// NO.18-021 | 08/2020

DISCUSSION PAPER

// SEBASTIAN CAMARERO GARCIA

Inequality of Educational Opportunities and the Role of Learning Intensity

Leibniz Association

ZEW

Inequality of Educational Opportunities and the Role of Learning Intensity*

Sebastian Camarero Garcia†
ZEW Mannheim ‡, University of Mannheim§ and CEP at LSE ¶

First Version: May 2018
This Version: August 2020

LATEST VERSION HERE

Abstract

Over the 2000s, many federal states in Germany shortened the duration of secondary school by one year while keeping the curriculum unchanged. The quasi-experimental variation arising from the staggered introduction of this reform allows me to identify the causal effect of increased learning intensity, the ratio of curricular content covered per year, on Inequality of Educational Opportunity (IEOp), the share in educational outcome variance explained by predetermined *circumstances* beyond a student's control. Findings show that higher learning intensity aggravated IEOp due to parental resources becoming more important through support opportunities like private tuition, adapting to an intensified educational process. The effect is stronger for mathematics/science than for reading, implying the existence of subject-dependent curricular flexibilities. My findings underscore the importance of accounting for distributional consequences when evaluating reforms aimed at increasing educational efficiency and point to the role of learning intensity for explaining changes in educational opportunities influencing social mobility.

Keywords: educational efficiency; human capital; inequality of opportunity; social mobility; school reform; compulsory education
JEL-Classification: D63, H75, I24, I28, J24, J62

---

*A first version of this paper was published as *ZEW Discussion Paper No.18-021* in May 2018 and entitled "*Inequality of Educational Opportunities and the Role of Learning Intensity: Evidence from a Quasi-Experiment in Germany* ." I would like to thank my supervisor Andreas Peichl. Moreover, I am thankful to Philipp Dörrenberg, Christina Felfe, Lenka Fiala, Cung T. Hoang, Kilian Huber, Paul Hufe, Julien Lafortune, Eckhard Janeba, Stephen Kastroyano, Stephen Machin, Daniel Mahler, Panos Mavrokonstantis, Federico Rossi, David Schönholzer, Sebastian Siegloch, Konrad Stahl, Holger Stichnoth, Michèle Tertilt as well as participants at the Public Economics, Macroeconomics and CDSE-Seminar at ZEW/University of Mannheim, the CEP Labour Workshop at LSE, the ECINEQ Winter School in Canazei, and at conferences of the ECINEQ at CUNY, New York City, the IARIW in Copenhagen, the VfS in Freiburg and the EALE in Lyon for their helpful comments and discussions. I would also like to thank the IQB and the Research Data Center in Berlin for granting me permission to conduct this analysis and for their support. I acknowledge financial support for pursuing my PhD studies by the Cusanuswerk. Moreover, I appreciate the fellowship of the German National Academic Foundation. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. I declare that I have neither material interests, nor other conflicts of interests that relate to the findings of this paper. My thanks also go to Lukas Bauer, Yasemin Karamik, Marc Stadtherr, and Markus Teichmann.

†E-Mail: Sebastian.CamareroGarcia@zew.de or sebastian.camarerogarcia@gmx.com.
‡ZEW – Leibniz Centre for European Economic Research (ZEW Mannheim), L7 1, 68161 Mannheim, Germany
§University of Mannheim, Department of Economics, L7 3-5, 68161 Mannheim, Germany
¶Centre for Economic Performance (CEP), London School of Economics, Houghton Street, London WC2A 2AE

1 Introduction

In modern societies, the general belief that, by working and studying hard, everyone has a fair chance at climbing the social ladder has been central to maintaining social cohesion and political stability. However, in an era of relatively high inequality of wealth and incomes compared to the post-war decades in most developed countries (Piketty & Zucman, 2014), an increase in the number of both citizens who fear that their children could be worse off in the future (fear of *downward* mobility) and groups in society who believe that the "game is rigged" (fear of a lack of *upward* mobility) might explain rising political polarization. For these reasons, the reduction in social mobility[1] is becoming an increasingly important issue when it comes to understanding recent trends of inequality within society. As education tends to be the main vehicle for *upward* mobility, it is of key policy interest to analyze educational systems in terms of equality and to detect drivers of *Inequality of Opportunity* in particular (as Chetty et al. (2020) for US colleges). Yet in times of public spending constraints, accelerating growth of knowledge and higher economic competition, political attention has shifted onto making a country's educational system more efficient (Machin, 2014). In fact, reforms have started to focus on compressing educational processes, i.e. on increasing learning intensity.

