns 4-6, I report results obtained using words that are mentioned in at least 2%, 2.5%, and 5% of speeches within a topic. Results are robust to these different thresholds.

5.6 Polarizing Issues in Education Debates

5.6.1 Polarization Score

In the previous section, I showed that polarization in education debates increased as a consequence of the PISA shock. I have shown that the effect was mainly driven by the two main center-right and center-left parties, the CDU/CSU and SPD, respectively. In this section, I provide suggestive evidence on what are the most polarizing issues in education debates. I focus on the two main parties that drove the increase in polarization, the CDU/CSU and the SPD, and on debates about education. For each term in $w \in W$, where W denotes the vocabulary of terms uttered by MPs of either the CDU/CSU or SPD in debates about education, I develop a polarization score $p(w)$, which is defined as follows:

$$p(w) \equiv \frac{f\,(w_{CDU}) - f\,(w_{SPD})}{f\,(w_{CDU}) + f\,(w_{SPD})} \times \ln\left(f\,(w_{CDU}) + f\,(w_{SPD})\right), \tag{5.5}$$

where $f\,(w_{CDU})$ ($f\,(w_{SPD})$) denotes the total number of times the term w is mentioned by the CDU/CSU (SPD). The first part of the score varies between -1 and 1, where 1 (-1) indicates terms that have only been mentioned by MPs that belong to the CDU/CSU (SPD). This part is weighted by the natural logarithm of the total number of times the term w has been mentioned by either the CDU/CSU or the SPD.

The rationale for this polarization score is simple. In absolute value, terms that display high polarization scores are those that *(i)* tend to be mentioned more often by one party and *(ii)* are mentioned often. Terms that are uttered the same number of times by both parties will get a polarization score of 0. Terms that are uttered more often by one party but are relatively infrequent will be pushed toward zero.[28] Hence, the

[28] Note that, in the extreme case where a term w is mentioned only once and, therefore, is mentioned only by one party, $p(w) = 0$, since $f\,(w_{CDU}) + f\,(w_{SPD}) = 1$ and $\ln(1) = 0$.

polarization score of rare terms, for which there is a higher probability that they are uttered only or mostly by one party just by chance,[29] will be pushed toward 0.

The polarization score closely mirrors the polarization measure that I use throughout the analysis, which is based on the cosine similarity between a corpus of speeches from a benchmark party and speeches from the other parties. Similar to the cosine similarity, the polarization score depends on both the frequency with which one term is used by one party and on its absolute frequency. This ensures that terms that have high polarizing scores are also those that drive the polarization measure in the education debates.

5.6.2 Polarizing Issues in Education

I focus on the 10,000 most frequent terms uttered by either member of the CDU/CSU and SPD in education speeches, after removing uninformative terms such as stopwords, names, and numbers. This ensures that these terms are unlikely to obtain large polarization scores just by chance (see footnote 29). On average, these terms are mentioned 254,7 times and 50% of the terms are mentioned at least 88 times. The minimum frequency of a term is 35. I report the distribution of the polarization score in Figure A5.8. I rescaled the polarization score to have a zero mean and divided it by $\max(|p(w)|)$, so that $-1 \leq p(w) \leq 1$. The distribution is quite concentrated around the mean; the standard deviation of the distribution is 0.16 and for 50% of the terms $|p(w)| \leq 0.1$.

I focus on the 250 terms with the largest polarization score for each party, or the top 5% of polarizing terms. For the CDU/CSU (SPD), these terms lie in the black (red)-shaded area in Figure A5.8. These terms have a polarization score $|p(w)| \geq 0.32$. I display the 250 terms with the highest CDU/CSU (SPD) polarization score in Figure A5.9 (Figure A5.10), translated in English (Panel (a)) and in the original language (German, Panel (b)).

[29] To formalize this intuition, let us consider a generic term w, for which $f(w_{CDU}) + f(w_{SPD}) = N$, with $N \in \mathbb{N}, N > 0$. Let us assume that w is a neutral term, i.e., that each realization of $f(w_{CDU})$, without loss of generality, is equally likely: $f(w_{CDU}) \sim U(0, N)$ and $E[f(w_{CDU})] = \frac{N}{2}$. Hence, the probability that the term w is uttered by only one party is: $P(f(w_{CDU}) = 0) + P(f(w_{CDU}) = N) = \frac{1}{N} + \frac{1}{N} = 2/N$. Thus, the smaller the N, or, equivalently, the rarer the term w, the higher the probability that w is uttered only by MPs of one party just by chance, since $P(f(w_{CDU}) = n)$ strictly decreases in N. More generally, for an arbitrarily small $n \in \mathbb{N}$, $P(f(w_{CDU}) \leq n) = n/N$. Thus, the larger the N, the lower the probability that terms are mentioned primarily by MPs of one party just by chance.

A variety of findings emerge from the most polarizing terms. I primarily focus on those issues that were particularly relevant in the aftermath of the PISA shock. As mentioned in Section 5.2.1, the three most important issues that emerged from the PISA shock were: developing a monitoring system with common educational standards and central examination, expanding "all-day school" offers, and reforming the tracking system. It is interesting to notice that terms related to these issues can be found among the most polarizing terms in Figure A5.9 and Figure A5.10.

Concerning a monitoring system with common educational standards and central examination, the term "state exams" (*"Landesprüfungen"*) appears as a strongly polarized term favored by the CDU/CSU. Conversely, the term "learning assessments" ("*Lernstandserhebungen"*) is a strongly polarized term favored by the SPD. This terminology suggests polarized views on the ways to monitor the education systems: while the CDU/CSU favored a testing regime of central state exams, which are typically high-stake exams for students, the SPD seemingly favored a testing regime aimed at monitoring student achievement in a low-stake environment. As a matter of fact, state exams, in particular those at the end of high school in Germany, have been introduced in most states in the years after the PISA shock. In 2000, only 7 states had a central upper secondary school leaving examination ("*Zentralabitur"*).[30] From 2004 to 2008, this examination was gradually rolled out to all German states except Rhineland-Palatinate (Helbig and Nikolai 2015). At the same time, a plan to establish a new set of common standards was also implemented. In 2004, the Institute for Quality Development in Education (IQB) was created to develop math, reading, writing and foreign-language standards and accompanying tests (Neumann, Fischer, and Kauertz 2010; OECD 2011).

A second issue concerns the expansion of all-day schooling. Again, a term linked to this concept can be found among strongly polarized terms: the term "all-day elementary school" ("*Ganztagsgrundschule"*) is a polarized term favored by the SPD. This picture is in line with the account by Kuhlmann and Tillmann (2009), according to which the SPD was promoting the expansion of the all-day schooling offer since the end of 2001, as it considered it an effective policy to improve equal opportunities for students.[31] Conversely, the CDU/CSU considered all-day schooling as a threat to the

[30] The states are: Bavaria, Baden-Württemberg, Mecklenburg-Vorpommern, Saarland, Saxony, Saxony-Anhalt.

[31] Relatedly, the term "equality of opportunity ("*Chancengleichheit"*) also features among the strongly polarized terms favored by the SPD.

family and, therefore, hindered its expansion for a long time. Despite the different stances toward this issue, the offer of all day-schools in Germany was rapidly expanded thanks to large subsidies granted by the German national government through the investment program "*Future, Education, and Care* (2003-2009)" (IZBB).[32]

A third relevant issue was the tracking system. Until 2000, the large majority of students in all German states were tracked into three main different ability schools at the age of 10. Given the large educational inequality highlighted by the PISA shock across German students with different socio-economic and, in particular, migration backgrounds, strong arguments were made against the existing three-tiered early-tracking system. Again, terms related to this concept can be easily found among the most polarized terms. For example, the terms "sorting" or "selecting" ("*sortieren"* and "*aussortieren*", respectively) are terms typically used by the SPD. Conversely, the term "comprehensive school" ("*Einheitsschule*") is a strongly polarized term used by the CDU/CSU. A comprehensive school is opposed to the three-tier school system typical of Germany. While some states enacted reforms to reduce the segregation induced by the early-tracking system,[33] the distinction between three hierarchical school tracks has been mostly left intact (Henninges, Traini, and Kleinert 2019).

Overall, this section offers suggestive evidence on three possible issues that might have led to an increase in the polarization of education debates. It is interesting to note that polarization in two of these topics, the monitoring system and the all-day schooling, was accompanied by important and substantial reforms on these issues. Conversely, the tracking system was not largely addressed by the reforms.

5.7 Conclusion

The rise of polarization observed in many democracies has fueled a lively debate on the causes of such phenomenon. While research on the determinants of polarization in the electorate abounds, much less is known about what drives polarization in political speech. In this paper, I shed light on topic salience as a possible determinant

[32] Detail of the program can be found at https://www.ganztagsschulen.org/de/service/izbb-programm/das-investitionsprogramm-zukunft-bildung-und-betreuung-izbb (last accessed: 16 December 2022).

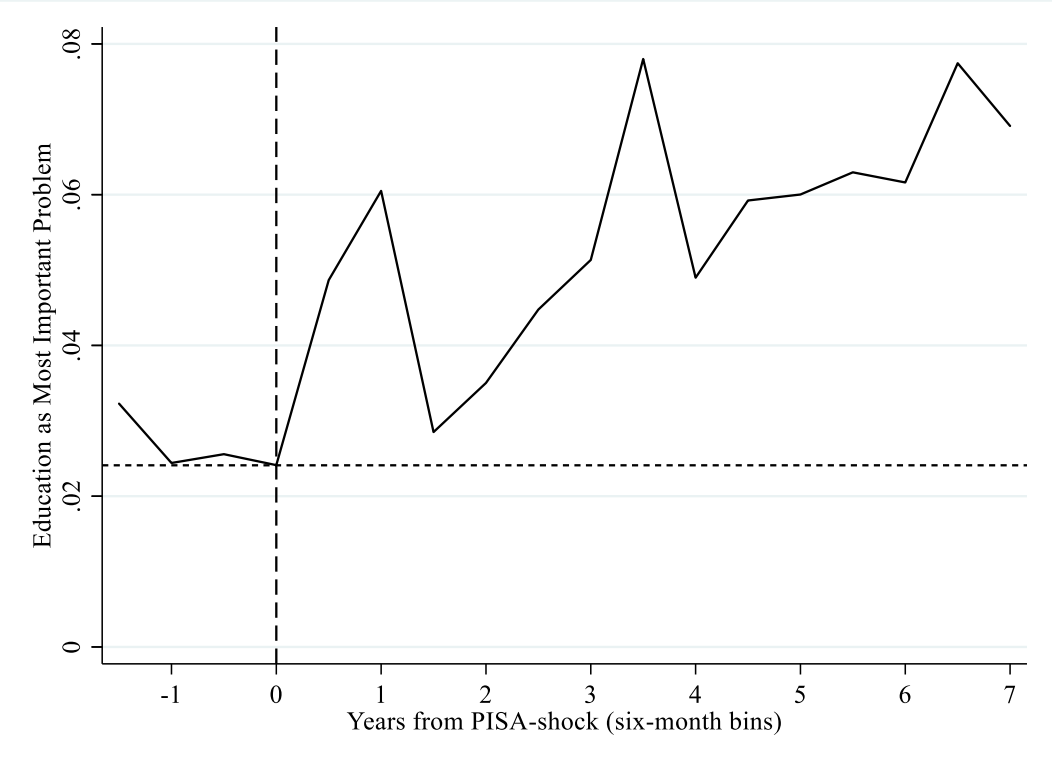
[33] For example, some states have merged the two lower-level tracks ("*Realschule*" and "*Hauptschule*") into one school, called regional schools ("*Regionalschulen*") (Davoli and Entorf 2018). Despite this trend toward a two-tier education system, the issue of access to the academic track ("*Gymnasium")*, which constitutes the main route to a tertiary degree, has not been addressed (Henninges, Traini, and Kleinert 2019).

of polarization in parliamentary debates. I find that the sharp increase in the salience of education induced by the PISA shock in Germany had a strong and long-lasting impact on the polarization of debates about education. I do not find heterogeneities across states, despite considerable differences in the performance of students in different states revealed by the PISA shock. These results lend support to the cleavage theory of political behavior as opposed to a convergence toward the median platform, whereby MPs amplify their ideological distinctiveness with respect to a salient topic. The results are robust to different measures of polarization, to different numbers of topics in the counterfactual group and to a variety of robustness checks. I also find an increase in the number of initiated bills about education, which is driven by rejected bills. The simultaneous increase in polarization and in the number of initiated bills is an interesting pattern which challenges previous findings in the literature, that have often linked high polarization with gridlocks in parliament.

I also provide suggestive evidence that issues related to developing a monitoring system with common educational standards and central examination, expanding all-day school offers, and reforming the tracking system led to the increase in polarization of education debates. While the first two topics were subject to substantial reforms in the aftermath of the PISA shock, the tracking system was not largely addressed. This provides further evidence that polarization in parliamentary debates and the legislative process do not necessarily overlap.

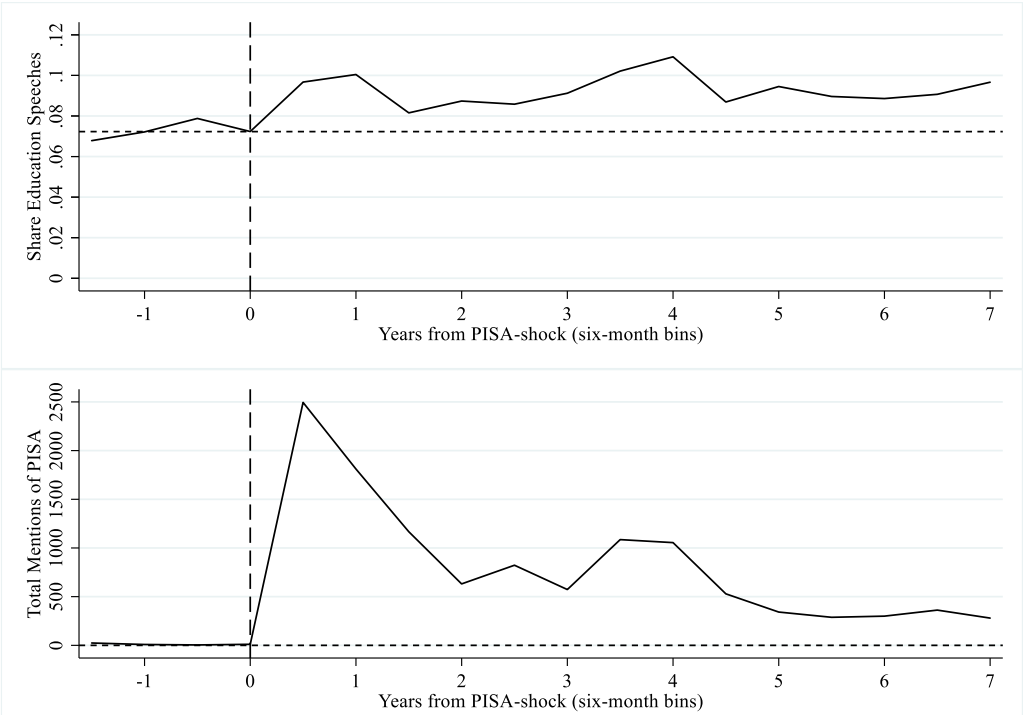
Figures and Tables

Figure 5.1: Education as Most Important Problem



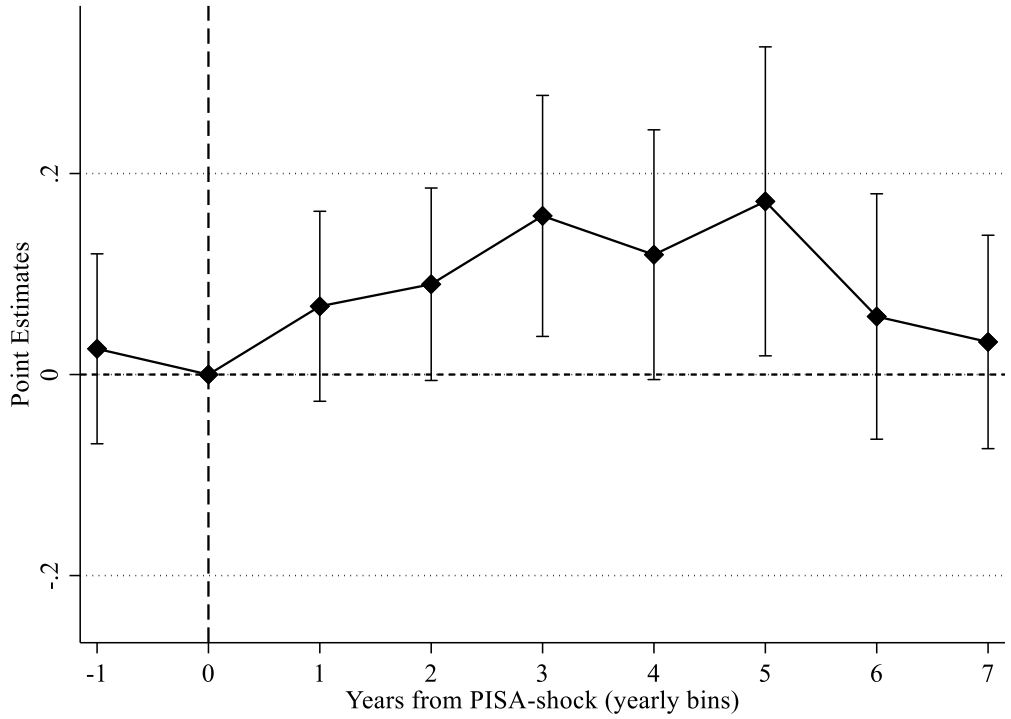
Note: Data source: Politbarometer (Forschungsgruppe Wahlen 2019). The y-axis reports the share of respondents that indicated education as the most or second most important problem in Germany. The x-axis reports the distance (in years) from the PISA shock, which occurred on the 4th of December 2001. Data are aggregated into six-month bins.

Figure 5.2: The "Tsunami"-like Impact of PISA



Note: The figure reports the share of education speeches in parliamentary debates in the upper panel and the total number of mentions of the term "PISA" in parliamentary debates in the lower panel. The x-axis reports the distance (in years) from the PISA shock, which occurred on the 4th of December 2001. Data are aggregated into six-month bins.