This paper contributes to the issue of how the trend in intensification of education may explain decreased social mobility by analyzing the question of how increasing learning intensity affects Inequality of Educational Opportunity (IEOp). Thus, I shift focus onto the distributional concerns and the potential unintended consequences for social mobility that arise from compressing educational processes. If, for instance, higher intensity made it harder to learn the curriculum through schooling alone, educational opportunities could become more dependent on a student's parental support resources. In this context, I adopt the concept as illustrated by Roemer and Trannoy (2015), stating that society has achieved *Equality of Opportunity* if what individuals achieve regarding a desirable objective is determined by their *efforts* (e.g. how hard they study), instead of by *circumstances* that are beyond individual control (e.g. sex). Thus, IEOp[2] is defined as inequality in the distribution of educational outcomes that can only be attributed to *circumstances* through either their direct or indirect (via changing *efforts*) impact on outcomes. It is a relative measure of educational mobility.

This paper is among the first to provide an analysis of IEOp in a quasi-experimental setting that goes beyond its pure measurement. As Ramos and Van de gaer (2016) point out, the understanding of how institutions influence IEOp is still limited. Therefore, my contribution to this issue consists in providing evidence on the role of learning intensity as a relevant policy dimension that causally affects IEOp. From a social welfare perspective, it is interesting to reveal the effects of increasing learning intensity on both academic achievement and IEOp. Pareto-improvements might be realized if intense curricula proved to be an instrument to overcome the trade-off between educational spending and schooling outcomes.

[1] For instance, Chetty et al. (2017) provide evidence for falling absolute income mobility. Organization of Economic Co-operation and Development (OECD) data from 2012 confirm low absolute educational mobility. In particular, Germany reaches only below average social mobility rates in terms of the percentage of 25-64 year-old non-students whose educational attainment is higher (*upward* mobility) or lower (*downward* mobility) than that of their parents (Graph A.4.3 in OECD (2014)).

[2] Inequality of Opportunity (IOp) and Equality of Opportunity (EOp) refer to the same concept, placing emphasis on either the unfair or fair part within the distribution of opportunities. If opportunities depend less on factors beyond individual control but more on *efforts*, IOp (EOp) will decrease (increase). In line with Brunori et al. (2012), instead of IOp (EOp) in education, I use the expression IEOp (Equality of Educational Opportunity (EEOp)). I will only use IOp or IEOp for ease of interpretation.

To identify the causal effect of (increased) learning intensity on IEOp, I analyze an educational reform in Germany. During the last decade, Germany's federal states shortened secondary school for the academic track (*Gymnasium*) from nine to eight years at staggered time points between 2001 and 2008. This so called Gymnasium-8 reform (G-8 reform) reduced school duration by one year but kept the curriculum unchanged for the affected (treated) student cohorts. Due to the implementation of the reform, there were two cohorts who would finish school together while one cohort entered one year earlier than the other, leading to differences in years of schooling (9 vs. 8 years). As both cohorts had to take the same final exams in the same year, treated students had less time to learn the same material, thus experiencing higher learning intensity. This staggered introduction of the reform across federal states generates quasi-experimental variation that allows the application of a Difference-in-Difference estimation approach (DiD) to derive the causal effect of the increase in learning intensity on IEOp, comparing the respective treatment and control groups over time.

For the purpose of measuring IEOp, I use Program for International Student Assessment (PISA) data to construct a representative sample of students in the ninth grade. The data include standardized test scores in reading, mathematics, and science which are comparable across time and federal states unlike grading schemes that vary with year and state (for more details on the data I use, see Appendix A.4.1). Moreover, these data contain a rich set of family background variables which allow me to define relevant *circumstances*. I also apply a new machine learning approach to cross-validate my theory-driven choice of variables. Ultimately, IEOp reflects the coefficient of determination when regressing test scores on these *circumstances* variables.

The analysis yields three main findings. First, the estimated size of IEOp, 20-35% of the variance in cognitive test scores that can only be attributed to *circumstances*, corresponds to the levels of common estimates for inequality of opportunity in income. Second, the reform-induced increase in learning intensity led to a significant rise in IEOp, by at least 10 percentage points of the explained test score variance for affected (treated) students. Given the initial size of IEOp and the fact that this paper's IEOp measures are lower bound estimates, this corresponds to relative increases in IEOp of at least 25%. Third, the results provide some evidence for the existence of subject-dependent curricular flexibilities. In fact, less flexible skills in mathematics and science are more responsive to changes in curricular intensity. Conversely, reading competency is trained predominantly through its usage in everyday life and thus less dependent on schooling. Finally, the results can be rationalized by differential compensation possibilities for higher learning intensity depending on parental resources, specifically the capacity to finance private tuition or to provide academic support themselves. This shows that there are important distributional concerns with respect to providing equal opportunities (cf. Andreoli et al. (2018)) that must be taken into account when designing reforms altering the intensity of educational processes.

This paper offers several contributions to the existing literature. First, I contribute to the strand of research on measuring Inequality of Opportunity (IOp) with respect to educational outcomes by adding empirical evidence on how Inequality of Educational Opportunity (IEOp) has changed over time in a developed country. So far, papers dealing with IOp have focused on measurement issues, using income as the main outcome variable (e.g. Almås et al. (2011)). Concerning IOp in educational outcomes, most studies focus on measuring IEOp for developing countries (e.g. Gamboa and Waltenberg (2012)). The few

papers on developed countries mostly follow a cross-country comparison approach using PISA data to achieve comparability of educational achievement measures over time and across countries (e.g. Ferreira and Gignoux (2013)). Instead, my study estimates IEOp for Germany exploiting quasi-experimental within-country variation (as Cantoni et al. (2017) for China). Such settings allow going beyond measuring IEOp to actually estimate the causal effects of specific policies on IEOp. For instance, some studies analyze IEOp in the context of reforms that changed tertiary education systems (e.g. Brunori et al. (2012) on Italy). They find that both expanding higher education through opening more sites as well as reducing the length of getting a first-level degree to have a positive effect on Equality of Educational Opportunity (EEOp). However, only a few studies investigate the impact of school reforms on IEOp (e.g. Edmark et al. (2014) for Sweden). In this paper, I add evidence on how IEOp changed over time in Germany and focus on estimating the causal effect of increasing learning intensity on IEOp for the academic track in the secondary school system.

Second, this work contributes to a strand of the literature analyzing educational policy reforms to identify the underlying role of different input factors in the human capital accumulation process. Even though the G-8 reform shows that changing school intensity is an important consideration in educational policy-making, research on such reforms is still limited. To begin with, empirical work has analyzed the effects of variations in pure schooling quantity without considering learning intensity. In that context, most studies focus on reforms that increase educational participation, such as policies raising compulsory minimum duration of schooling. They usually find the returns of additional schooling on earnings to be positive (e.g. Angrist and Krueger (1991); Grenet (2013); Aakvik et al. (2010); Eble and Hu (2019)). Furthermore, the impact of differences in instructional time on academic performance has been investigated. Relying on either cross-national or within-country variation in instructional time, most studies find a positive impact of additional time on standardized test scores (e.g. Aksoy and Link (2000), Marcotte (2007), Lavy (2015)). However, only a few studies have analyzed the impact of variations in instructional time when curricular content can be assumed to remain constant. In this context, reforms that shortened the duration of schooling while keeping curricular content unchanged allow for evaluating the impact of increasing learning intensity. For instance, analyzing a similar school reform in parts of Canada, Krashinsky (2014) finds only small long-term effects on wages. This suggests that increased learning intensity might not affect earnings permanently.[3] The results are in line with Pischke (2007), who exploits a German reform in the 1960s that changed the start of the school year to autumn by implementing two short school years. The reform led to a significant increase in the number of students repeating a grade, but only small effects on earnings persisted.

Despite the resulting public controversy that has even led some federal states to reverse the reform, only a few studies have evaluated the G-8 reform and its effects on educational outcomes (A.3.1 in Appendix A.3 provides an overview of the related literature). Additionally, the findings of those studies vary depending on the chosen educational outcome measure. For instance, Huebener and Marcus (2017) find that the reform had, on average, a significantly negative effect on GPA (grades) of students. On the contrary, the reform tends to have a positive effect on cognitive test scores as measured by PISA data (see, e.g. Huebener et al. (2017)). Furthermore, the results of Marcus and Zambre (2019) indicate

[3] Whether this is true due to schooling working primarily as a signal or because increased intensity may compensate for less schooling and maintain the human capital accumulation process is unclear.

that the reform led to falling enrolment rates at university.[4] Moving away from the explicit effect on direct educational outcome variables, the analysis in my paper shifts focus in the evaluation of the G-8 reform onto distributional concerns. In other words, I evaluate whether the reform is *inclusive*, i.e. it decreases IEOp while maintaining at least test score results, or *selective*, i.e. it increases IEOp, (Checchi & van de Werfhorst, 2018). In particular, my findings are relevant for policy suggestions on designing curricula that take the effect of learning intensity on both cognitive skill formation and IEOp into account. Implementing a whole-day school system, for instance, might limit the necessity for parents to help students deal with compressed schooling curricula.