Figure 5.3: The Impact of the PISA Shock on Polarization in Education Debates: Event-Study Graph



Note: Event-study estimates of the impact of the PISA shock on polarization with 95% confidence intervals. The estimated equation takes the following form: $y_{islt,r\neq b} = \theta_s + \alpha PostPISA_t + \sum_{\tau\in(-1,7),\tau\neq 0} \beta_\tau Ed_s \times \mathbb{I}(t+\tau) + \gamma' X_{ilt,r\neq b} + \sigma_l + \varepsilon_{islt,r\neq b}$. This event-study setup takes the same form as Equation (5.4), but instead of pooling years before and after shock, I interact the education dummy with an indicator variable for each year $(t+\tau)$. I label t the year before the PISA shock, which I consider the reference year. The pre-shock period covers the years $t-1$ and t, while the post-shock period covers the years from $t+1$ to $t+7$ (until August 2008). Standard errors have been clustered at the state level. The dependent variable is the standardized polarization. The x-axis reports the distance (in years) from the PISA shock, which occurred on the 4th of December 2001. The year prior to the PISA shock is the excluded category. The *p*-values of the joint hypothesis tests of zero pre- and post-event effects are 0.603 and 0.001, respectively.

Table 5.1: Descriptive Statistics

	Mean (1)	SD (2)	Min/Max (3)
Word Count	663.57	(622.52)	100.0-17503.0
Share CDU/CSU	0.34	(0.47)	0.0-1.0
Share SPD	0.27	(0.44)	0.0-1.0
Share GREENS	0.14	(0.34)	0.0-1.0
Share FDP	0.11	(0.32)	0.0-1.0
Share Ministers	0.24	(0.42)	0.0-1.0
Share Gov. Speeches	0.53	(0.50)	0.0-1.0
Share Education Speeches	0.09	(0.28)	0.0-1.0
# Observations		210,006	
# States		16	
# Parl. Sessions		3,277	

Note: Descriptive statistics of speeches from parliamentary debates. The share of speeches is reported separately only for parties for which the total number of speeches is larger than 10% of the entire corpus of speeches. The number of observations coincides with the number of speeches.

Table 5.2: PISA Shock and Political Polarization in Education Debates – Difference-in-Differences

	(1)	(2)	(3)
PISA shock × Education	0.080*	0.111**	0.088**
	(0.042)	*(0.042)*	*(0.038)*
Topic FE	Yes	Yes	Yes
State-Legislative Period FE	No	Yes	Yes
Party, Year FE	No	Yes	Yes
Controls	No	No	Yes
R^2	0.148	0.260	0.535
Observations	137,820	137,820	137,820

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include topic fixed effects and a dummy for whether the speeches occurred after the PISA shock. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 5.3: Heterogeneity by State-Specific Performance

	(1)	(2)	(3)	(4)
PISA shock *(Federal)* × Education	0.058	0.058	0.048	0.049
	(0.038)	*(0.038)*	*(0.033)*	*(0.033)*
PISA shock *(State)* × Education	0.033	0.033	0.033	0.033
	(0.036)	*(0.036)*	*(0.024)*	*(0.024)*
PISA shock *(State)* × Education × PISA Published Score		0.000		
		(0.017)		
PISA shock *(State)* × Education × Med. Perf. Perf.			0.009	
			(0.017)	
PISA shock *(State)* × Education × High Perf.			-0.022	
			(0.026)	
PISA shock *(State)* × Education × PISA Perf./ 100				-0.070
				(0.057)
Topics, State-Legisl. Period, Party, Year	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
R^2	0.628	0.620	0.635	0.635
Observations	137,820	137,820	119,462	119,462

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include a dummy for whether the speeches occurred after the federal or state PISA shock, topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The variable, "PISA shock *(Federal)*" is a dummy variable which takes value one if a speech occurred after 4th December 2001. The variable "PISA shock *(State)*" is a dummy variable that takes value one if a speech occurred after 26th June 2002. The medium performance variable takes value one if the performance of the respective state is in the middle tercile, while high performance takes value one if the performance is in the upper tercile. In Column 3, the omitted category is the lower tercile. The variable "PISA Performance" is the performance of each state in the PISA 2000 reading test. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 5.4: Heterogeneity by Party

	(1)	(2)	(3)	(4)	(5)
PISA shock × Education	0.088**	0.056	0.086*	0.094*	-0.003
	(0.038)	*(0.038)*	*(0.044)*	*(0.048)*	*(0.087)*
PISA shock × Education × SPD		0.071			0.130*
		(0.051)			*(0.062)*
PISA shock × Education × FDP			0.015		0.103
			(0.059)		*(0.101)*
PISA shock × Education × GREENS				-0.028	0.069
				(0.076)	*(0.103)*
Topics, State-Legisl. Period, Party,	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
R^2	0.535	0.535	0.535	0.535	0.535
Observations	137,820	137,820	137,820	137,820	137,820

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include a dummy for whether the speeches occurred after the PISA shock, topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 5.5: PISA Shock and Bills about Education – Difference-in-Differences

	All Bills			Adopted	Rejected
	(1)	(2)	(3)	(4)	(5)
PISA shock × Education	0.201**	0.211**	0.162**	0.121	0.261*
	(0.079)	*(0.077)*	*(0.073)*	*(0.081)*	*(0.130)*
Topic FE	Yes	Yes	Yes	Yes	Yes
State FE	No	Yes	Yes	Yes	Yes
Year FE	No	No	Yes	Yes	Yes
R^2	0.240	0.250	0.255	0.238	0.307
Observations	2,931	2,931	2,931	2,510	547

Note: Difference-in-differences estimate of the impact of the PISA shock on the number of bills about education in all German states. The dependent variable is the natural logarithm of the total number of bills about each topic in a state, topic, six-month-bin cell. Each observation corresponds to a state-topic-six-month-bin cell. All regressions include a dummy for whether the speeches occurred after the PISA shock. In Columns 1-3, all the bills are used, regardless of their status. In Column 4 and 5, I restrict the sample to accepted bills and rejected bills, respectively. The data include all proposed bills from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 5.6: Main Results with Different Benchmark Parties or Factions

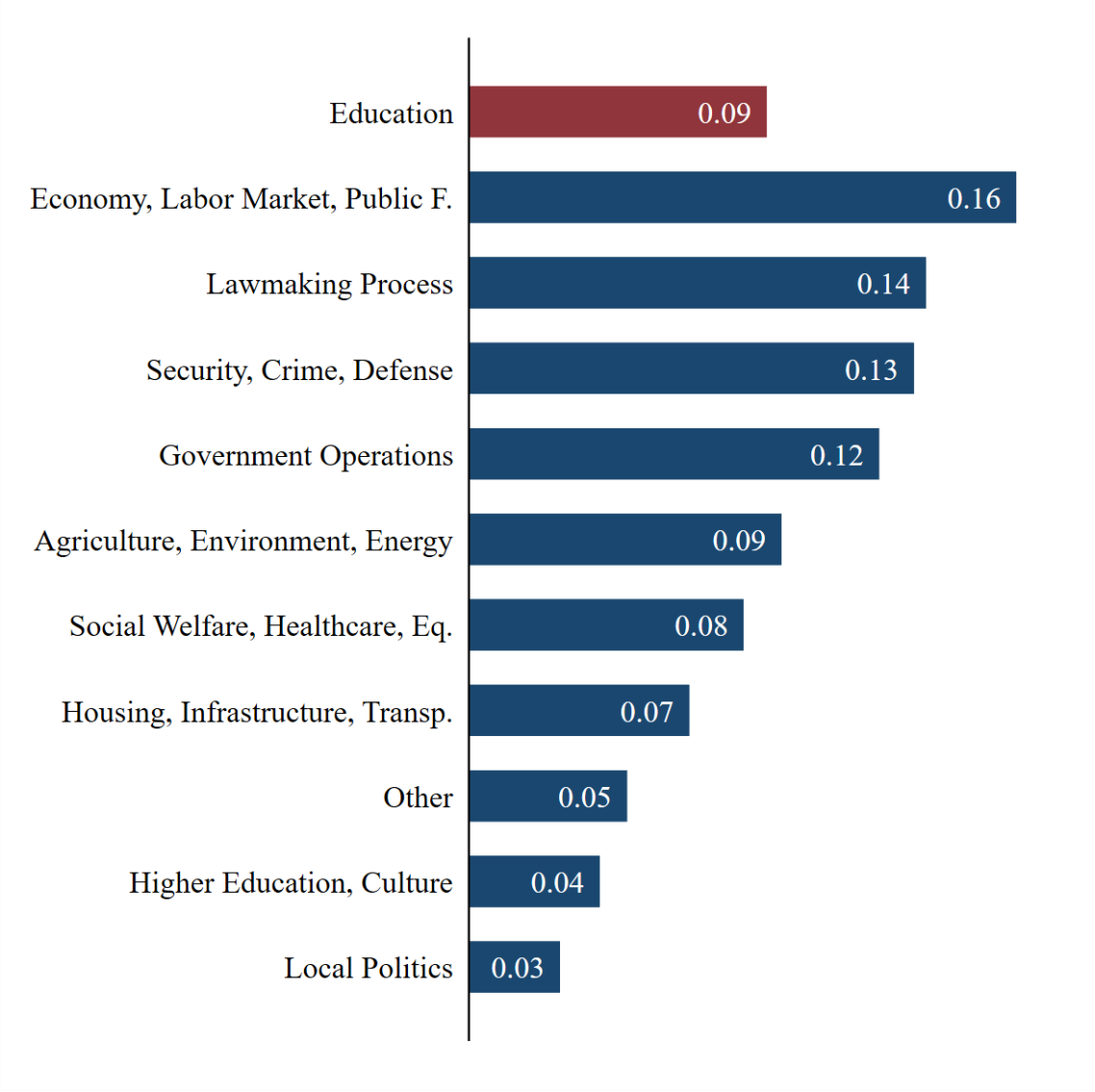
	SPD (1)	Left/Right (2)	Gov./Opp. (3)
PISA shock × Education	0.090** *(0.038)*	0.086** *(0.033)*	0.077* *(0.037)*
Topic, State-Legisl. Period, Party, Year FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes
R^2	0.535	0.553	0.560
Observations	152,464	205,160	209,459

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. The dependent variable is the standardized polarization with SPD as the benchmark party in Column 1, speeches of parties of the opposite wing (i.e., left or right) as the benchmark corpus in Column 2, and speeches of opposite the coalition (i.e., governing or opposition) as the benchmark corpus in Column 3. All regressions include a dummy for whether the speeches occurred after the PISA shock, topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Appendix

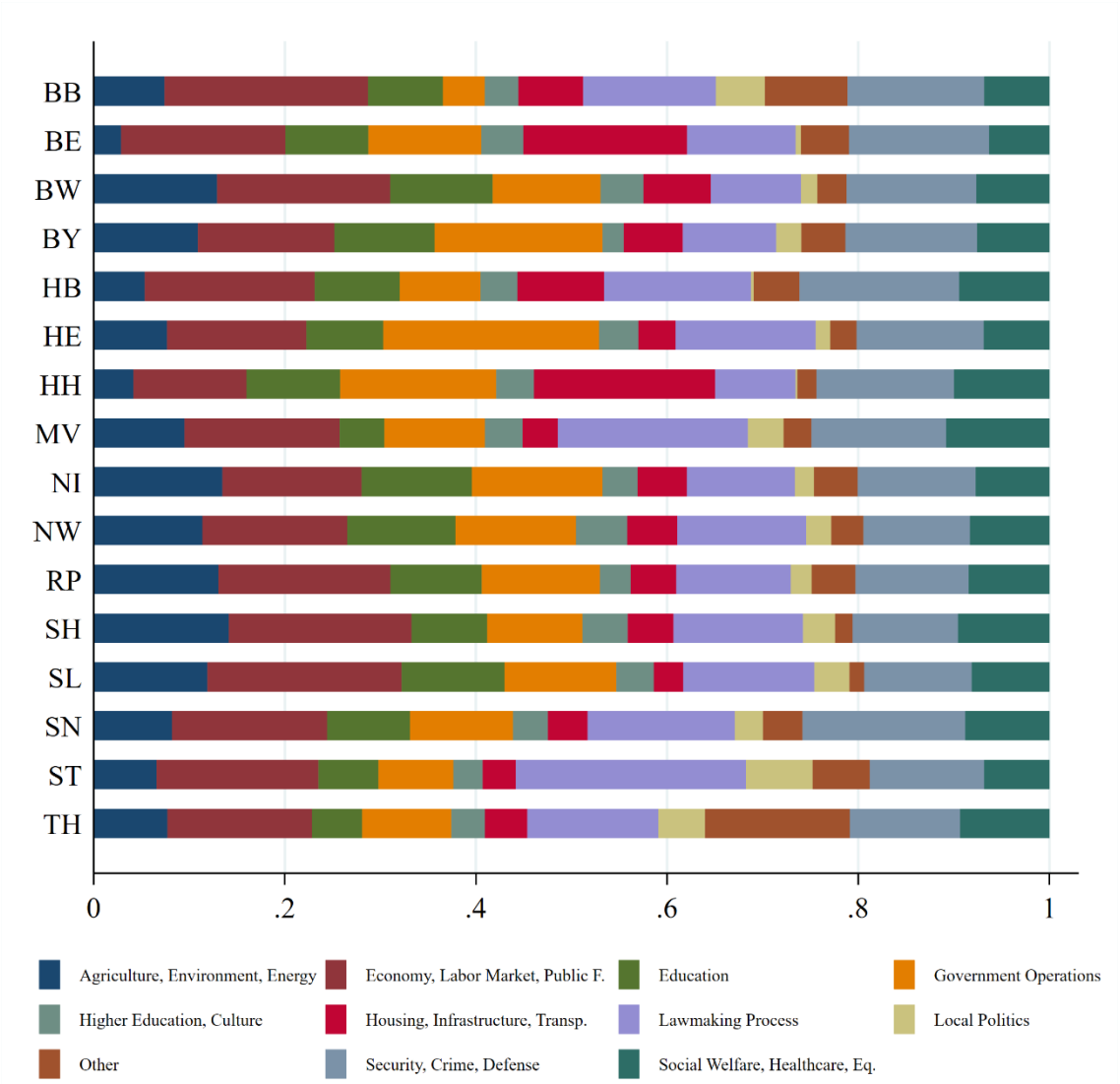
Appendix A: Additional Figures and Tables

Figure A5.1: Share of Speeches' Topics



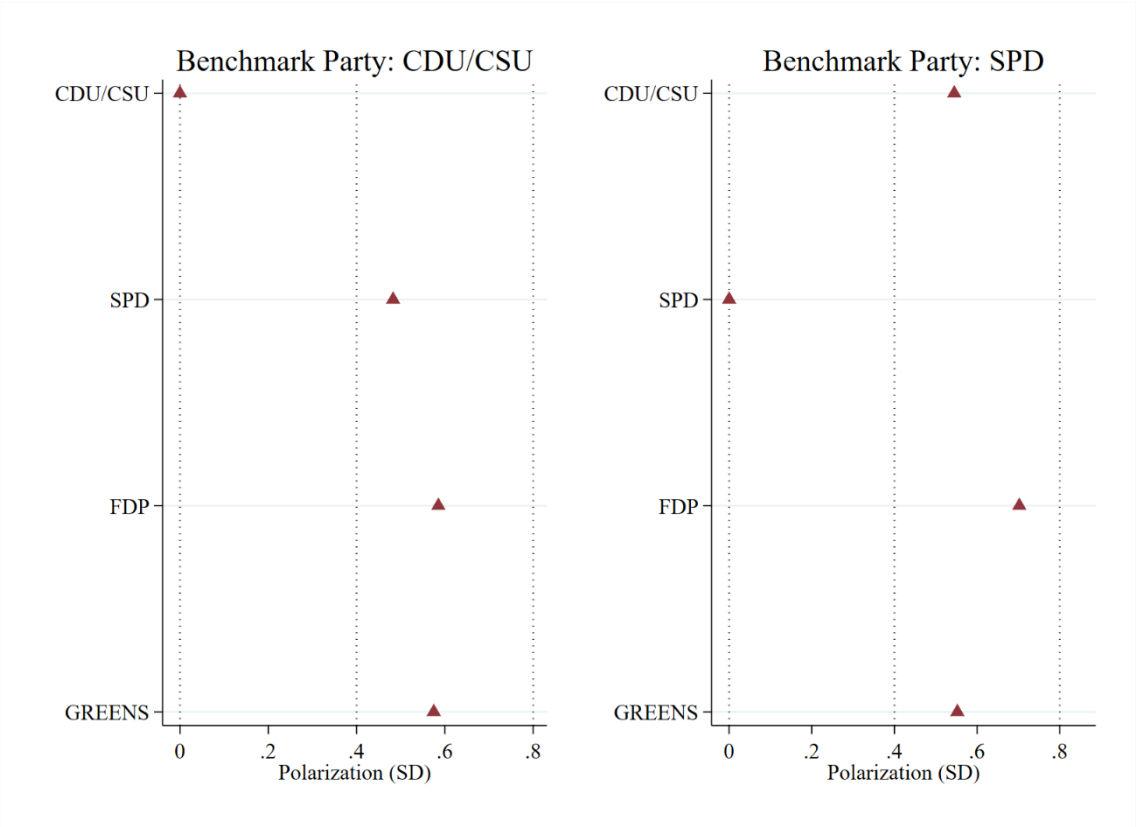
Note: I classified the topic "education", in red, with a supervised machine learning algorithm. The remaining topics have been classified using an unsupervised machine learning algorithm, namely correlated topic modelling, and assigning the topic with the highest weight to each speech. Details of the classification task are provided in Appendix C.