Thirdly, my paper relates to the emerging literature on finding drivers of inequality in educational outcomes which are key determinants of recent trends in decreased social mobility (e.g. Chetty et al. (2020); Philippis and Rossi (2019); Rothstein (2019); Boneva and Rauh (2018)). I contribute to this strand of research by providing evidence that the previously neglected factor of learning intensity might be a relevant policy channel for both the effectiveness of (non-)cognitive skill formation and the importance of *circumstances* for educational outcomes. While my analysis mainly focuses on exploiting a school reform to derive causal estimates on how intensified instruction affects IEOp, the interpretation of these results in terms of potential mechanisms complements explanations delivered by this most recent strand of literature. Although a complete model of learning intensity, IEOp, and its connection to social mobility is beyond the scope of this study, I provide evidence on which future research tackling this big picture question can base itself. This also supports the integration of learning intensity as a key factor into the human capital literature.

The remainder of this paper is organized as follows. Section 2 illustrates the institutional background and the G-8 reform on which the identification strategy relies. Section 3 explains how IEOp is measured given the data in this study. In Section 4, the empirical strategy is illustrated. Section 5 provides the results with robustness checks and a discussion on their implications. Section 6 concludes.

2 Institutional Setting: the "G-8 reform"

This section explains the institutional background and implementation of the G-8 reform which can be exploited as a quasi-experiment to analyze the role of increased learning intensity on IEOp.

2.1 Institutional Background: the German School System and Reform Debate

Like the United States, Germany has a federal structure. Education policy strictly falls under the remit of the 16 federal states (*Länder*), yet most features are comparable across states. School usually starts at the age of six, when students enter primary school for a period of four years. Afterwards, students follow a tripartite secondary school system, where the choice of track is determined by their previous academic performance.[5]

[4] Further related work on the G-8 reform includes Andrietti and Su (2019); Thiel et al. (2014); Büttner and Thomsen (2015); T. Meyer and Thomsen (2016); T. Meyer et al. (2018) as explained in A.3.1 in Appendix A.3.

[5] Primary schools issue recommendations for each student regarding which secondary school track the student should enter (Dustmann et al., 2017). Based on a student's performance in primary school, recommendations were binding in federal states for the time considered in this study. An overview of the regulations on the transition from primary to secondary education for the period studied here is available on `https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2006/2006_03_01-Uebergang-Grundschule-Sek1.pdf`.

Both the shortest track of secondary school, *Hauptschule*, and the intermediary track, *Realschule*, allow graduates to pursue apprenticeship programs after a total of nine or ten years of schooling. Only the academic track, *Gymnasium*, which this paper focuses on, leads to a diploma granting access to university (*Abitur*). On average, the largest share of all students in secondary school (about 40 percent of each cohort) attended this track in the time period 2000 until 2012. Traditionally, the academic track lasted nine years (for a total of 13 years including primary school) in West Germany. However, the former German Democratic Republic (GDR) had a different school system: All students were taught together for ten years, after which they could either follow vocational training or complete two additional years of *Gymnasium* to obtain the *Abitur*. Following reunification, most East German federal states adopted the West German standard, the Gymnasium-9 model (G-9 model), but two states, Saxony and Thuringia, maintained the Gymnasium-8 model (G-8 model).[6]

Later, in the early 2000s, the nine years were perceived as a competitive disadvantage for the economy because they contributed to the comparatively advanced age at which Germans entered the labor market after school and/or university. Moreover, the long duration of the academic track was criticized for hindering the creation of a more comparable, harmonized framework for tertiary education in the European Higher Education Area (EHEA). Thus, in order to adjust school duration to the average among OECD countries of twelve years, federal states decided to shorten the *Gymnasium* to eight years without reducing the curriculum, also known as the Gymnasium-8 reform (G-8 reform).[7]

2.2 Implementation of the Reform: Increasing Learning Intensity

After 2001, all 14 federal states with a G-9 model shortened their academic secondary school track from nine to eight years. With the graduation of a *double cohort* consisting of both the first G-8 model and the last G-9 model student cohort that together had to pass the same final exams (*Abitur*) in the same year, the reform process took eight years to transform all grades of *Gymnasium*.

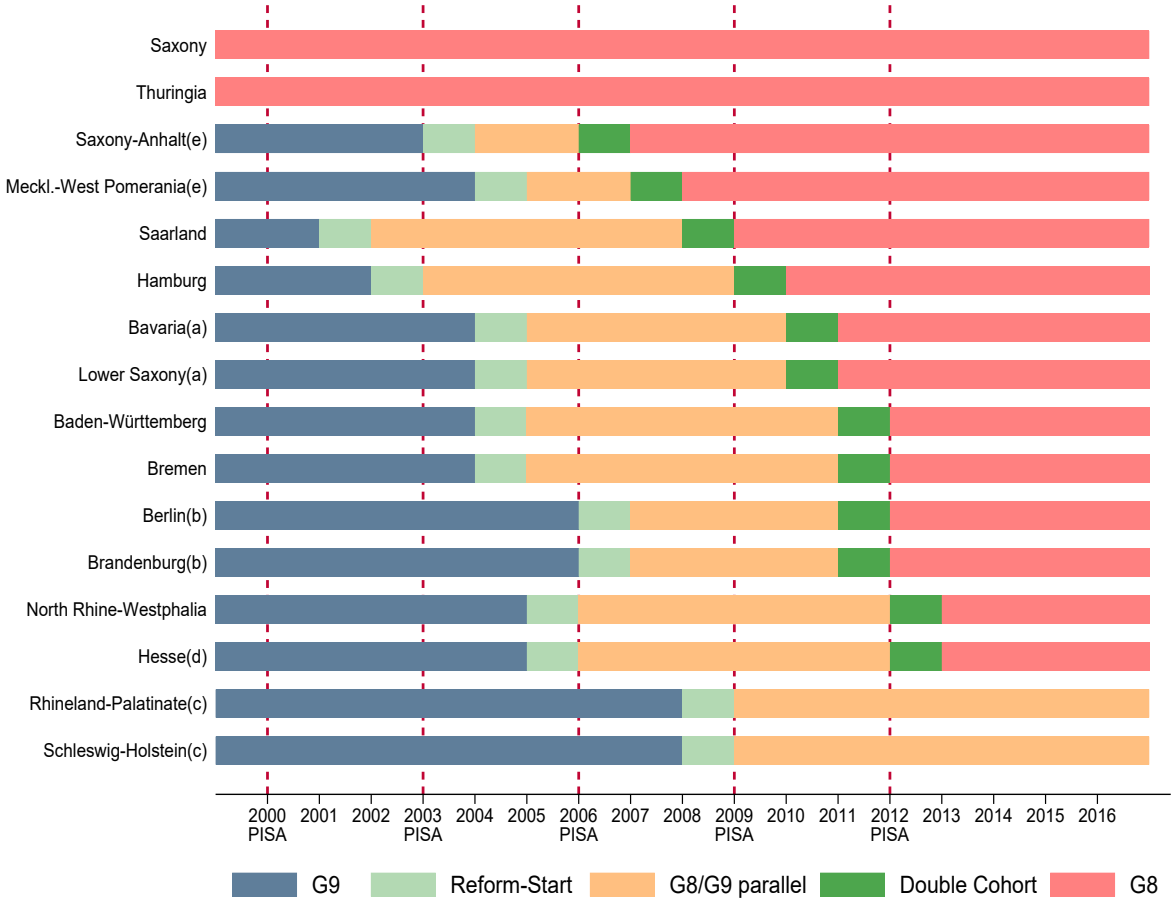
For the purpose of this paper, two features of the reform are particularly important. First, as shown in Figure 1, not all federal states started the reform process at the same time. Some of them began in school year 2001/2002 whereas others waited until school year 2008/2009, creating *double cohorts* which graduated between 2006/2007 and 2015/2016. Second, although the academic track was reduced by one school year, the curricular content remained at the original level. In fact, education ministers decided that standards for the university access diploma (*Abitur*) were not to be lowered in response to the reform. The minimum number of 265 instruction hours per school year over all grade levels was maintained, as was the total number of lessons required to graduate from the *Gymnasium* KMK (2016). This should ensure comparable standards nationwide for university access diplomas despite the differences in school duration. Adding more content to the last two years of the *Gymnasium* was perceived to be difficult as the first G-8 model and the last G-9 model cohort had to complete those grades together. Only marks during