Figure A5.2: Share of Speeches' Topics, by State



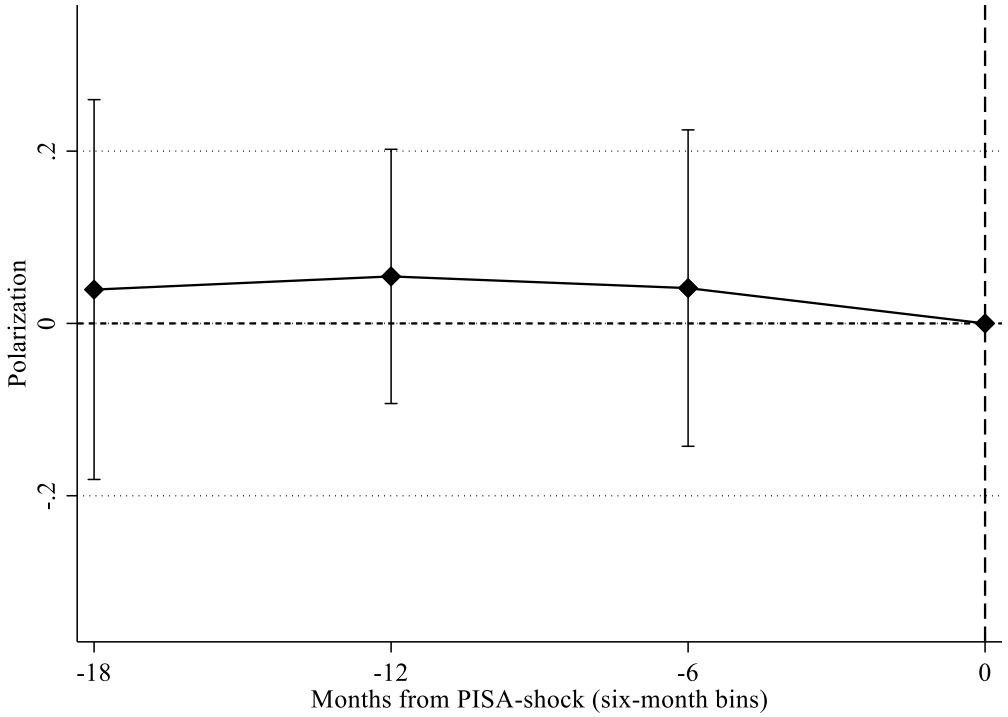
Note: Share of speeches' topics by state. Details of the topic classification task are provided in Appendix C. The codes identifying German states on the y-axes are the official 2-letter acronyms and correspond to the following states: Brandenburg (BB), Berlin (BE), Baden-Württemberg (BW), Bavaria (BY), Bremen (HB), Hessen (HE), Hamburg (HH), Mecklenburg-Vorpommern (MV), Lower Saxony (NI), North Rhine-Westphalia (NW), Rhineland-Palatinate (RP), Schleswig-Holstein (SH), Saarland (SL), Saxony (SN), Saxony-Anhalt (ST), Thuringia (TH).

Figure A5.3: Polarization by Party



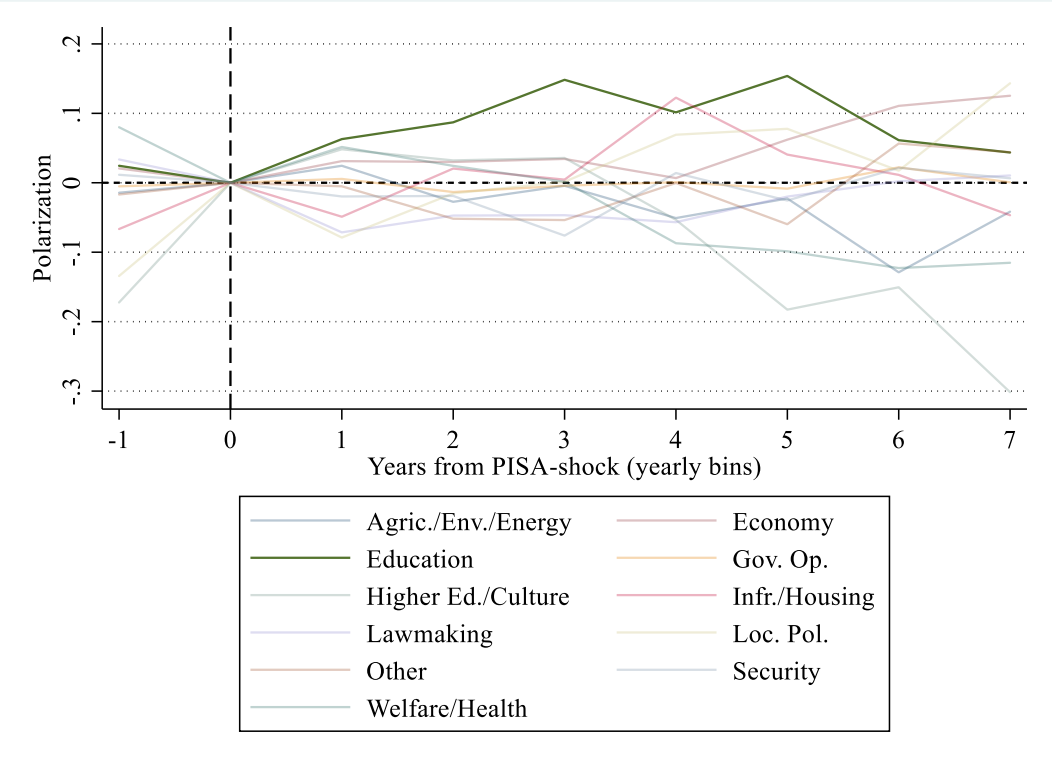
Note: The figure reports the average polarization measure aggregated at the party level with respect to a benchmark party. The polarization measured has been divided by its standard deviation and recentered around the average polarization of the CDU/CSU party on the left panel and around the average polarization of the SPD party on the right panel. The x-axis can therefore be interpreted in terms of standard deviation. The polarization measure consists of the opposite of the cosine similarity between all the speeches from a benchmark party (CDU/CSU in the left panel and SPD in the right panel) and all the other speeches in the same topic and legislative period.

Figure A5.4: Pre-Trends in Polarization



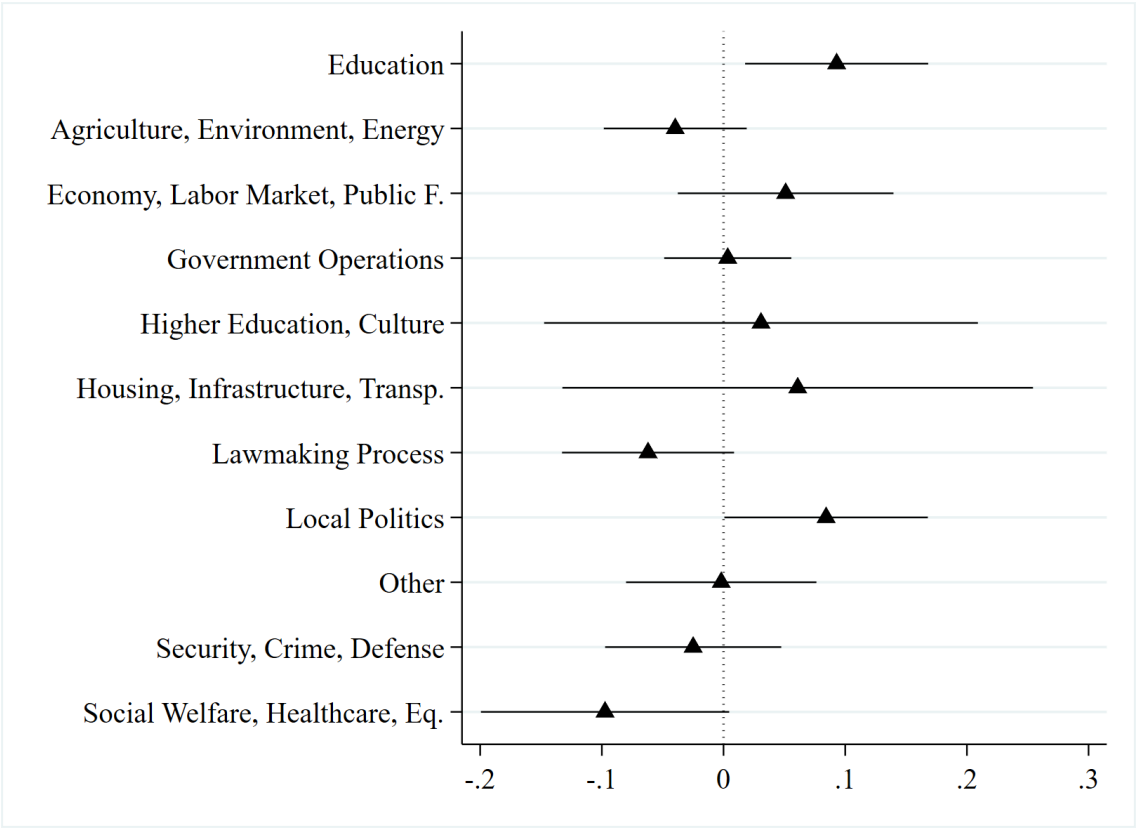
Note: The graph plots coefficients and 95% confidence interval from the interaction between the dummy variable indicating whether a speech is about education and six-month fixed effects. Standard errors have been clustered at the state level. The dependent variable is the standardized polarization. Only pre-trends are reported. The x-axis reports the distance (in six-month bins) from the PISA shock, which occurred on the 4th of December 2001. The six-month bin prior to the PISA shock is the excluded category. Standard errors are clustered at the state level. The *p*-values of the joint hypothesis test of the pre-trend coefficients being different from 0 is .901.

Figure A5.5: Trends in Residualized Polarization by Topic



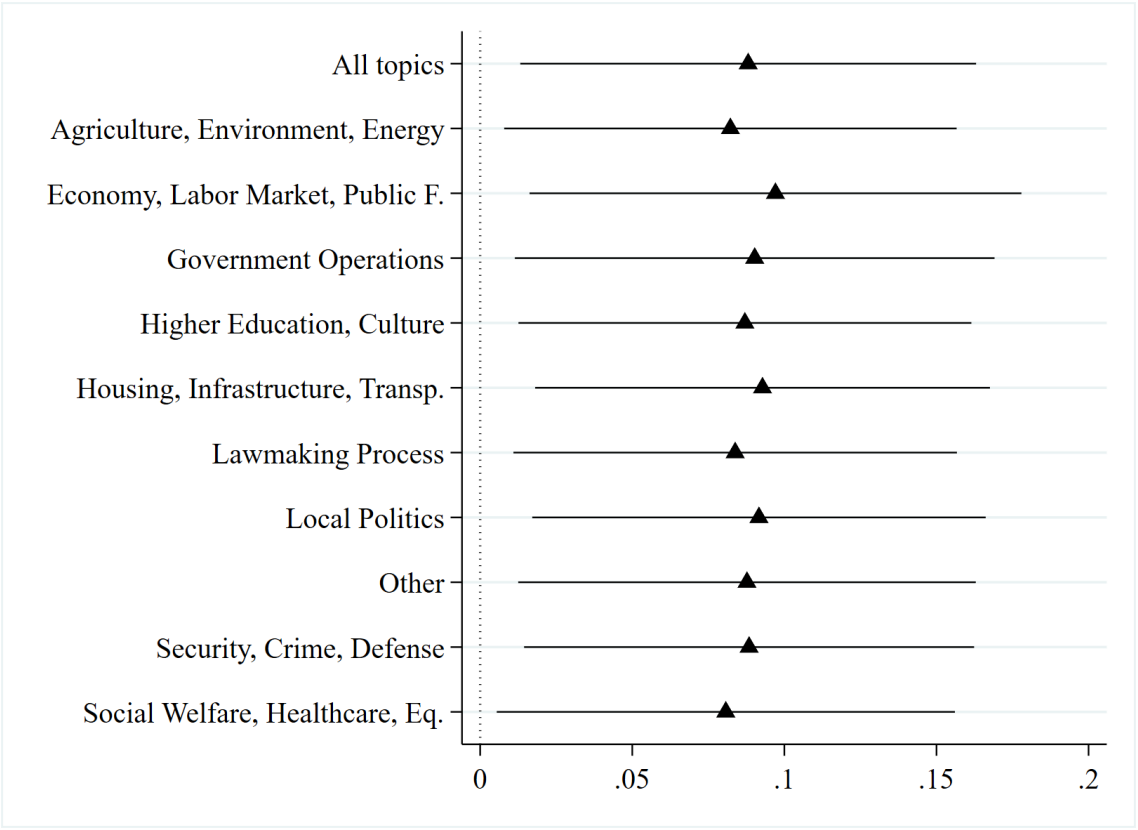
Note: The figure reports the average standardized and residualized polarization measure over time for each topic. The polarization measure has been residualized of the controls and fixed effects in Equation (5.4). The measure has been normalized to 0 in the year before the shock for each topic.

Figure A5.6: Placebo with Other Topics



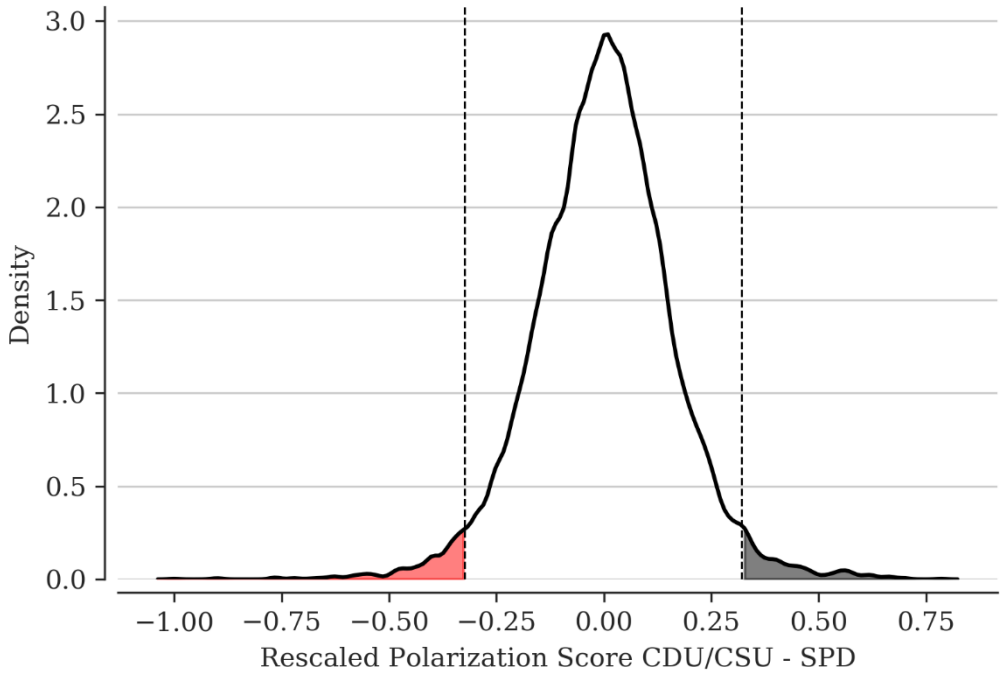
Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of the topic indicated in each row with 95% confidence intervals. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include a dummy for whether the speeches occurred after the PISA shock, topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors have been clustered at the state level. In the first row, I report the coefficient of the impact of the PISA shock on the polarization of education speeches, i.e, the "true" shock.

Figure A5.7: Leave-One-Topic-Out



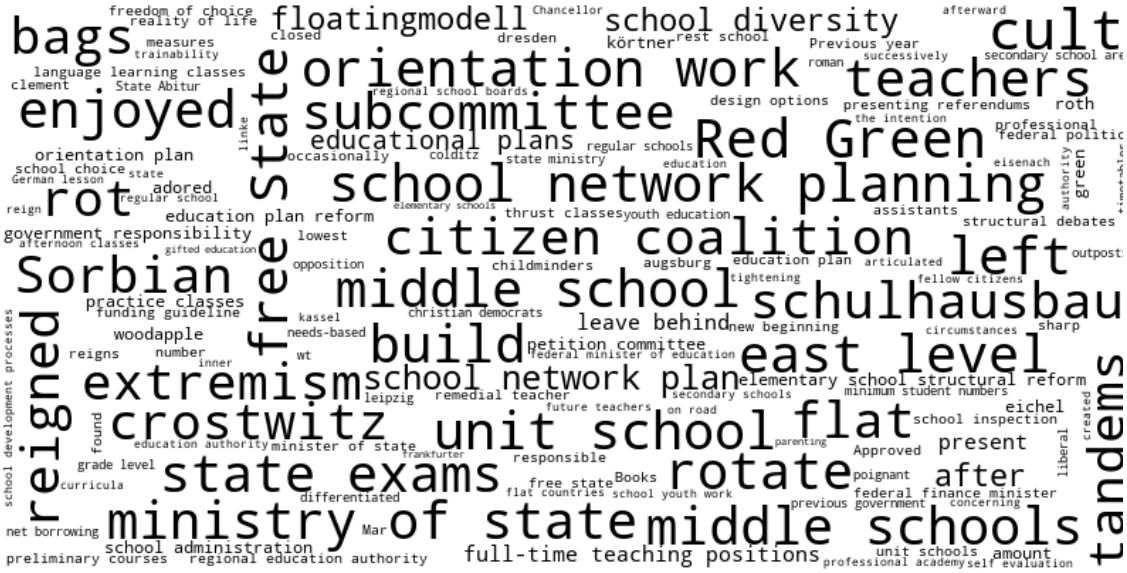
Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of the education topic obtained by dropping the topic indicated in each row with 95% confidence intervals. In the first row, all topics are included. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include a dummy for whether the speeches occurred after the PISA shock, topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors have been clustered at the state level.

Figure A5.8: Density Plot of Rescaled Polarization Score (CDU/CSU – SPD)



Notes: The figure reports the density plot of the rescaled polarization score between the CDU/CSU and the SPD. Positive (negative) values indicate terms that are uttered more often by MPs of the CDU/CSU (SPD) with respect to members of the SPD (CDU/CSU). The score has been centered around zero and divided by the maximum of the absolute value of the polarization score $\max(|p(w)|)$, so that $-1 \leq p(w) \leq 1$. Terms with polarization scores in the black-(red-)shaded area are the top-250 terms in terms of polarization score for the CDU/CSU (SPD) and are depicted in Figure A5.9 (Figure A5.10).

Figure A5.9: Most Polarizing Words - CDU/CSU

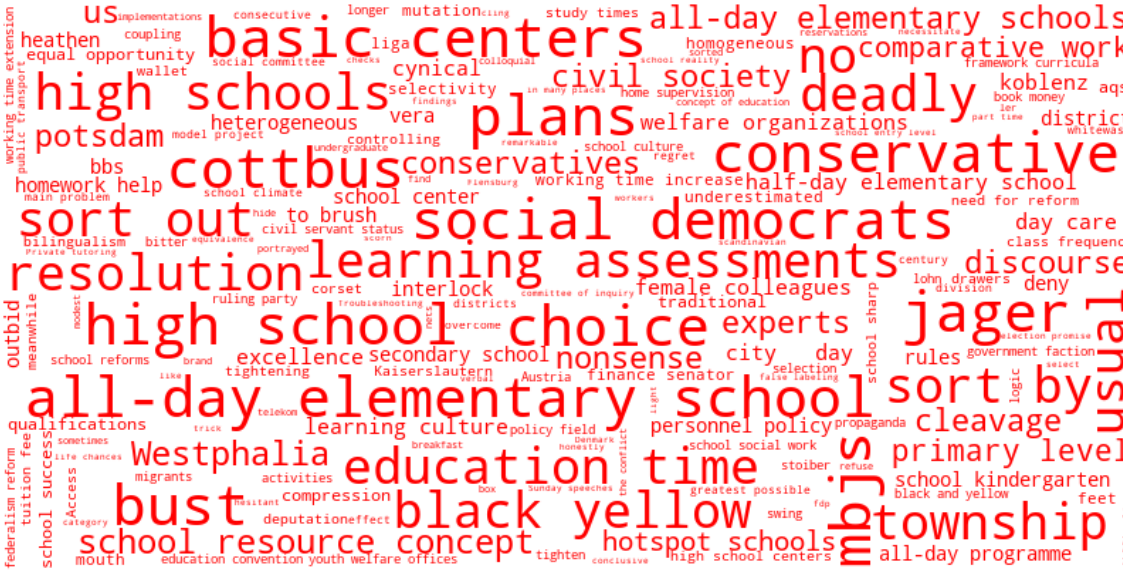


(a): English (translated)

(b): German (original)

Notes: The figure reports the 250 most polarizing words for the CDU/CSU. In Panel (a), the words have been translated into English using the Python package *deep_translator*. In Panel (b), I report the original German terms. The font size of the words increases with the polarization score $p(w)$ of each term.

Figure A5.10: Most Polarizing Words - SPD



(a): English (translated)

(b): German (original)

Notes: The figure reports the 250 most polarizing words for the SPD. In Panel (a), the words have been translated into English using the Python package *deep_translator.* In Panel (b), I report the original German terms. The font size of the words increases with the polarization score $p(w)$ of each term.

Table A5.1: State-Specific Results in PISA 2000

State	PISA State Score Reading (1)	Deviation from federal mean (2)	Position in international PISA ranking (3)
Bavaria	510	26	11
Baden-Württemberg	500	16	18
Saxony	491	7	23
Rhineland-Pfalz	485	1	25
Saarland	484	0	27
North Rhine-Westphalia	482	-2	29
Thuringia	482	-2	30
Schleswig-Holstein	478	-6	33
Hessen	476	-8	34
Lower Saxony	474	-10	36
Mecklenburg-Vorpommern	467	-17	38
Brandenburg	459	-25	40
Saxony-Anhalt	455	-29	42
Bremen	448	-36	44

Note: The table reports the average performance in reading of each German state in Column 1, the distance from the average German performance in Column 2, and position in the international PISA ranking in Column 3. Data have been taken from Artelt et al. (2002). Results for Berlin and Hamburg were not made public due to these states not meeting the prescribed threshold of sample size.

Table A5.2: Bills by Topic and Status

Topic	Num-	Shar	Topic	Num-	Shar
Education	525	0.10	Taxes & Dues	67	0.01
Political System & Parties	507	0.09	Social Matters	67	0.01
Other	426	0.08	Justice and Laws	51	0.01
Communal Matters	376	0.07	Social Welfare	49	0.01
State Budget	269	0.05	Housing	44	0.01
Justice and Security	232	0.04	Europe	43	0.01
Government Officials	230	0.04	Regional Planning	41	0.01
Health	228	0.04	Immigration & Integration	40	0.01
Economy	223	0.04	Lottery/Gambling Industry	40	0.01
Environment	185	0.03	Culture	35	0.01
Labor	177	0.03	Animals	33	0.01
Media	175	0.03	Data Protection	29	0.01
Family/Children/Youth	164	0.03	Civic Duties	29	0.01
Administration	140	0.03	Data	28	0.01
Taxes & Finances	138	0.03	Pension/Seniority/Retirement	27	0.01
Judicial System	112	0.02	Agriculture	25	0
Construction	110	0.02	Religion	23	0
Finances	99	0.02	Technology	19	0
Equality	89	0.02	Energy	15	0
Civil rights	83	0.02	Community Financing	11	0
Traffic and Transportation	75	0.01	Defense	6	0
Society	70	0.01	International Matters	1	0

Total Number of Bills	5,356
Total Number of Accepted	4,116
Total Number of Rejected Bills	821
Total Number of Bills with	419

Note: The table reports the total number and share of bills by topic, and the total number of bills by status (accepted, rejected, or with "other" status) for the period January 2000 – August 2008. The data come from the "Patterns of Lawmaking in the German Lander" dataset (Stecker, Kachel, and Paasch 2021). The original topic names in German can be found in Stecker, Kachel, and Paasch (2021).

Table A5.3: The Effect of the PISA Shock on the Share of Education Speeches

	(1)	(2)	(3)	(4)	(5)
PISA shock (*Federal*)	0.018***	0.024***	0.024***	0.021***	0.022***
	(0.004)	*(0.006)*	*(0.005)*	*(0.006)*	*(0.006)*
PISA shock (*State*)		-0.008*	0.004	-0.009	-0.117
		(0.004)	*(0.008)*	*(0.009)*	*(0.110)*
PISA shock *(State)* × PISA Published Score			-0.014*		
			(0.007)		
PISA shock (*State*) × Med. Perf.				-0.002	
				(0.009)	
PISA shock (*State*) × High Perf.				0.009	
				(0.009)	
PISA shock (*State*) × PISA Perf./100					0.023
					(0.023)
Mean DV (Pre-shock)			0.073		
State-Legisltative Period FE	Yes	Yes	Yes	Yes	Yes
Party FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
R^2	0.009	0.009	0.009	0.009	0.009
Observations	210,006	210,006	210,006	185,729	185,729

Note: The table reports OLS estimate of the impact of the PISA shock on the share of education speeches. The dependent variable is a dummy variable indicating whether a speech is about education. The variable, "PISA shock (*Federal*)" is a dummy variable which takes value one if a speech occurred after 4th December 2001, when the first PISA results were released. The variable "PISA shock (*State*)" is a dummy variable that takes value one if a speech occurred after 26th June 2002, the date on which state specific results were released. The PISA Published Score dummy variable takes value zero for the states of Berlin and Hamburg (for which state-specific results were not published) and one otherwise. The medium performance variable takes value one if the performance of the respective state is in the middle tercile, while high performance takes value one if the performance is in the upper tercile. The "PISA Performance" variable represents the performance on each state in the reading test as reported in Table A5.1 (Column 1). Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A5.4: Heterogeneity by Party – Left-Right Polarization Measure

	(1)	(2)	(3)	(4)	(5)
PISA shock × Education	0.068	0.078**	0.090**	0.093**	0.012
	(0.039)	*(0.031)*	*(0.033)*	*(0.036)*	*(0.056)*
PISA shock × Education × CDU/CSU	0.047				0.103**
	(0.030)				*(0.048)*
PISA shock × Education × SPD		0.026			0.092*
		(0.038)			*(0.045)*
PISA shock × Education × FDP			-0.038		0.040
			(0.041)		*(0.060)*
PISA shock × Education × GREENS				-0.053	0.028
				(0.057)	*(0.068)*
Topics, State-Legisl. Period, Party,	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
R^2	0.553	0.553	0.553	0.553	0.553
Observations	205,160	205,160	205,160	205,160	205,160

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. The dependent variable is the standardized cosine similarity between each speech from the left (right) parties and all the speeches from the right (left) in the same topic, state and legislative period. All regressions include a dummy for whether the speeches occurred after the PISA shock, topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A5.5: Symmetric Time Window (2000-2004)

	(1)	(2)	(3)
PISA shock × Education	0.076*	0.101**	0.096**
	(0.041)	*(0.036)*	*(0.036)*
Topic FE	Yes	Yes	Yes
State-Legislative Period FE	No	Yes	Yes
Party, Year FE	No	Yes	Yes
Controls	No	No	Yes
R^2	0.139	0.263	0.550
Observations	72,784	72,784	72,784

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include a dummy for whether the speeches occurred after the PISA shock and topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till December 2004. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A5.6: Different Number of Topics

	Baseline	Number of Topics				
		9	10	11	13	15
	(1)	(2)	(3)	(4)	(5)	(6)
PISA shock × Education	0.088**	0.090**	0.077**	0.083**	0.071*	0.063*
	(0.038)	*(0.037)*	*(0.036)*	*(0.037)*	*(0.035)*	*(0.035)*
Topic, State-Legisl. Period, Party,	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.535	0.559	0.561	0.555	0.553	0.546
Observations	137,820	138,274	138,124	137,850	137,427	137,238

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. In Column 1, I report the baseline estimate using the CTM estimated with 30 topics, aggregated into 11 topics. In Column 2, 3, 4, 5 and 6, I report the estimate using the CTM estimated using 9, 10, 11, 13, and 15 topics, respectively. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include a dummy for whether the speeches occurred after the PISA shock and topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A5.7: Symmetric Time Window (2000-2004) with Education and Placebo Topics (Local Politics and Social Welfare/Healthcare)

	Education (1)	Local Politics (2)	Social Wel- (3)
PISA shock × Topic Dummy	0.096**	0.042	-0.014
	(0.036)	*(0.064)*	*(0.041)*
Topic, State-Legisl. Period, Party, Year FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes
R^2	72,784	72,784	72,784
Observations	137,820	137,820	137,820

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of speeches about education (Column 1), about local politics (Column 2), and about social welfare/healthcare (Column 3). The dependent variable is the standardized polarization with CDU/CSU as the benchmark party. All regressions include a dummy for whether the speeches occurred after the PISA shock and topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till December 2004. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A5.8: Sensitivity to Different Thresholds of Term Frequency

	Lower Bound of Term Frequency					
	> 20	> 30	> 40	> 2%	> 2.5%	> 5%
	(1)	(2)	(3)	(4)	(5)	(6)
PISA shock × Education	0.088**	0.094**	0.097**	0.083**	0.080**	0.063**
	(0.039)	*(0.038)*	*(0.037)*	*(0.036)*	*(0.037)*	*(0.029)*
Topic, State-Legisl. Period, Party, Year	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.536	0.537	0.537	0.470	0.459	0.417
Observations	137,820	137,820	137,820	137,820	137,820	137,820

Note: Difference-in-differences estimate of the impact of the PISA shock on the polarization of education speeches. The dependent variable is the standardized polarization with CDU/CSU as the benchmark party computed using only words that appear in at least 20 speeches in a topic (Column 1), at least 30 speeches (Column 2), 40 (Column 3), 2% (Column 4), 2.5% (Column 5) and 5% (Column 6). All regressions include a dummy for whether the speeches occurred after the PISA shock, topic, state-legislative period, party, and year fixed effects. Controls include the length of a speech, the shares obtained at the latest state election by the two main parties, CDU/CSU and SPD, and a dummy variable for whether a speech is given by a minister or a state secretary, and if the MPs belongs to the governing coalition, and the distance from the next election. The data include all parliamentary debates from January 2000 till August 2008. Standard errors (in parentheses) have been clustered at the state level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Appendix B: Corpus Collection

The main data source for this project consists of parliamentary debates from the 16 German state parliaments for the period 2000-2008. The transcripts of such debates are not available in a structured format. They can be retrieved from each state-parliaments' website as PDF documents.[1] I have obtained these documents by web scraping each state-parliaments' website. Each document contains the transcript of a single plenary debate. Most transcripts of debates held in the period 2000-2008 are available as machine-readable PDF documents (93%). The remaining transcripts (7%) need to be first converted into machine-encoded text through an Optical Character Recognition (OCR) software. The share of documents for which OCR is necessary increases dramatically for debates held before 2000. This step is error-prone and renders the parsing of documents less reliable. Further, the availability of these documents on state-parliaments' websites decreases for debates held before 2000. For these reasons, I limit my analysis to debates from January 2000. In total, I have collected 3,302 PDF documents, 206.4 per state on average. In Figure B5.1, I report an example of a page from a plenary debate in the state of Baden-Württemberg that occurred on the 13th of December 2001. The raw text is clearly readable, but it lacks a formal structure. For my analysis, it is necessary to systematically identify and process the different features of the document, such as the name and role of the speaker, the party to which she belongs, the speech, the interruptions, the header, the page number etc.

In the example, the first speaker is Ms. Renate Rastätter, written in bold, a member of the parliament, as denoted by the abbreviation *Abg.* (*Abgeordnete*, member of parliament in German), of the Green Party (*GRÜNE*). The speech starts directly thereafter, and interruptions are reported in parentheses and are indented. At the end of the page, the President of the session and a Minister also speak. It can be noted how the way speaker names are reported also depends on their role.

These features constitute only some of the challenges that need to be addressed. The process to transform the PDF documents into a structured dataset suitable for my

[1] The only exception is Saarland, which does not provide the transcript of the parliamentary debates for the period 2000-2008 on its website. These debates were made available for my research upon request to the administration of the parliament.

analysis involves four main steps, which I now briefly describe.[2] Each step described has to be run separately for each state, as the process needs to be adapted to the different structure of the PDF documents in each state. For example, the way speakers and interruptions are reported in the transcripts differ substantially across states.

B1. – Layout scan.

The aim of the first step of the pipeline is to identify the coordinates that identify the location of the different elements in the document. This allows me to process the different features of the text correctly. To this purpose, I scan the layout of all the PDFs in a state using the Python package *layout_collector*. During this process, the coordinates of each text box that contains any content in each page of the documents are recorded. After all the coordinates have been recorded, I analyze the distribution of the coordinates of the text boxes to infer the relevant coordinates for the main text and interruptions in the left and right column, the header and the footer. I record such coordinates, which will be used in the next steps.

B2. – Conversion into XML files

All the PDFs are converted into XML files using the Python package *pdf2txt*. The XML version of each PDF stores information for each element in the PDF files, such as position, font, font size, boldness etc.

B3. – Conversion into plain text files

In this step, I convert the XML documents into plain text files enriched by features of the original PDF document recorded in the XML version. I use the coordinates for the headers and page numbers recorded in step B1 to drop the headers and page numbers while reconstructing the text files. Similarly, I use the coordinates recorded for the main text to ensure that the text on the left column of each page precedes the text on the right column. I also insert tags to denote interruptions and words in bold. Figure B5.2 shows the outcome of this step for the page depicted in Figure B5.1. One can observe that the plain text file does not contain headers nor page numbers, and that the text in the left column precedes the text in the right column. The text files are also enriched with tags to identify interruptions and speakers. These are added every time the text is indented with respect to the column in which it is located or when it is

[2] The process builds upon the publicly available GitHub repository https://github.com/panoptikum/plenary_record_parser , which contains the codes used by Felix Idelberger to perform the same task for German state parliamentary debates for the period 2008-2018.

written in bold in the original PDF document (<*indentation*> and <*poi_begin*> followed by <*poi_end*>), respectively.

B4. – Parsing

In the last step, I process each plain text file line-by-line. The processing script contains an exhaustive collection of regular expressions which capture the features of the documents, like the name of the speaker, party, role etc. and processes them accordingly. The script collects the speeches of each speaker as well as interruptions. Each speech or interruption is assigned to the speaker who utters them or, in the case of interruptions, under which they occur. Together with the name of the speaker, a variety of metadata such as the party (if reported), role, date, state, legislative session etc. are also collected. All the speeches are then aggregated into a single dataset where each observation corresponds to either a speech or an interruption and all the corresponding metadata. For the purpose of my analysis, I drop all the interruptions and aggregate speeches as a single utterance. In Figure B5.2 below, for example, an observation would correspond to the entire speech by Ms. Renate Rastätter until the speech of the President, stripped of all the interruptions.

B5. – Aggregation

Finally, the processing script aggregates all the speeches into a unified corpus. The process occurs hierarchically. At the end of each plenary debate, all the speeches are stored into a dataset. Once all the plenary debates in a state have been processed, all the separate debates are aggregated into a single dataset, which consists of all the speeches uttered in the period 2000-2008 in a single state and the related metadata (speaker, party affiliation, role, date, state etc.). As a final step, all the speeches of all 16 states in the period 2000-2008 are aggregated into one corpus, which provides the main data source for this project.

Figure B5.1: Example of PDF Document

wie das jetzt bei vielen der Fall ist. Zum Beispiel fordert der CDU-Politiker Rüttgers jetzt einen Sprachtest für alle Dreijährigen. Oder ich nenne die Aussage des Rektors eines Gymnasiums in Nordbaden, der jetzt fordert: Wir brauchen mehr Pauken von Faktenwissen statt Orientierungswissen. Oder es gibt die Äußerung von Ministerpräsident Stoiber, der seine politische Forderung nach einem Einwanderungsstopp jetzt mit den schlechten Schulleistungen von ausländischen Kindern begründet. Das, meine Damen und Herren, betrachte ich als eine infame Instrumentalisierung

(Zuruf von der CDU)

der PISA-Studie für politische Zwecke.

(Beifall bei den Grünen und Abgeordneten der SPD)

Diese Äußerung wird auch wider besseres Wissen gemacht. Denn wenn die PISA-Studie eines gezeigt hat, dann dies, dass es in der Bundesrepublik Deutschland nicht gelingt, gerade ausländische Schüler in den vorschulischen Einrichtungen und in der Grundschule ausreichend zu fördern, sodass sie in unserem System bessere Bildungschancen haben.