[6] In addition to the three different school tracks, federal states have recently started to provide a comprehensive school (*Integrierte Gesamtschule*), in which students are not channeled into specific academic paths after primary school, but can graduate after 9, 10, or 13 years. However, this option played a negligible role for the considered time period (2000-2012) as the vast majority of students achieving *Abitur* still attended *Gymnasium*. See Figure A.1 in Appendix A.1 for further details on the German education system.

[7] For further arguments discussed during the reform debate, please refer to A.3.2 in Appendix A.3.

the final two years and in the *Abitur* exams count towards the final GPA. Therefore, authorities focused the compression on the first years of *Gymnasium*, squeezing the material originally taught in the seven years between grades 5 and 11 into the six years between grades 5 and 10. Students in the G-8 model were supposed to enter the final two years of *Gymnasium* as if they had completed the original 11^{th} grade.

Figure 1: Implementation of the G-8 Reform across Federal States



Notes: This figure illustrates for each federal state whether the graduating cohort in each school year of the *Gymnasium* was in a G-8 model, G-9 model, consisted of the *double cohort* or whether due to the reform implementation process both models existed parallel with younger grades already in a G-8 model and older ones still in a G-9 model.
Notes on some states:

a In Bavaria and Lower Saxony, the 5^{th} and 6^{th} grades were allocated into the G-8 model in the same school year. However, the 9^{th} graders in 2009 were affected by the reform from the 5^{th} grade onward.

b Berlin and Brandenburg, where primary school lasts six years, introduced the reform for 7^{th} grade onward.

c Rhineland-Palatinate and Schleswig-Holstein planned to introduce the G-8 reform for school year 2008/09 to be completed by 2015/16. In the end, both kept the G-9 model for all grades and over all PISA waves.

d Hesse introduced the reform over 3 years: The "main" *double cohort* covering 60% of schools is shown.

e Mecklenburg-West Pomerania and Saxony-Anhalt introduced the reform directly for the 9^{th} grade onward.

Source: The figure has been constructed based on facts as provided in Table 2 and the regulations explained in Table A.1 in Appendix A.2. This figure corresponds to the geographical maps illustrating the implementation of the reform across time and space in Figure A.2 in Appendix A.1.

To keep the required total minimum weekly lessons unchanged for the new G-8 model, instructional time increased by about two hours a week during grades 5-10 for G-8 model students compared to previous cohorts in the G-9 model.[8] However, the total loss in time of one school year was not fully compensated by additional instructional time per week: In order to limit the amount of afternoon schooling in 5^{th} and 6^{th} grade, hours originally planned for revision (beyond the minimum required) were dropped and instead used to already teach new curricular content at an earlier point in time compared to the G-9 model. Therefore, it is plausible to assume that total curricular content was not reduced for the first student cohorts affected by the G-8 reform that are in the focus of this study, in any of the federal states. As curricular content in the G-8 model began to change in the years after 2012 (cf. Table A.1 in Appendix A.2), this assumption would not necessarily hold for later G-8 model cohorts. By using data of ninth graders tested in 2012 or before, I focus on the very first cohorts affected by the reform and these later changes do not affect the analysis.

In conclusion, the G-8 reform exogenously led to a considerable increase in learning intensity over the first years of the *Gymnasium*. That is, the amount of material covered per week increased for each grade level (excluding the final two grade levels).

3 Data and Measuring Inequality of Educational Opportunity

In this section, I first focus on which specific PISA data are used for my analysis.[9] Second, I explain how one can measure IEOp, the main outcome variable, based on the related literature and the educational data available for the main test domains in mathematics, reading and science. Third, I provide some descriptive analysis on the *circumstances* variables defined for this paper.