(Zuruf von der CDU)

Auch Sie, Frau Ministerin Schavan, wissen sofort, worin die Ursachen der Misere liegen. Sie zählen wieder alle Rezepte auf, alle Ihre Reformprojekte: frühere Einschulung, achtjähriges Gymnasium, reformierte Oberstufe, Englisch an Grundschulen, Neuorientierung des Unterrichts. Gleichzeitig pflegen Sie die altbekannten Vorurteile: Gesamtschulen seien schlecht, „Kuschelecken" ersetzten nicht das nachhaltige Lernen.

(Beifall bei Abgeordneten der CDU – Zuruf des Abg. Oelmayer GRÜNE)

Das, meine Damen und Herren, halte ich für ein populistisches Ausspielen

(Unruhe)

von Leistung und Sich-wohl-Fühlen. Es ist eine Diskriminierung von Grundschullehrerinnen in unserem Land. Diese nehmen nämlich den Erziehungs- und Bildungsauftrag des Lehrplans der Grundschule ernst, der fordert, Grundschulen als Lern- und Lebensorte auszugestalten, als Orte, an denen sich Kinder auch wohl fühlen können.

(Beifall bei den Grünen und des Abg. Zeller SPD)

Dabei gibt es durchaus Reformprojekte, die wir unterstützen, Frau Ministerin: zum Beispiel die frühere Einschulung gekoppelt mit dem „Schulanfang auf neuen Wegen", die jahrgangsübergreifenden Klassen, bei denen das Prinzip gilt: differenzieren und fördern.

Wenn man sich allerdings anschaut, wie die Entwicklung verläuft, muss man sagen: Nachdem Sie sich bundesweit schulen haben dieses Projekt tatsächlich durchgeführt.

Ich vermisse somit die Bereitschaft, Frau Kultusministerin, innezuhalten und auch einmal kritisch zu fragen: Mache ich, machen wir in diesem Bundesland eigentlich alles richtig? Man darf nicht immer nur sagen: Wir sind Spitze, wir können alles.

(Beifall der Abg. Heike Dederer GRÜNE – Abg. Pfister FDP/DVP: Außer Hochdeutsch!)

Nachdenklichkeit, meine Damen und Herren, ist allerdings auch in diesem Hause, bei uns selbst, angesagt. Ich würde es für ein gutes Zeichen halten, wenn sich der Landtag entschließen könnte, eine Enquetekommission zum Thema „Weiterentwicklung von Schule und Unterricht" einzusetzen. Ich möchte daran erinnern – Frau Kollegin Rudolf hat dies ja bereits angesprochen –: Wir sind sehr zufrieden damit, dass wir in der letzten Legislaturperiode die Enquetekommission „Jugend – Arbeit – Zukunft" hatten. Sie hat genau zu der Erkenntnis geführt, dass ungefähr 20 % der Jugendlichen aus sozial benachteiligten Familien nicht mehr die Leistungen erbringen, die für eine berufliche Integration notwendig sind.

Die Jugendenquetekommission hat im Ergebnis ein Bündel von Maßnahmen empfohlen. Unter anderem hat sie die Regierungskoalition endlich davon überzeugt, wie dringend notwendig die Schulsozialarbeit in diesem Land ist. Sie hat dafür gesorgt, dass in Baden-Württemberg Jugendagenturen eingerichtet wurden, die den Jugendlichen helfen, den schwierigen Übergang von der Schule in den Beruf zu meistern.

Deshalb, sage ich, würde es uns gut anstehen, zunächst einmal genau hinzuschauen und zu klären: Wo liegen denn die Schwächen? Vor allem sollten wir aber klären: Welche ganz konkreten Handlungsperspektiven müssen aufgebaut werden, damit alle Jugendlichen, von den sozial benachteiligten bis zu den höchstbegabten, die Bildung bekommen und die Kompetenzen entwickeln können, die sie von ihren Potenzialen her mitbringen?

Ich bedanke mich.

(Beifall bei den Grünen und des Abg. Kaufmann SPD)

Präsident Straub: Das Wort erteile ich Frau Ministerin Schavan.

Ministerin für Kultus, Jugend und Sport Dr. Annette Schavan: Herr Präsident, meine sehr verehrten Damen und Herren! 1997 hat sich die Kultusministerkonferenz in Konstanz entschieden, künftig deutsche Schulen an internationalen Vergleichsstudien zu beteiligen. In den letzten zehn Tagen habe ich mich an diese Situation, an die damalige Sitzung und die Wochen und Monate danach erinnert, und ich habe mich übrigens auch an manche schul- und bildungspolitische Debatte der letzten Jahre in diesem Haus erinnert.

(Abg. Röhm CDU: Jetzt kommts!)

738

Figure B5.2: Plain Text Representation of PDF Document

<poi_begin>Abg. Renate Rastätter<poi_end> GRUNE: Herr Präsident, meine Damen und Herren! Die Ergebnisse der PISA-Studie sind in der Tat Grund zur Sorge und Anlass zum Handeln. Nicht hilfreich ist es allerdings, in Panik und Angst zu verfallen, wie das jetzt bei vielen der Fall ist. Zum Beispiel fordert der CDU-Politiker Rüttgers jetzt einen Sprachtest für alle Dreijährigen. Oder ich nenne die Aussage des Rektors ei-nes Gymnasiums in Nordbaden, der jetzt fordert: Wir brauchen mehr Pauken von Faktenwissen statt Orientierungs-wissen. Oder es gibt die Äußerung von Ministerpräsident Stoiber, der seine politische Forderung nach einem Ein-wanderungsstopp jetzt mit den schlechten Schulleistungen von ausländischen Kindern begründet. Das, meine Damen und Herren, betrachte ich als eine infame Instrumentalisierung

<indentation>(Zuruf von der CDU)

der PISA-Studie für politische Zwecke.

<indentation>(Beifall bei den Grünen und Abgeordneten der
<identation>SPD)

Diese Äußerung wird auch wider besseres Wissen ge-macht. Denn wenn die PISA-Studie eines gezeigt hat, dann dies, dass es in der Bundesrepublik Deutschland nicht ge-lingt, gerade ausländische Schüler in den vorschulischen Einrichtungen und in der Grundschule ausreichend zu för-dern, sodass sie in unserem System bessere Bildungschancen haben.

<indentation>(Zuruf von der CDU)

Auch Sie, Frau Ministerin Schavan, wissen sofort, worin die Ursachen der Misere liegen. Sie zählen wieder alle Rezepte auf, alle Ihre Reformprojekte: frühere Einschulung, achtjähriges Gymnasium, reformierte Oberstufe, Englisch an Grundschulen, Neuorientierung des Unterrichts. Gleich-zeitig pflegen Sie die altbekannten Vorurteile: Gesamtschulen seien schlecht, „Kuschelecken“ ersetzten nicht das nachhaltige Lernen.

<indentation>(Beifall bei Abgeordneten der CDU – Zuruf des
<indentation>Oelmayer GRÜNE)

Das, meine Damen und Herren, halte ich für ein populisti-sches Ausspielen

<indentation>(Unruhe)

von Leistung und Sich-wohl-Fühlen. Es ist eine Diskrimi-nierung von Grundschullehrerinnen in unserem Land. Diese nehmen nämlich den Erziehungs- und Bildungsauftrag des Lehrplans der Grundschule ernst, der fordert, Grundschulen als Lern- und Lebensorte auszugestalten, als Orte, an denen sich Kinder auch wohl fühlen können.

<indentation>(Beifall bei den Grünen und des Abg. Zeller SPD)

Dabei gibt es durchaus Reformprojekte, die wir unterstüt-zen, Frau Ministerin: zum Beispiel die frühere Einschulung gekoppelt mit dem „Schulanfang auf neuen Wegen“, die jahrgangsübergreifenden Klassen, bei denen das Prinzip gilt: differenzieren und fördern.

Wenn man sich allerdings anschaut, wie die Entwicklung verläuft, muss man sagen: Nachdem Sie sich bundesweit

damit profiliert haben, haben Sie das Interesse verloren. Das Reformprojekt dümpelt vor sich hin. Stattdessen hätte es ein Schlüsselprojekt für die Weiterentwicklung der Grundschule werden können. Nur 4 % der 2 500 Grundschulen haben dieses Projekt tatsächlich durchgeführt.

Ich vermisse somit die Bereitschaft, Frau Kultusministerin, innezuhalten und auch einmal kritisch zu fragen: Mache ich, machen wir in diesem Bundesland eigentlich alles richtig? Man darf nicht immer nur sagen: Wir sind Spitze, wir können alles.

<indentation>(Beifall der Abg. Heike Dederer GRÜNE – Abg.
<indentation>Pfister FDP/DVP: Außer Hochdeutsch!)

Nachdenklichkeit, meine Damen und Herren, ist allerdings auch in diesem Hause, bei uns selbst, angesagt. Ich würde es für ein gutes Zeichen halten, wenn sich der Landtag ent-schließen könnte, eine Enquetekommission zum Thema „Weiterentwicklung von Schule und Unterricht“ einzuset-zen. Ich möchte daran erinnern – Frau Kollegin Rudolf hat dies ja bereits angesprochen –: Wir sind sehr zufrieden da-mit, dass wir in der letzten Legislaturperiode die Enquetekommission „Jugend – Arbeit – Zukunft“ hatten. Sie hat genau zu der Erkenntnis geführt, dass ungefähr 20 % der Jugendlichen aus sozial benachteiligten Familien nicht mehr die Leistungen erbringen, die für eine berufliche Integration notwendig sind. Die Jugendenquetekommission hat im Ergebnis ein Bündel von Maßnahmen empfohlen.
Unter anderem hat sie die Re-gierungskoalition endlich davon überzeugt, wie dringend
notwendig die Schulsozialarbeit in diesem Land ist. Sie hat dafür gesorgt, dass in Baden-Württemberg Jugendagenturen eingerichtet wurden, die den Jugendlichen helfen, den schwierigen Übergang von der Schule in den Beruf zu meistern. Deshalb, sage ich, würde es uns gut anstehen, zunächst ein-
mal genau hinzuschauen und zu klären: Wo liegen denn die Schwächen? Vor allem sollten wir aber klären: Welche ganz konkreten Handlungsperspektiven müssen aufgebaut werden, damit alle Jugendlichen, von den sozial benachteiligten bis zu den höchstbegabten, die Bildung bekommen und die Kompetenzen entwickeln können, die sie von ihren Potenzialen her mitbringen? Ich bedanke mich.

<indentation>(Beifall bei den Grünen und des Abg. Kaufmann
<indentation>SPD)

<poi_begin>Präsident Straub:<poi_end> Das Wort erteile ich Frau Ministerin Schavan.

<poi_begin>Ministerin für Kultus, Jugend und Sport Dr. Annette Schavan:<poi_end> Herr Präsident, meine sehr verehrten Damen und Herren! 1997 hat sich die Kultusministerkonferenz in Kon-stanz entschieden, künftig deutsche Schulen an internationalen Vergleichsstudien zu beteiligen. In den letzten zehn Tagen habe ich mich an diese Situation, an die damalige Sitzung und die Wochen und Monate danach erinnert, und ich habe mich übrigens auch an manche schul- und bildungspolitische Debatte der letzten Jahre in diesem Haus erinnert.

<indentation>(Abg. Röhm CDU: Jetzt kommts!)

Appendix C: Topic Classification

Topic classification is crucial for my analysis. I have used both supervised and unsupervised machine learning methods to achieve this task. The rationale for combining these methods is to obtain a reliable classification of the topics in the corpus at a relatively low cost. Supervised machine learning methods allow the researcher to have more control over the classification task but require manual labelling of a subset of the data, which is labor- and time-intensive. I have adopted this method to classify the most important topic for my analysis: education. Conversely, unsupervised machine learning methods for topic classification have the advantage that they do not require any manual labelling and do not require the researcher to know all the topics of the corpus in advance, but are harder to interpret. I have used this method to classify the topics of all the speeches that I have classified as not being about education. In the following sections, I provide a brief description of the classification task separately for the supervised and unsupervised machine learning methods.

C1. – Classification of Education Topic

Supervised machine learning (SML) methods require labelled data to learn the relationship between the outcome of interest and the available explanatory variables. I have therefore instructed two research assistants to manually classify 48 plenary sessions for a total of 3,346 speeches. The sessions were picked randomly from each state to ensure representativeness of the labelled dataset. The selection of sessions was slightly adjusted to favor sessions that discussed education topics. Specifically, randomly selected sessions were discarded if the word “school” did not appear in the entire session. It is important to remind that plenary sessions tend to be quite lengthy and deal with plenty of issues. Thus, favoring the sessions in which education topics are discussed does not prevent other topics from being adequately represented. The research assistants classified each speech in a binary way: whether it is about education or not. For the purpose of my analysis, I instructed the research assistants to consider speeches as being about education if they concern any education-related topic at the elementary, primary and secondary school level. Higher education was not considered part of the education theme, as it has a different legal basis and tends to be mandated to different ministries. The research assistants classified 571 speeches about education, 17.1% of the total number of speeches classified, while the other 2,775 were classified as not being about education.

At this stage, I face a binary classification task. The aim is to learn the conditional expectation function $Y(X)$, where $Y \in \{0,1\}$ denotes a binary indicator of whether the speech is about education and X denotes a vector representation of the speech, that governs the relationship between the label and content of the speech. I will then use the estimation of such function to predict the label of the entire corpus. I transform the speeches into a vector representation in three steps. I first perform standard preprocessing steps such as removal of stopwords, punctuation and numbers. Then, I apply a term frequency-inverse document frequency ($tf-idf$) transformation of the entire corpus of speeches. In this case, I apply a standard $tf-idf$ transformation of each speech according to the following formula:

$$tf-idf_d \equiv \frac{c_{dw}}{\sum_{k\in d} c_{dk}} \times \ln\left(\frac{D}{\sum_{n\in D} \mathbb{I}(c_{nw}>0)}\right), \tag{5.6}$$

Differently from the $tf-idf$ transformation used described in Section 5.3.5, in this case I do not perform a topic version of the $tf-idf$. The $tf-idf$ transformation of the documents upweights words that are specific to certain documents and downweights words that occur in many documents. I exclude words that occur in more than 50% of the documents and in less than 1% of the documents, as these terms are either too common or too rare. Given the limited size of the labelled dataset, I perform a further step to reduce the dimensionality of the explanatory variables. I implement the topic modelling algorithm Latent Dirichlet Analysis (LDA). LDA is a machine learning algorithm that identifies topics in corpora of texts in an unsupervised way based on the frequency with which words co-occur together. The crucial input parameter for this algorithm is the number of topics, which is unknown to the researcher and affects the broadness and interpretability of the topics. In this specific application, where I use LDA as a step for dimensionality-reduction purposes, I chose 15 topics as this was the number of topics that provided the highest accuracy in the classification task.

I then split the manually labelled sample into a train (80%) and test (20%) sample, stratifying by the binary outcome to ensure that both the train and test sample contain an equal share or speeches about education. I use a logistic regression classifier with 5-fold cross-validation and tune the hyper-parameters of the classifier using grid search over the type of penalty and strength of the regularization.[3] The best estimator is a logistic regression with an L1 type of penalty and a regularization hyper-parameter equal to 100. I report the evaluation metrics of the classification task in

[3] I have also tried other classifiers, such as Random Forest, Lasso, XGBoost and Gradient Boosting and achieved equivalent results in the classification task.

Table C5.1. Overall, figures show that the logistic regression achieves very good results in the classification of education vs non-education speeches. The F1 score, a metric which combines the precision and recall of the classifier, is close to 1, the maximum value for such metric.

I then use the machine learning model trained on the labeled dataset to make out-of-sample predictions for the entire corpus. I classify 18,701 speeches as being about education, or 8.9% of the speeches. I provide further descriptive evidence to corroborate the reliability of the classification task. In Table C5.2, I report the average number of times a set of words typical of the education context are mentioned in a speech. I select the terms "school", "teacher", "education", and "lesson". On average, the term "school" in Column 1 is mentioned 8.29 times in speeches classified as being about education, while only 0.55 in non-education speeches. Similarly, the terms "teacher", "education" and "lesson" are mentioned much more often in education speeches.

C2. – Classification of Other Topics

In cases where at least some of the topics in the corpus are unknown to the researcher and the corpus has not been manually labelled, topic models offer a fast and cheap solution to the classification task. Topic models are latent variable models that exploit the correlations among the words and latent semantic themes. For the classification task in this paper I apply the correlated topic model (Blei and Lafferty 2007)[4] to the corpus after the standard pre-processing steps described in the previous section. The correlated topic model (CTM) has the advantage over the more common Latent Dirichlet allocation (LDA) topic model of explicitly modeling the correlation between the latent topics in the corpus. I take advantage of this feature to aggregate similar topics, as topics do not be overly narrow to compute the polarization measure.

Like all topic models, the key tuning parameter of the CTM is the number of latent topics *K.* The outcome of the CTM depends largely on this parameter, which is mainly set depending on the size of the corpus, prior knowledge of the researcher about the corpus, and the downstream task the researcher wants to achieve. A low *K* might induce the CTM to aggregate unrelated topics, whereas a large *K* might split a single topic into excessively narrow sub-topics. In my setting, the corpus is relatively large, but I am not interested in narrowly defined topics. Ideally, topics should be of similar

[4] I used the R package *stm,* which enables a fast implementation of the correlated topic model (for further details, see Roberts, Stewart, and Tingley (2019))

size to the education topic classified using the SML algorithm in the previous step, which is around 9% of the corpus.

Once the CTM with K topics is estimated, the researcher needs to assess the outcome of the model by manual inspection of the identified topics and the provided metrics. A CTM with 30 topics provided good outcomes in terms of interpretability of the topics. I report the estimated topics in Table C5.3, with the 5 most relevant words for each topic and the manually assigned label. I aggregate topics which are either semantically similar or display a high correlation and identify 11 distinct topics. This step ensures that the size of the topics is similar to the education topic. I provide a graphical representation of the correlation among topics in Figure C5.1. Further, the heatmap places correlated topics next to each other and clusters of topics can be identified by looking at the dendrogram built on top of the heatmap. For example, the heatmap places topic 16 and 17 next to each other, as they display a high level of correlation. These topics concern discussion about housing and infrastructures, as it can be inferred from the most representative words reported in Table C5.3. Thus, I aggregate these topics into a single topic labelled "Housing, Infrastructure and Transportation". One topic, labelled "Other", does not have a clear interpretation and is not highly correlated with any of the other topics.

In principle, I could avoid aggregating topics while still obtaining large enough topics by using lower *K*, as the average size of the topics strictly decreases in *K.* However, using lower *K* led to worse performance of the CTM in terms of interpretability of the topics. Further, the final size of the topics was largely heterogenous, with some very small-sized topics and some relatively large topics. Nonetheless, it is reassuring that the number of topics *K* set at this stage does not affect my analysis in a substantial way, as I show in Table A5.6 in Section 5.5.5.

For each speech in the corpus, the CTM provides the estimated weight of each latent topic, with weights summing up to one. I assign the topic with the largest weight to each speech, which allows me to obtain a categorical classification of the corpus.

As a final step, I combine the estimation of the education topic described in the previous section and the topics obtained with the CTM. I assign the topic "education" to all the speeches that the SML algorithm classifies as being about education. For the remaining speeches, I assign the aggregated topics reported in Table C5.3. As expected, the CTM also identifies two topics that are clearly about education, namely

topic 9 and topic 18, since I applied the CTM to the entire corpus.[5] However, I do not assign the education topic identified by the CTM to any speech, as this *de facto* contradicts the more reliable SML classification, which predicted such speeches as not being about education. Given the high correlation between the SML classification of speeches about education and the CTM education topics (see footnote 5), such conflicts are rare: they concern less than 1% of the speeches. In these cases, I assign the second largest topic to these speeches instead. I report the share of aggregated topics for the main analysis in Figure A5.1. It can be noted that the size of the education topic lies at the median of the distribution.

[5] Alternatively, I could apply the CTM only to those speeches classified as not being about education by the SML algorithm. However, doing so did not resolve the issue in this case, as often the CTM identified an education-related topic nonetheless. Note that this might not only be due to the SML classification being imperfect, but also to the fact that speeches often touch upon different topics. For the purpose of my analysis, the SML only classifies speeches in a binary way, whereas in reality the distinction is fuzzier than such classification might suggest. Thus, speeches classified as not being about education might still reference to education-related issues. A further reason to apply the CTM to the entire corpus is that it gives me the opportunity to compare the SML classification of the education topic with the CTM classification of education topics. Reassuringly, the SML classification of the education topic is highly correlated with the education topics identified by the CTM (.71, p-value < .01). This increases the credibility of my classification task.

Table C5.1: Confusion Matrix – Classification of Education Speeches

	Precision	Recall	F1 Score	Support
	(1)	(2)	(3)	(4)
Non-Education	0.96	0.98	0.97	556
Education	0.88	0.81	0.84	114
Weighted Average / Total	0.95	0.95	0.95	670

Note: Confusion matrix of the classification of non-education and education speeches using a Logistic Classifier. In Column 1, I report the precision rate of the classification task, in Column 2 the recall rate, in Column 3 the F1 score. In Column 4, I report the sample size of both categories and the total sample size of the test sample.

Table C5.2: Average Frequency of Education Terms

	School	Teacher	Education	Lesson
	(1)	(2)	(3)	(4)
Non-Education	0.55	0.08	0.72	0.06
Education	8.29	2.82	3.95	1.77

Note: Average frequency of the terms "school" (Column 1), "teacher" (Column 2), "education" (Column 3), and "lesson" (Column 4) by speeches classified as being *not* about education and speeches about education in the entire corpus (210,006 speeches). The original German terms searched are "Schule", "Lehrer", "Bildung", and "Unterricht".

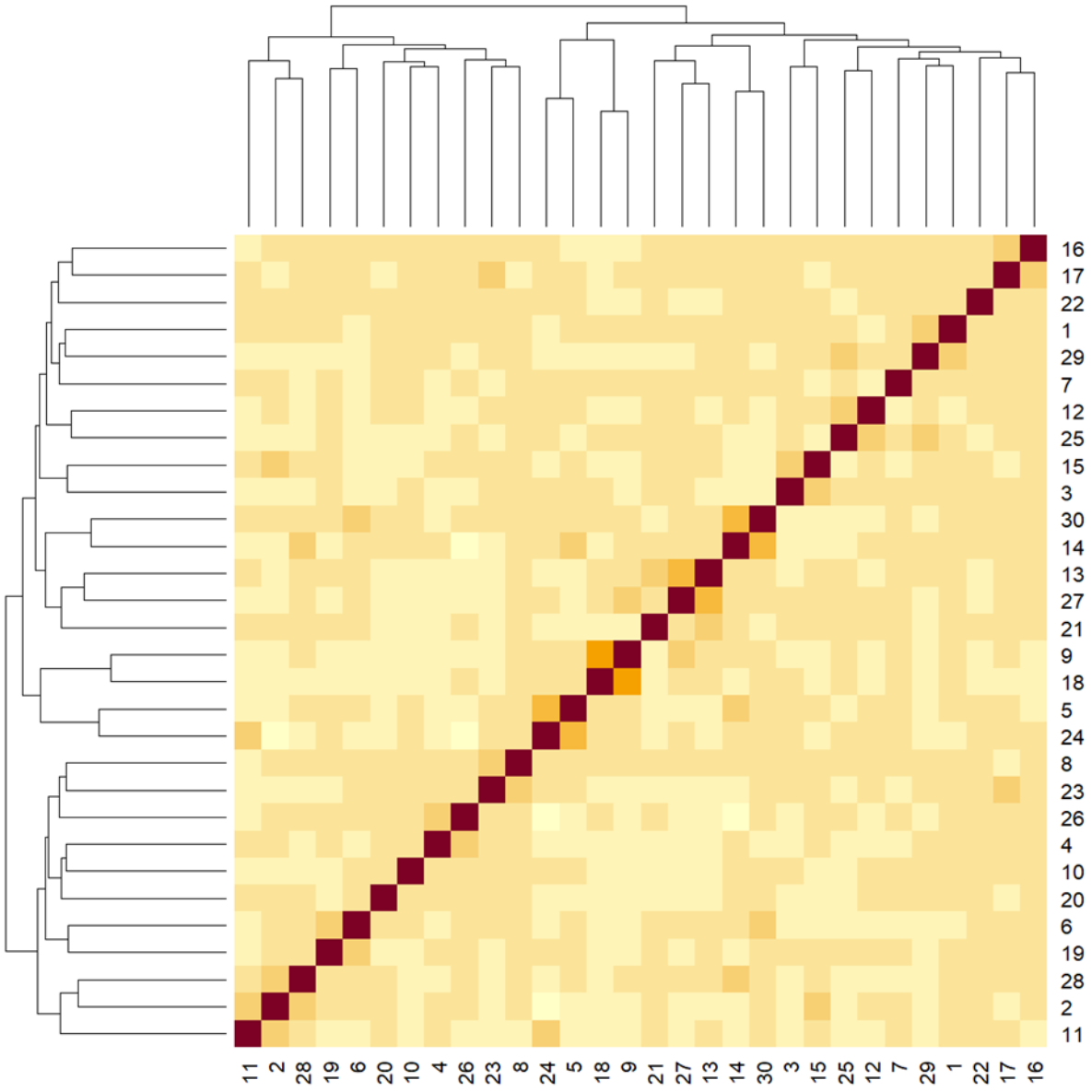
Table C5.3: Topic Labels and Top 5 FREX Words

Topic Number	Example FREX Words	Topic
1	museums, culture foundation, cultural policy, orchestra, theater	Higher Education, Culture
2	resolution recommendation, bill, bill, bill, legal	Lawmaking Process
3	supplementary budget, debt, net neuverschuldung, net borrowing, debts	Economy, Labor Market, Public Finance
4	national park, waters, water framework directive, national parks, forest owner	Agriculture, Environment, Energy
5	koch, committee of inquiry, roland, prime minister, investigative committee	Government Operations
6	preventive detention, rehabilitation, juvenile detention, prisoners, jva	Security, Crime, Defense
7	(university) students, courses, study places, higher education law	Higher Education, Culture
8	state government, called, weiss, measures, big	Government Operations
9	religious education, Pisa study, pisa, integration policy, comportment grades	Education
10	energies, renewable, renewable, nuclear energy, power plants	Agriculture, Environment, Energy
11	application, to discuss, application, think, points	Lawmaking Process
12	minimum wage, minimum wages, wages, collective agreements, moonlighting	Economy, Labor Market, Public Finance
13	gender, mainstreaming, men, women politics, women	Social Welfare, Healthcare, Equality
14	people's union, war, soldiers, sed, ddr	Security, Crime, Defense
15	municipal, municipal, circle-free, municipal financial reform, connection principle	Local Politics
16	renter, urban redevelopment, housing company, housing industry, housing market	Housing, Infrastructure, Transportation
17	temple courtyard, bust, schönefeld, bbi, lb	Housing, Infrastructure, Transportation
18	school board, school boards, class failure, principal, student numbers	Education
19	civil servants, officials, district courts, Officer, fire brigades	Security, Crime, Defense
20	agricultural policy, genetically, bse, modulation, animal meal	Agriculture, Environment, Energy
21	care insurance, insured, dependent, patients, statutory health insurance	Social Welfare, Healthcare, Equality
22	savings banks, broadcasting, broadcasting amendment state treaty, ard, state banks	Economy, Labor Market, Public Finance
23	federal traffic route plan, road charge, freight transport, passengers, long-distance	Housing, Infrastructure, Transportation
24	tell, rode, talked, outside, people	Government Operations
25	apprenticeship places, apprenticeship fee, labor market policy, long-term unemployment	Economy, Labor Market, Public Finance
26	answer, answer, exam, answered, documents	Other
27	child poverty, child protection, day care centers, family policy, childminders	Social Welfare, Healthcare, Equality
28	hardship commission, petition committee, petitioners, petition commission, non smoking pro-	Security, Crime, Defense
29	economic promotion, dockyard, structural change, north hesse, biotechnology	Economy, Labor Market, Public Finance
30	right-wing extremism, extremism, far right, right-wing extremist, right-wing extremists	Security, Crime, Defense

Note: List of 30 topics identified by the CTM model. For each topic, I report the five most representative words according to the Frequency-Exclusivity (FREX, see Roberts, Stewart, and Tingley (2019)) metric and the manually assigned topic. The FREX words have been translated into English using the *deep translator* package in Phython.

Determinants and Consequences of Student Test Scores 219

Figure C5.1: Ordered Heatmap of Topic Correlation



Note: The ordered heatmap depicts correlation between each topic and all the other topics. Topics are reordered using a clustering algorithm which arranges topics by similarity, thus placing more correlated topics next to each other. The overlayed dendrogram arranges cluster of topics by their correlation with each other. Darker colors indicate higher correlation.

6 References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. “Teachers and Student Achievement in the Chicago Public High Schools.” Journal of Labor Economics 25 (1): 95–135. https://doi.org/10.1086/508733.

Aguiar, Luis, Christian Peukert, Maximilian Schäfer, and Hannes Ullrich. 2022. “Facebook Shadow Profiles.” CESifo Working Paper.

Alan, Sule, and Seda Ertac. 2018. “Fostering Patience in the Classroom: Results from Randomized Educational Intervention.” *Journal of Political Economy* 126 (5): 1865–1911.

Alesina, Alberto, and Paola Giuliano. 2014. “Family Ties.” Vol. 2, 177–215. Handbook of Economic Growth: Elsevier.

Alesina, Alberto, and Paola Giuliano. 2015. “Culture and Institutions. Journal of Economic Literature 53 (4): 898-944.” *Journal of Economic Literature* 53 (4): 898–944.

Altonji, Joseph G., and Charles R. Pierret. 2001. “Employer Learning and Statistical Discrimination.” *The Quarterly Journal of Economics* 116 (1): 313–50.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. “Maimonides' Rule Redux.” *American Economic Review: Insights* 1 (3): 309–24. https://doi.org/10.1257/aeri.20180120.

Artelt, Cordula, Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, and Manfred Weiss, eds. 2002. *PISA 2000 – Die Länder Der Bundesrepublik Deutschland Im Vergleich:* Leske + Budrich, Opladen. Accessed Accessed February 18, 2022. www.grundschulpaedagogik.uni-bremen.de/archiv/pisa/PISA2000Lesen/PISA-E.pdf.

References

Ash, Elliott, Massimo Morelli, and Richard van Weelden. 2017. “Elections and Divisiveness: Theory and Evidence.” *The Journal of Politics* 79 (4): 1268–85. https://doi.org/10.1086/692587.

Ashenfelter, Orley, and John Ham. 1979. “Education, Unemployment, and Earnings.” *Journal of Political Economy* 87 (5, Part 2): S99-S116. https://doi.org/10.1086/260824.

Athey, Susan, and Guido W. Imbens. 2019. “Machine Learning Methods That Economists Should Know About.” *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Autor, David, David Dorn, Gordon Hanson, and Kaveh Majlesi. 2020. “Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure.” *American Economic Review* 110 (10): 3139–83. https://doi.org/10.1257/aer.20170011.

Bailey, Michael, Drew Johnston, Martin Koenen, Theresa Kuchler, Dominic Russel, and Johannes Stroebel. 2022. “The Social Integration of International Migrants: Evidence from the Networks of Syrians in Germany.” CESifo Working Paper.

Bastian, Kevin C., and C. Kevin Fortner. 2020. “Is Less More? Subject-Area Specialization and Outcomes in Elementary Schools.” *Education Finance and Policy* 15 (2): 357–82. https://doi.org/10.1162/edfp_a_00278.

Becker, Gary S. 1962. “Investment in Human Capital: A Theoretical Analysis.” *Journal of Political Economy* 70 (5, Part 2): 9–49. https://doi.org/10.1086/258724.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education.* New York: National Bureau of Economic Research.

Bergbauer, Annika B., Eric A. Hanushek, and Ludger Woessmann. 2021. “Testing.” *Journal of Human Resources*, 0520-10886R1. https://doi.org/10.3368/jhr.0520-10886R1.

Berger, Lara M. 2022. “How Digital Media Markets Amplify News Sentiment.”.

Bieber, Tonia, and Kerstin Martens. 2011. “The OECD PISA Study as a Soft Power in Education? Lessons from Switzerland and the US.” *European Journal of Education* 46 (1): 101–16.

Bietenbeck, Jan. 2014. “Teaching Practices and Cognitive Skills.” *Labour Economics* 30 (3): 143–53. https://doi.org/10.1016/j.labeco.2014.03.002.

Bietenbeck, Jan, and Matthew Collins. 2023. “New Evidence on the Importance of Instruction Time for Student Achievement on International Assessments.” *Journal of Applied Econometrics.* https://doi.org/10.1002/jae.2957.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2018. “Africa’s Skill Tragedy.” *Journal of Human Resources* 53 (3): 553–78. https://doi.org/10.3368/jhr.53.3.0616-8002R1.

Binder, Sarah A. 2004. *Stalemate: Causes and Consequences of Legislative Gridlock:* Brookings Institution Press.

Bird, Edward J. 2001. “Does the Welfare State Induce Risk-Taking?” *Journal of Public Economics* 80 (3): 357–83.

Bisin, Alberto, and Thierry Verdier. 2011. “The Economics of Cultural Transmission and Socialization.” Vol. 1, 339–416. Handbook of Social Economics: Elsevier.

Blei, David M., and John D. Lafferty. 2007. “A Correlated Topic Model of Science.” *The Annals of Applied Statistics* 1 (1). https://doi.org/10.1214/07-AOAS114.

Bleich, Erik, and A. Maurits van der Veen. 2021. “Media Portrayals of Muslims: A Comparative Sentiment Analysis of American Newspapers, 1996–2015.” *Politics, Groups, and Identities* 9 (1): 20–39. https://doi.org/10.1080/21565503.2018.1531770.

Blömeke, Sigrid, Gabriele Kaiser, and Rainer Lehmann, eds. 2010. *TEDS–M 2008: Professionelle Kompetenz Und Lerngelegenheiten Angehender*

Mathematiklehrkräfte Für Die Sekundarstufe I Im Internationalen Vergleich [Cross-National Comparison of the Professional Competency of and Learning Opportunities for Future Secondary School Teachers of Mathematics]. Münster: Waxmann.

Bonica, Adam. 2013. "Ideology and Interests in the Political Marketplace." *American Journal of Political Science* 57 (2): 294–311. https://doi.org/10.1111/ajps.12014.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2022. "Salience." *Annual Review of Economics* 14 (1): 521–44. https://doi.org/10.1146/annurev-economics-051520-011616.

Bos, Wilfried, and T. Neville Postlethwaite. 2002. "Internationale Schulleistungsforschung: Ihre Entwicklungen Und Folgen Für Die Deutsche Bildungslandschaft." In *Leistungsmessungen in Schulen.*, edited by Franz E. (H.) Weinert. 2nd: Weinheim: Beltz.

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. 2022. "Cross-Country Trends in Affective Polarization." *The Review of Economics and Statistics*, 1–60. https://doi.org/10.1162/rest_a_01160.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James Wyckoff. 2008. "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-poverty Schools." *Journal of Policy Analysis and Management* 27 (4): 793–818. https://doi.org/10.1002/pam.20377.

Breakspear, Simon. 2012. "The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance." *OECD Education Working Papers.* https://doi.org/10.1787/5k9fdfqffr28-en.

Buddin, Richard, and Gema Zamarro. 2009. "Teacher Qualifications and Student Achievement in Urban Elementary Schools." *Journal of Urban Economics* 66 (2): 103–15. https://doi.org/10.1016/j.jue.2009.05.001.

Burroughs, Nathan, Jacqueline Gardner, Youngjun Lee, Siwen Guo, Israel Touitou, Kimberly Jansen, and William Schmidt. 2019. "A Review of the Literature on

Teacher Effectiveness and Student Outcomes." In *Teaching for Excellence and Equity*. Vol. 6, edited by Nathan Burroughs, Jacqueline Gardner, Youngjun Lee, Siwen Guo, Israel Touitou, Kimberly Jansen, and William Schmidt, 7–17. IEA Research for Education. Cham: Springer International Publishing.