3.1 PISA Data

For Germany, two types of PISA test data are available, the version conducted for international comparisons (PISA-I) and a national extension (PISA-E). The PISA-I data result from students who take the test on the same day and are selected in a two-stage sampling procedure. In the first stage, schools from the 16 federal states of Germany are randomly selected. In the second stage, for each school, about 25 students of age 15 are randomly taken for the test (*age-based* sample). Additionally, within already selected schools, two classes of ninth graders with a minimum of 25 students are randomly chosen (*grade-based* sample). In total, the *grade-based* PISA-I sample consists of about 10,000 students from about 225 schools (Table A.2). Thus, its sample size is about twice as large as that of the *age-based* sample. While comparisons across countries are best carried out at a given age, for the strategy pursued in this paper, a comparison among ninth graders is more appropriate because the G-8 reform affected students based on their grade in a certain school year.

[8] However, this is only an approximation for an average student; the exact changes depend on the federal state. Huebener et al. (2017) have collected binding timetable regulations for each federal state and show the changes in the distribution of average weekly instruction hours. This confirms the interpretation of the G-8 reform: on average hours per grade increased by about 2 hours a week, i.e. by about 8-10% of weekly lessons per year during grades 5-10.

[9] Some background information on the OECD PISA data, its advantages for measuring educational outcomes, and the representativeness of these data across states, schools and over time is provided in Appendix A.4.1.

Moreover, national PISA extensions (PISA-E) were conducted for the years 2000, 2003, and 2006. Each of them consists of about 40,000 students. By oversampling less populated federal states, these extensions allow for a more robust comparison of educational performance between the German federal states.[10] However, PISA-E was discontinued in 2009 and replaced by the *federal state comparison test* which is conducted by the Institut zur Qualitätsentwicklung im Bildungswesen (IQB). This new comparison test aims to assess national educational standards determined by the Standing Conference of the Ministers of Education and Cultural Affairs (SC) of all federal states instead of the OECD. Since then, each extension of this comparison test covers only a particular domain (reading in 2009, mathematics and science in 2012), which prohibits their use for analyzing the entire period considered in this study (until 2012).

Nevertheless, Andrietti and Su (2019) or Huebener et al. (2017) use data from the national PISA extensions for the years 2000, 2003, and 2006. They complement them with single waves of PISA-I from the year 2009 and 2012. Only *grade-based* PISA-I samples provide all three domains consistently for each test year. Therefore, in order to have consistent comparability across the studies used, this paper avoids mixing PISA-E and PISA-I datasets and focuses on PISA-I data from the waves 2000, 2003, 2006, 2009, and 2012 only.[11]

As this paper focuses on the academic track (*Gymnasium*), only schools of this type are included in the sample. They make up more than one third of the *grade-based* PISA-I sample which corresponds approximately to the real share of students in *Gymnasium*. Finally, the analysis is restricted to variables derived from the questionnaire answered by students and their parents (the *student*-dataset). Thus, this paper relies on the *grade-based* PISA-I sample to construct a representative repeated cross-section of students in grade nine of the *Gymnasium*. This allows me to analyze the increase in IEOp due to the G-8 reform by using variables based on PISA test scores and the tested students' available background characteristics.

Descriptive Statistics Regarding the main outcome variables, PISA test scores in the domains of reading, mathematics, and science are above the German average when focusing on students in the academic track of secondary school. A typical ninth grader in *Gymnasium* achieves results that are about 60 points higher than for the average German ninth grader. This difference corresponds to about an entire *proficiency level*, that is, the value-added of two school years (compare Appendix A.4.1). With respect to the three testing areas, students perform worst in reading literacy. The reading skills score average stagnated or even slightly deteriorated between 2000 and 2012. This observation is in line with reports on German PISA results for the 2000s which show that students perform better in mathematical and scientific than reading tests (e.g. Klieme et al. (2011)). The mean scores in mathematics (about 580) exceed those in reading (about 570). Students perform best in science, on average, achieving up to 590 points (see Table A.3 in Appendix A.2).

[10] For this purpose, one day after the students for the PISA-I samples had taken their test, additional students in each federal state were randomly selected to undergo the same testing procedures for the PISA-E test in which they had to answer an additional national questionnaire.

[11] In 2000, there was no specific *grade-sample based* PISA-I sample available from the IQB, but PISA-2000 - being the PISA-2000-E dataset - is *ninth grade-based* (Baumert et al., 2002). Only one instead of the 80 usual replication weights is provided.