Cabañas, José González, Ángel Cuevas, and Rubén Cuevas. 2018. "Facebook Use of Sensitive Data for Advertising in Europe." http://arxiv.org/pdf/1802.05030v1.

Canen, Nathan, Chad Kendall, and Francesco Trebbi. 2020a. *Political Parties as Drivers of U.S. Polarization: 1927-2018.* Cambridge, MA: National Bureau of Economic Research.

Canen, Nathan, Chad Kendall, and Francesco Trebbi. 2020b. "Unbundling Polarization." *Econometrica* 88 (3): 1197–1233. https://doi.org/10.3982/ECTA16756.

Card, David. 1999. "The Causal Effect of Education on Earnings." Vol. 3, 1801–63. Handbook of Labor Economics: Elsevier.

Carnevale, Anthony P., Nicole Smith, and Michelle Melton. 2011. "STEM: Science Technology Engineering Mathematics. State-Level Analysis." *Tech. rep.* Georgetown University Center on Education and the Workforce. http://cew.georgetown.edu/stem/.

Castillo, Marco, Jeffrey L. Jordan, and Ragan Petrie. 2019. "Discount Rates of Children and High School Graduation." *The Economic Journal* 129 (619): 1153–81. https://doi.org/10.1111/ecoj.12574.

Cattaneo, Maria A., Chantal Oggenfuss, and Stefan C. Wolter. 2017. "The More, the Better? The Impact of Instructional Time on Student Performance." *Education Economics* 25 (5): 433–45. https://doi.org/10.1080/09645292.2017.1315055.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79. https://doi.org/10.1257/aer.104.9.2633.

References

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong et al. 2022. “Social Capital I: Measurement and Associations with Economic Mobility.” *Nature* 608 (7921): 108–21. https://doi.org/10.1038/s41586-022-04996-4.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. “Teacher-Student Matching and the Assessment of Teacher Effectiveness.” *Journal of Human Resources* XLI (4): 778–820. https://doi.org/10.3368/jhr.XLI.4.778.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. “Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects.” *Economics of Education Review* 26 (6): 673–82. https://doi.org/10.1016/j.econedurev.2007.10.002.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2010. “Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects.” *Journal of Human Resources* 45 (3): 655–81. http://www.jstor.org/stable/25703472.

Coenen, Johan, Ilja Cornelisz, Wim Groot, Henriette van den Maassen Brink, and Chris van Klaveren. 2018. “Teacher Characteristics and Their Effects on Student Test Scores: A Systematic Review.” *Journal of Economic Surveys* 32 (3): 848–77. https://doi.org/10.1111/joes.12210.

Croninger, Robert G., Jennifer King Rice, Amy Rathbun, and Masako Nishio. 2007. “Teacher Qualifications and Early Learning: Effects of Certification, Degree, and Experience on First-Grade Student Achievement.” *Economics of Education Review* 26 (3): 312–24. https://doi.org/10.1016/j.econedurev.2005.05.008.

Cutler, David, and Adriana Lleras-Muney. 2006. *Education and Health: Evaluating Theories and Evidence.* Cambridge, MA: National Bureau of Economic Research.

Davoli, Maddalena, and Horst Entorf. 2018. “The PISA Shock, Socioeconomic Inequality, and School Reforms in Germany.”.

Deaton, Angus S., and Christina Paxson. 2001. “Mortality, Education, Income, and Inequality Among American Cohorts.” In *Themes in the Economics of Aging*, edited by David A. Wise. A National Bureau of Economic Research project report. Chicag, London: University of Chicago Press.

Dee, Thomas S. 2005. “A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?” *American Economic Review* 95 (2): 158–65. https://doi.org/10.1257/000282805774670446.

Dee, Thomas S. 2007. “Teachers and the Gender Gaps in Student Achievement.” *Journal of Human Resources* XLII (3): 528–54. https://doi.org/10.3368/jhr.XLII.3.528.

Dee, Thomas S., and Brian A. Jacob. 2011. “The Impact of No Child Left Behind on Student Achievement. 30 (3): 418-446.” *Journal of Policy Analysis and Management* 30 (3): 418–46.

DellaVigna, S., and E. Kaplan. 2007. “The Fox News Effect: Media Bias and Voting.” *The Quarterly Journal of Economics* 122 (3): 1187–1234. https://doi.org/10.1162/qjec.122.3.1187.

Der Spiegel. 2001. “Sind Deutsche Schüler Doof?” *Der Spiegel*, December 13, 2001. https://www.spiegel.de/lebenundlernen/schule/die-pisa-analyse-sind-deutsche-schueler-doof-a-172357.html.

Dieterle, Steven G. 2015. “Class-Size Reduction Policies and the Quality of Entering Teachers.” *Labour Economics* 36:35–47. https://doi.org/10.1016/j.labeco.2015.07.005.

Downs, Anthony. 1957. *An Economic Theory of Democracy:* New York: Harper and Brothers.

Ebenrett, Heinz J., Dieter Hansen, and Klaus J. . Puzicha. 2003. “Verlust Von Humankapital in Regionen Mit Hoher Arbeitslosigkeit.” *Aus Politik und Zeitgeschichte* B6 (7).

References

Ertl, Hubert. 2006. “Educational Standards and the Changing Discourse on Education: The Reception and Consequences of the PISA Study in Germany.” *Oxford Review of Education* 32 (5): 619–34. http://www.jstor.org/stable/4618685.

European Commission: DG Employment Social Affairs and Inclusion. 2020. “European Skills Agenda for Sustainable Competitiveness, Social Fairness and Resilience.” News release. 2020. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2020:274:FIN.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. “Global Evidence on Economic Preferences.” *The Quarterly Journal of Economics* 133 (4): 1645–92. https://doi.org/10.1093/qje/qjy013.

Falorsi, Piero Demetrio, Roberto Ricci, and Patrizia Falzetti. 2019. *Le Metodologie Di Campionamento E Scomposizione Della Devianza Nelle Rilevazioni Nazionali Dell'invalsi: Le Rilevazioni Degli Apprendimenti A.S. 2018-2019.* Milano: Franco Angeli.

FAZ. 2001. “Katastrophales Zeugnis Für Deutsche Schüler.” *FAZ*, December 2, 2001.

Figlio, David, Paola Giuliano, Umut Özek, and Paola Sapienza. 2019. “Long-Term Orientation and Educational Performance.” *American Economic Journal: Economic Policy* 11 (4): 272–309. https://doi.org/10.1257/pol.20180374.

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

Fiva, Jon H., Oda Nedregård, and Henning Øien. 2022. “Polarization in Party-Centered Environments: Evidence from Parliamentary Debates.” Unpublished manuscript, last modified Accessed 3rd, February 2022. https://www.jon.fiva.no/docs/FivaNedregardOien.pdf.

Forschungsgruppe Wahlen, Mannheim. 2019. “Politbarometer - Gesamtkumulation.”.

Fryer, Roland G. 2018. “The “Pupil” Factory: Specialization and the Production of Human Capital in Schools.” *American Economic Review* 108 (3): 616–56. https://doi.org/10.1257/aer.20161495.

Funke, Manuel, Moritz Schularick, and Christoph Trebesch. 2016. “Going to Extremes: Politics After Financial Crises, 1870–2014.” *European Economic Review* 88 (3): 227–60. https://doi.org/10.1016/j.euroecorev.2016.03.006.

Galey-Horn, Sarah, Sarah Reckhow, Joseph J. Ferrare, and Lorien Jasny. 2020. “Building Consensus: Idea Brokerage in Teacher Policy Networks.” *American Educational Research Journal* 57 (2): 872–905. https://doi.org/10.3102/0002831219872738.

Galor, Oded, and Ömer Özak. 2016. “The Agricultural Origins of Time Preference.” *American Economic Review* 106 (10): 3064–3103. https://doi.org/10.1257/aer.20150020.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2019. “Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech.” *Econometrica* 87 (4): 1307–40. https://doi.org/10.3982/ECTA16566.

Giannetti, Daniela, and Michael Laver. 2005. “Policy Positions and Jobs in the Government.” *European Journal of Political Research* ` 44 (1): 91–120. https://doi.org/10.1111/j.1475-6765.2005.00220.x.

Global Education Monitoring Report Team. 2017. *Accountability in Education: Meeting Our Commitments; Global Education Monitoring Report, 2017/8:* Paris: UNESCO.

Goet, Niels D. 2019. “Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015.” *Political Analysis* 27 (4): 518–39. https://doi.org/10.1017/pan.2019.2.

Goldhaber, Dan, and Emily Anthony. 2007. “Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching.” *Review of Economics and Statistics* 89 (1): 134–50. https://doi.org/10.1162/rest.89.1.134.

Goldhaber, Dan D., and Dominic J. Brewer. 1997. “Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity.” *The Journal of Human Resources* 32 (3): 505. https://doi.org/10.2307/146181.

Goldhaber, Dan D., and Dominic J. Brewer. 2000. “Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement.” *Educational Evaluation and Policy Analysis* 22 (2): 129. https://doi.org/10.2307/1164392.

Goldin, Claudia. 2016. “Human Capital.” In *Handbook of Cliometrics*, edited by Claude Diebolt and Michael Haupert, 55–86. Heidelberg, Germany: Springer Verlag.

Golsteyn, Bart H.H., Hans Grönqvist, and Lena Lindahl. 2014. “Adolescent Time Preferences Predict Lifetime Outcomes.” *The Economic Journal* 124 (580): F739-F761. https://doi.org/10.1111/ecoj.12095.

Gong, Jie, Yi Lu, and Hong Song. 2018. “The Effect of Teacher Gender on Students’ Academic and Noncognitive Outcomes.” *Journal of Labor Economics* 36 (3): 743–78. https://doi.org/10.1086/696203.

Grek, Sotiria. 2009. “Governing by Numbers: The PISA ‘Effect’ in Europe.” *Journal of Education Policy* 24 (1): 23–37. https://doi.org/10.1080/02680930802412669.

Gruber, Karl H. 2006. “The German "PISA-Shock": Some Aspects of the Extraordinary Impact of the OECD's PISA Studyon the German Education System.” In *Cross-National Attraction in Education: Accounts from England and Germany*, edited by Hubert Ertl, 195–208: Symposium Books.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2004. “The Role of Social Capital in Financial Development.” *American Economic Review* 94 (3): 526–56.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2006. “Does Culture Affect Economic Outcomes?” *Journal of Economic Perspectives* 20 (2): 23–48. https://doi.org/10.1257/jep.20.2.23.

Hanushek, Eric A. 1971. “Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data.” *American Economic Review* 61 (2): 280–88.

Hanushek, Eric A. 1986. “The Economics of Schooling: Production and Efficiency in Public Schools.” *Journal of Economic Literature* 24 (3): 1141–77.

Hanushek, Eric A. 2016. “What Matters for Achievement: Updating Coleman on the Influence of Families and Schools. 16 (2): 22-30.” *Education Next* 16 (2): 22–30.

Hanushek, Eric A., Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann. 2022. “Patience, Risk-Taking, and Human Capital Investment Across Countries.” *The Economic Journal* 132 (646): 2290–2307. https://doi.org/10.1093/ej/ueab105.

Hanushek, Eric A., Marc Piopiunik, and Simon Wiederhold. 2019. “The Value of Smarter Teachers.” *Journal of Human Resources* 54 (4): 857–99. https://doi.org/10.3368/jhr.54.4.0317.8619R1.

Hanushek, Eric A., and Margaret E. Raymond. 2005. “Does School Accountability Lead to Improved Student Performance?” *Journal of Policy Analysis and Management* 24 (2): 297–327.

Hanushek, Eric A., and Steven G. Rivkin. 2004. “How to Improve the Supply of High-Quality Teachers.” *Brookings Papers on Education Policy*.

Hanushek, Eric A., Jens Ruhose, and Ludger Woessmann. 2017. “Knowledge Capital and Aggregate Income Differences: Development Accounting for U.S. States.” *American Economic Journal: Macroeconomics* 9 (4): 184–224.

Hanushek, Eric A., and Ludger Woessmann. 2006. “Does Educational Tracking Affect Performance and Inequality? Differences- In-Differences Evidence Across

Countries." *The Economic Journal* 116 (510): C63-C76. https://doi.org/10.1111/j.1468-0297.2006.01076.x.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46 (3): 607–68. https://doi.org/10.1257/jel.46.3.607.

Hanushek, Eric A., and Ludger Woessmann. 2012. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17 (4): 267–321. https://doi.org/10.1007/s10887-012-9081-x.

Hanushek, Eric A., and Ludger Woessmann. 2015. *Universal Basic Skills:* OECD.

Harris, Douglas N., and Tim R. Sass. 2011. "Teacher Training, Teacher Quality and Student Achievement." *Journal of Public Economics* 95 (7-8): 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103 (6): 2052–86. https://doi.org/10.1257/aer.103.6.2052.

Heckman, James J., Lance J. Lochner, and Petra E. Todd. 2006. "Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond." In *Handbook of the Economics of Education Volume 1*. Vol. 1, 307–458. Handbook of the Economics of Education: Elsevier.

Heckman, James J., and Rodrigo Pinto. 2015. "Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs." *Econometric reviews* 34 (1-2): 6–31. https://doi.org/10.1080/07474938.2014.944466.

Helbig, Marcel, and Rita Nikolai. 2015. *Die Unvergleichbaren: Der Wandel Der Schulsysteme in Den Deutschen Bundesländern Seit 1949.* Bad Heilbrunn: Verlag Julius Klinkhardt.

Henninges, Miriam, Claudia Traini, and Corinna Kleinert. 2019. “Tracking and Sorting in the German Educational System.” *Leibniz Institute for Educational Trajectories (LIfBi) Working Paper No* 83.

Hermes, Henning, Philipp Lergetporer, Frauke Peter, and Simon Wiederhold. 2021. “Behavioral Barriers and the Socioeconomic Gap in Child Care Enrollment.”.

Herzog, Alexander, and Kenneth Benoit. 2015. “The Most Unkindest Cuts: Speaker Selection and Expressed Government Dissent During Economic Crisis.” *The Journal of Politics* 77 (4): 1157–75. https://doi.org/10.1086/682670.

Hopmann, Stefan, Gertrude Brinek, and Martin Retzl. 2007. *PISA Zufolge PISA: Hält PISA, Was Es Verspricht? = PISA According to PISA : Does PISA Keep What It Promises?* Schulpädagogik und pädagogische Psychologie Bd. 6. Wien: Lit.

Howitt, Peter, and Philippe Aghion. 1998. “Capital Accumulation and Innovation as Complementary Factors in Long-Run Growth.” *Journal of Economic Growth* 3 (2): 111–30. https://doi.org/10.1023/A:1009769717601.

Ichino, Andrea, and Giovanni Maggi. 2000. “Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm.” *The Quarterly Journal of Economics* 115 (3): 1057–90.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. “The Origins and Consequences of Affective Polarization in the United States.” *Annual Review of Political Science* 22 (1): 129–46. https://doi.org/10.1146/annurev-polisci-051117-073034.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2016. “The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms.” *The Quarterly Journal of Economics* 131 (1): 157–218. https://doi.org/10.1093/qje/qjv036.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. “Teacher Effects and Teacher-Related Policies.” *Annual Review of Economics* 6 (1): 801–25. https://doi.org/10.1146/annurev-economics-080213-040845.

Jackson, C. Kirabo, Cora Wigger, and Heyu Xiong. 2021. “Do School Spending Cuts Matter? Evidence from the Great Recession.” *American Economic Journal: Economic Policy* 13 (2): 304–35. https://doi.org/10.1257/pol.20180674.

Jackson, Kirabo, and Alexey Makarin. 2018. “Can Online Off-the-Shelf Lessons Improve Student Outcomes? Evidence from a Field Experiment.” *American Economic Journal: Economic Policy* 10 (3): 226–54. https://doi.org/10.1257/pol.20170211.

Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Wilse-Samson Laurence. 2012. “Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech.” *Brookings Papers on Economic Activity*, 1–81. Accessed 9th, February 2022.

Jepsen, Christopher, and Steven G. Rivkin. 2009. “Class Size Reduction and Student Achievement.” *Journal of Human Resources* 44 (1): 223–50. https://doi.org/10.3368/jhr.44.1.223.

Jerrim, John, Luis Alejandro Lopez-Agudo, Oscar D. Marcenaro-Gutierrez, and Nikki Shure. 2017. “What Happens When Econometrics and Psychometrics Collide? An Example Using the PISA Data.” *Economics of Education Review* 61 (5): 51–58. https://doi.org/10.1016/j.econedurev.2017.09.007.

Jones, David R. 2001. “Party Polarization and Legislative Gridlock.” *Political Research Quarterly* 54 (1): 125–41. https://doi.org/10.1177/106591290105400107.

Jung, Dawoon, Tushar Bharati, and Seungwoo Chin. 2021. “Does Education Affect Time Preference? Evidence from Indonesia.” *Economic Development and Cultural Change* 69 (4): 1451–99. https://doi.org/10.1086/706496.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. “What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City.” *Economics of Education Review* 27 (6): 615–31. https://doi.org/10.1016/j.econedurev.2007.05.005.

Kayser, Mark A., and Michael Peress. 2021. “Does the Media Cover the Economy Accurately? An Analysis of Sixteen Developed Democracies.” *Quarterly Journal of Political Science* 16 (1): 1–33. https://doi.org/10.1561/100.00019098.

Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk. 2020. “The Formation of Prosociality: Causal Evidence on the Role of Social Environment.” *Journal of Political Economy* 128 (2): 434–67. https://doi.org/10.1086/704386.

Kuhlmann, Christian, and Klaus-Jürgen Tillmann. 2009. “Mehr Ganztagsschulen Als Konsequenz Aus PISA? Bildungspolitische Diskurse Und Entwicklungen in Den Jahren 2000 Bis 2003.” In *Ganztagsschule Als Symbolische Konstruktion*, edited by Fritz-Ulrich Kolbe, Sabine Reh, Till-Sebastian Idel, Bettina Fritzsche, and Kerstin Rabenstein, 23–45. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kukla-Acevedo, Sharon. 2009. “Do Teacher Characteristics Matter? New Results on the Effects of Teacher Preparation on Student Achievement.” *Economics of Education Review* 28 (1): 49–57. https://doi.org/10.1016/j.econedurev.2007.10.007.

Ladd, Helen F., and Lucy C. Sorensen. 2015. “Do Master's Degrees Matter? Advanced Degrees, Career Paths, and the Effectiveness of Teachers.” *National Center for Analysis of Longitudinal Data in Education Research (CALDER)* (Working Paper 136).

Lapinski, John S. 2008. “Policy Substance and Performance in American Lawmaking, 1877–1994.” *American Journal of Political Science* 52 (2): 235–51. https://doi.org/10.1111/j.1540-5907.2008.00310.x.

Lauderdale, Benjamin E., and Alexander Herzog. 2016. “Measuring Political Positions from Legislative Speech.” *Political Analysis* 24 (3): 374–94. https://doi.org/10.1093/pan/mpw017.

Lavy, Victor. 2015. “Do Differences in Schools’ Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries.” *The Economic Journal* 125 (588): F397-F424. https://doi.org/10.1111/ecoj.12233.

Lemieux, Thomas. 2006. “The “Mincer Equation” Thirty Years After Schooling, Experience, and Earnings.” In *Jacob Mincer a Pioneer of Modern Labor Economics*. Vol. 33, edited by Shoshana Grossbard, 127–45. Boston: Kluwer Academic Publishers.

Levendusky, Matthew S. 2013. “Why Do Partisan Media Polarize Viewers?” *American Journal of Political Science* 57 (3): 611–23. https://doi.org/10.1111/ajps.12008.

Lewandowsky, Marcel, Julia Schwanholz, Christoph Leonhardt, and Andreas Blätte. 2022. “New Parties, Populism, and Parliamentary Polarization: Evidence from Plenary Debates in the German Bundestag.” In *The Palgrave Handbook of Populism*. Vol. 36, edited by Michael Oswald, 611–27. Cham: Springer International Publishing.

Lim, Jaegeum, and Jonathan Meer. 2017. “The Impact of Teacher–Student Gender Matches.” *Journal of Human Resources* 52 (4): 979–97. https://doi.org/10.3368/jhr.52.4.1215-7585R1.

Lipset, Seymour M., and Stein Rokkan. 1967. *Party Systems and Voter Alignments: Cross-National Perspectives:* New York: The Free Press.

Lyon, Melissa Arnold, and Matthew A. Kraft. 2021. “Elevating Education in Politics: How Teacher Strikes Shape Congressional Election Campaigns.”.

Maltzman, Forrest, and Lee Sigelman. 1996. “The Politics of Talk: Unconstrained Floor Time in the U.S. House of Representatives.” *The Journal of Politics* 58 (3): 819–30. https://doi.org/10.2307/2960448.

Mankiw, N. Gregory, David Romer, and David Weil. 1992. “A Contribution to the Empirics of Economic Growth.” *The Quarterly Journal of Economics* 107 (2): 407–37. https://doi.org/10.2307/2118477.

Martens, Kerstin, and Dennis Niemann. 2013. “When Do Numbers Count? The Differential Impact of the PISA Rating and Ranking on Education Policy in Germany and the US.” *German Politics* 22 (3): 314–32. https://doi.org/10.1080/09644008.2013.794455.

Martin, M. O., I. V. S. Mullis, P. Foy, and M. Hooper. 2016. *TIMSS 2015 International Results in Science:* Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2015/international-results/.

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal. 2016. *Polarized America, Second Edition: The Dance of Ideology and Unequal Riches:* MIT Press.

McDonnell, Lorraine M., and M. Stephen Weatherford. 2013. “Organized Interests and the Common Core.” *Educational Researcher* 42 (9): 488–97. https://doi.org/10.3102/0013189X13512676.

Metzler, Johannes, and Ludger Woessmann. 2012. “The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation.” *Journal of Development Economics* 99 (2): 486–96. https://doi.org/10.1016/j.jdeveco.2012.06.002.

Mian, Atif, Amir Sufi, and Francesco Trebbi. 2014. “Resolving Debt Overhang: Political Constraints in the Aftermath of Financial Crises.” *American Economic Journal: Macroeconomics* 6 (2): 1–28. https://doi.org/10.1257/mac.6.2.1.

Mincer, Jacob. 1958. “Investment in Human Capital and Personal Income Distribution.” *Journal of Political Economy* 66 (4): 281–302. https://doi.org/10.1086/258055.

Mincer, Jacob. 1974. *Schooling, Experience, and Earnings.* New York: National Bureau of Economic Research.

Monk, David, and Jennifer King. 1994. “Multi-Level Teacher Resource Effects on Pupil Performance in Secondary Mathematics and Science: The Role of Teacher Subject Matter Preparation.” In *Contemporary Policy Issues: Choices and Consequences in Education*, edited by Ronald Ehrenberg, 29–58. Ithaca, NY: ILR.

Mullis, Ina V. S., and Michael O. Martin, eds. 2013. *TIMSS 2015 Assessment Frameworks.* Chestnut Hill MA: TIMSS & PIRLS International Study Center Lynch School of Education Boston College.

Mullis, Ina V. S., Michael O. Martin, and Tom Loveless, eds. 2016. *20 Years of TIMSS: International Trends in Mathematics and Science Achievement, Curriculum, and Instruction:* TIMSS & PIRLS International Study Center,Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, Mass. Ballinger Pub. Co.

Nagler, Markus, Marc Piopiunik, and Martin R. West. 2020. “Weak Markets, Strong Teachers: Recession at Career Start and Teacher Effectiveness.” *Journal of Labor Economics* 38 (2): 453–500. https://doi.org/10.1086/705883.

Neumann, Knut, Hans E. Fischer, and Alexander Kauertz. 2010. “From PISA to Educational Standards: The Impact of Large-Scale Assessments on Science Education in Germany.” *International Journal of Science and Mathematics Education* 8 (3): 545–63. https://doi.org/10.1007/s10763-010-9206-7.

Nickell, Stephen. 1979. “Education and Lifetime Patterns of Unemployment.” *Journal of Political Economy* 87 (5, Part 2): S117-S131. https://doi.org/10.1086/260825.

Obradovich, Nick, Ömer Özak, Ignacio Martín, Ignacio Ortuño-Ortín, Edmond Awad, Manuel Cebrián, Rubén Cuevas, Klaus Desmet, Iyad Rahwan, and Ángel Cuevas. 2022. “Expanding the Measurement of Culture with a Sample of Two Billion

Humans." *Journal of the Royal Society, Interface* 19 (190): 20220085. https://doi.org/10.1098/rsif.2022.0085.

OECD. 2011. *Lessons from PISA for the United States:* OECD.

OECD. 2014. *TALIS 2013 Results:* TALIS, OECD Publishing, Paris.

OECD. 2016a. "New Skills for the Digital Economy." *OECD Digital Economy Papers* 258. https://doi.org/10.1787/5jlwnkm2fc9x-en.

OECD. 2016b. *OECD Science, Technology and Innovation Outlook 2016:* OECD.

OECD. 2016c. *PISA 2015 Results (Volume I).* Paris: OECD.

OECD. 2017. *OECD Skills Outlook 2017: Skills and Global Value Chains.* Paris: OECD.

OECD. 2018. *Effective Teacher Policies.* PISA, OECD Publishing, Paris.

OECD. 2019. *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners.* TALIS, OECD Publishing, Paris.

OECD. 2022. *Education at a Glance 2022.* PISA, OECD Publishing, Paris: OECD.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204. https://doi.org/10.1080/07350015.2016.1227711.

Paredes, Valentina. 2014. "A Teacher Like Me or a Student Like Me? Role Model Versus Teacher Bias Effect." *Economics of Education Review* 39 (2): 38–49. https://doi.org/10.1016/j.econedurev.2013.12.001.

Peterson, Andrew, and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26 (1): 120–28. https://doi.org/10.1017/pan.2017.39.

President's Council of the Advisors on Science and Technology. 2012. "Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics." *Tech. rep.*

Prior, Markus. 2013. “Media and Political Polarization.” *Annual Review of Political Science* 16 (1): 101–27. https://doi.org/10.1146/annurev-polisci-100711-135242.

Proksch, Sven-Oliver, and Jonathan B. Slapin. 2012. “Institutional Foundations of Legislative Speech.” *American Journal of Political Science* 56 (3): 520–37. https://doi.org/10.1111/j.1540-5907.2011.00565.x.

Proksch, Sven-Oliver, and Jonathan B. Slapin, eds. 2015. *The Politics of Parliamentary Debate:* Cambridge University Press.

Putnam, Robert D. 1993. *Making Democracy Work: Civic Traditions in Modern Italy.* Princeton, NJ: Princeton University Press.

Resnjanskij, Sven, Jens Ruhose, Simon Wiederhold, and Ludger Woessmann. 2021. “Can Mentoring Alleviate Family Disadvantage in Adolscence? A Field Experiment to Improve Labor-Market Prospects.” *CESifo Working Paper*.

Rinne, Risto, Johanna Kallo, and Sanna Hokka. 2004. “Too Eager to Comply? OECD Education Policies and the Finnish Response.” *European Educational Research Journal* 3 (2): 454–85. https://doi.org/10.2304/eerj.2004.3.2.3.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools, and Academic Achievement.” *Econometrica* 73 (2): 417–58. https://doi.org/10.1111/j.1468-0262.2005.00584.x.

Rivkin, Steven G., and Jeffrey C. Schiman. 2015. “Instruction Time, Classroom Quality, and Academic Achievement.” *The Economic Journal* 125 (588): F425-F448. https://doi.org/10.1111/ecoj.12315.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. “Stm : An R Package for Structural Topic Models.” *Journal of Statistical Software* 91 (2). https://doi.org/10.18637/jss.v091.i02.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94 (2): 247–52. https://doi.org/10.1257/0002828041302244.

Romer, Paul M. 1990. “Endogenous Technological Change.” *Journal of Political Economy* 98 (5). http://www.jstor.org/stable/2937632.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2022. “What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature.” http://arxiv.org/pdf/2201.01194v3.

Ryu, Jung S. 1982. “Public Affairs and Sensationalism in Local TV News Programs.” *Journalism Quarterly* 59 (1): 74–137. https://doi.org/10.1177/107769908205900111.

Salla, Simola. 2020. “A Century of Partisanship in Finnish Political Speech.” *Working Paper*. https://sites.google.com/site/sallasimolaecon/home/. Accessed 3rd, February 2022.

Salmond, Rob. 2014. “Parliamentary Question Times: How Legislative Accountability Mechanisms Affect Mass Political Engagement.” *The Journal of Legislative Studies* 20 (3): 321–41. https://doi.org/10.1080/13572334.2014.895121.

Sancassani, Pietro. 2021. “The Effect of Teacher Characteristics on Students’ Science Achievement.” *ifo Working Paper* (No. 348).

Sansone, Dario. 2017. “Why Does Teacher Gender Matter?” *Economics of Education Review* 61 (6453): 9–18. https://doi.org/10.1016/j.econedurev.2017.09.004.

Schultz, Theodore W. 1961. “Investment in Human Capital.” *The American Economic Review* 51 (1): 1–17. http://www.jstor.org/stable/1818907.

Schwerdt, Guido, and Amelie C. Wuppermann. 2011. “Is Traditional Teaching Really All That Bad? A Within-Student Between-Subject Approach.” *Economics of Education Review* 30 (2): 365–79. https://doi.org/10.1016/j.econedurev.2010.11.005.

Settle, Jaime E. 2018. *Frenemies: How Social Media Polarizes America:* Cambridge University Press.

Slagter, Tracy Hoffman, and G. Loewenberg. 2007. “The Persistence of Procedural Consensus in the German Bundestag.” *Unpublished Manuscript*.

Sloane, Peter F.E, and Bernadette Dilger. 2005. “The Competence Clash – Dilemmata Bei Der Übertragung Des 'Konzepts Der Nationalen Bildungsstandards' Auf Die Berufliche Bildung.” Unpublished manuscript, last modified February 18, 2022. www.bwpat.de/ausgabe8/sloane_dilger_bwpat8.shtml.

Smith, Adam. [1776] 1979. *An Inquiry into the Nature and Causes of the Wealth of Nations.* Oxford: Clarendon Press.

Solow, Robert M. 1956. “A Contribution to the Theory of Economic Growth.” *The Quarterly Journal of Economics* 70 (1): 65. https://doi.org/10.2307/1884513.

Soroka, Stuart, Mark Daku, Dan Hiaeshutter-Rice, Lauren Guggenheim, and Josh Pasek. 2018. “Negativity and Positivity Biases in Economic News Coverage: Traditional Versus Social Media.” *Communication Research* 45 (7): 1078–98. https://doi.org/10.1177/0093650217725870.

Spirling, Arthur, and Iain McLean. 2007. “UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons.” *Political Analysis* 15 (1): 85–96. https://doi.org/10.1093/pan/mpl009.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. “Searching for Effective Teachers with Imperfect Information.” *Journal of Economic Perspectives* 24 (3): 97–118. https://doi.org/10.1257/jep.24.3.97.

Stecker, Christian, Jannis Kachel, and Jana Paasch. 2021. “Patterns of Lawmaking in the German Länder.”.

Sunde, Uwe, Thomas Dohmen, Benjamin Enke, Armin Falk, David Huffman, and Gerrit Meyerheim. 2022. “Patience and Comparative Development.” *The Review of Economic Studies* 89 (5): 2806–40. https://doi.org/10.1093/restud/rdab084.

Sunstein, Cass R. 2018. *#Republic: Divided Democracy in the Age of Social Media:* Princeton University Press.

Sutter, Matthias, Martin G. Kocher, Daniela Glätzle-Rützler, and Stefan T. Trautmann. 2013. “Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior.” *American Economic Review* 103 (1): 510–31. https://doi.org/10.1257/aer.103.1.510.

Swan, Trevor W. 1956. “Economic Growth and Capital Accumulation.” *Economic Record* 32 (2): 334–61. https://doi.org/10.1111/j.1475-4932.1956.tb00434.x.

Tatto, Maria Teresa, Ray Peck, John Schwille, Kiril Bankov, Sharon L. Senk, Michael Rodriguez, Lawrence Ingvarson, Mark Reckase, and Glenn Rowley. 2012. *Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M).* Amsterdam.

Taylor, Eric S., and John H. Tyler. 2012. “The Effect of Evaluation on Teacher Performance.” *American Economic Review* 102 (7): 3628–51. https://doi.org/10.1257/aer.102.7.3628.

TAZ. 2001. “Fast in Jeder Hinsicht Ein Desaster.” *TAZ*, December 5, 2001. https://taz.de/!1137421/.

Thorson, Kjerstin, Kelley Cotter, Mel Medeiros, and Chankyung Pak. 2021. “Algorithmic Inference, Political Interest, and Exposure to News and Politics on Facebook.” *Information, Communication & Society* 24 (2): 183–200. https://doi.org/10.1080/1369118X.2019.1642934.

Tillmann, Klaus-Jürgen. 2004. “Was ist eigentlich neu an PISA? Zum Verhältnis von erziehungswissenschaftlicher Forschung, öffentlicher Diskussion und

bildungspolitischem Handeln." Neue Sammlung 44 (2004) 4, S. 477-486. *Neue Sammlung* 44. https://doi.org/10.25656/01:2571.

Tresch, Anke. 2009. "Politicians in the Media: Determinants of Legislators' Presence and Prominence in Swiss Newspapers." *The International Journal of Press/Politics* 14 (1): 67–90. https://doi.org/10.1177/1940161208323266.

UNESCO. 2021. "World Development Indicators." Accessed November 02, 2021. https://databank.worldbank.org/reports.aspx?source=2&series=SE.SEC.DURS&country=.

United Nations. 2014. *World Economic Situation and Prospects 2014 - Country Classification.* New York. Accessed November 02, 2021. https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf.

Waldow, Florian. 2009. "What PISA Did and Did Not Do: Germany After the 'PISA-Shock'." *European Educational Research Journal* 8 (3): 476–83. https://doi.org/10.2304/eerj.2009.8.3.476.

Wayne, Andrew J., and Peter Youngs. 2003. "Teacher Characteristics and Student Achievement Gains: A Review." *Review of Educational Research* 73 (1): 89–122. https://doi.org/10.3102/00346543073001089.

Wedel, Katharina. 2021. "Instruction Time and Student Achievement: The Moderating Role of Teacher Qualifications." *Economics of Education Review* 85 (27): 102183. https://doi.org/10.1016/j.econedurev.2021.102183.

West, Martin R., and Ludger Woessmann, eds. 2021. *Public Opinion and the Political Economy of Education Policy Around the World.* CESifo seminar series. Cambridge Massachusetts: The MIT Press.

Winters, Marcus A., Robert C. Haight, Thomas T. Swaim, and Katarzyna A. Pickering. 2013. "The Effect of Same-Gender Teacher Assignment on Student Achievement in

the Elementary and Secondary Grades: Evidence from Panel Data." *Economics of Education Review* 34 (5): 69–75. https://doi.org/10.1016/j.econedurev.2013.01.007.

Woessmann, Ludger. 2010. "Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries." *Jahrbücher für Nationalökonomie und Statistik* 230 (2): 234–70. https://doi.org/10.1515/jbnst-2010-0206.

Woessmann, Ludger, Elke Luedemann, Gabriela Schuetz, and Martin R. West. 2009. *School Accountability, Autonomy and Choice Around the World.* Cheltenham: Edward Elgar.

Woessmann, Ludger, and Martin West. 2006. "Class-Size Effects in School Systems Around the World: Evidence from Between-Grade Variation in TIMSS." *European Economic Review* 50 (3): 695–736. https://doi.org/10.1016/j.euroecorev.2004.11.005.

World Bank. 2021. "World Development Indicators." Accessed November 02, 2021. https://databank.worldbank.org/reports.aspx?source=2&series=NY.GNP.PCAP.KD&country=.