Furthermore, in all three domains, the median exceeds mean test scores, indicating more variation at the lower end of the performance scale.The mean/median comparison and its development may be regarded as a first sign for whether IEOp changes over time. The data show that median and mean deviate only slightly more after the reform than before. The same applies to the variance of test scores which do not change significantly over time. Finally, the analysis dataset contains more than 60 schools per test year across all federal states and, on average, the number of students increases with each test cycle (see Table A.3 in Appendix A.2 for an overview). Moreover, Figure A.4 in Appendix A.1 provides a descriptive analysis based on the used *grade-based* PISA-I dataset for different subgroups. For instance, students from academic households achieve slightly higher scores than those from non-academic ones.

3.2 Outcome Measure: Inequality of Educational Opportunity (IEOp)

The idea that societies should distribute opportunities equally has a long-standing tradition within political philosophy. Following Rawls (1971) seminal contribution and its discussion (e.g. Sen (1980)), a prerequisite for measuring Inequality of Opportunity (IOp) is distinguishing whether a form of inequality is acceptable or not within a society.[12] However, these ideas only started to capture the more widespread attention of economists when scholars such as Roemer (1998) translated these philosophical concepts into a more formal theoretical framework. Since then, an empirical literature has emerged, proposing several methods on how to estimate IOp as shown in survey articles by Ramos and Van de gaer (2016) and Roemer and Trannoy (2015).

In the following, I formulate a model regarding how to measure IEOp in line with Ferreira and Gignoux (2011, 2013). To begin with, it is useful to define a set of conceptual notions:

- An *advantage*, y, denotes an individual achievement. Studies typically focus on income; in this paper, the achievement corresponds to educational outcomes as measured by PISA test scores.
- The vector of *efforts*, E, denotes the set of variables that influence the outcome variable (*advantage*) and over which the student has control (e.g. choice of time for studying).
- The vector of *circumstances*, C, denotes the set of individual characteristics which are beyond the student's control, e.g. their family household's Socio-Economic Status (SES), parental education, gender, ethnicity, or innate ability/talents.

Consider a sample of S students indexed by $i \in \{1,...,S\}$. Each student i can be described by a set of attributes $\{y, C_n, E_m\}$, where y denotes an *advantage* (here test scores), C_n is a vector of n discrete *circumstances* and E_m denotes the vector of m discrete *efforts*. Without loss of generality, this model could be extended to the case of having continuous elements in the vectors of *circumstances*/*efforts*. Thus, we can represent the population by an $(n \times m)$ matrix $[Y_{nm}]$ with a typical element (*cell*)

$$y_{nm} = g(C_n, E_m) | C \in \Omega, E \in \Theta, g : \Omega \times \Theta \Longrightarrow \mathbb{R}$$

being the *advantage* that is a function of both *circumstances* and *efforts*. After selecting the appropriate set of variables capturing *circumstances* characteristics relevant to educational achievement that constitute

[12] There is strong experimental evidence that people distinguish acceptable (fair) and unacceptable (unfair) income inequality (Cappelen et al., 2010; Almås et al., 2011). It tends to be acceptable if differences are due to individual responsibilities (*efforts*), but not acceptable if these are due to *luck* (*circumstances*). Lefranc and Trannoy (2017) show how *luck* can be incorporated as an intermediary category between *circumstances* and *efforts*.

the n different vectors C_i for each student i, the sample can be split into n distinct groups of students sharing the same *circumstances* (they are of the same *type*). Similarly, the sample can be split into m distinct groups of students exerting the same level of *efforts*, but having different *circumstances* (they belong to the same *tranche*). Together *types* and *tranches* form the *cells*.

In the context of this paper, when assuming talents to be distributed normally across the whole population, the concept of Inequality of Educational Opportunity (IEOp) can be translated as follows. Students who work harder and put in greater *efforts* should be rewarded by achieving better educational results regardless of their specific *circumstances*. Hence, IEOp corresponds to differences in educational achievement between students who put in the same *efforts* but only differ in terms of their *circumstances* (*compensation principle*). In contrast, disparities in test results driven by variations in individual *efforts* are acceptable (*reward principle*). Thus, IEOp resembles differences in *advantages* between students that can only be attributed to *circumstances*.

Deriving a measure of IEOp involves two steps: An *Estimation Phase* to transform the original distribution $[Y_{nm}]$ into a smoothed one $[\tilde{Y}_{nm}]$ reflecting only the unfair inequality in $[Y_{nm}]$ and the *Measurement Phase*, which thereon applies a measure of inequality. Following the IOp literature, I apply an *ex-ante*, *between-types inequality* measurement approach.[13] As *efforts* are not directly observable, this is also in line with the *indirect* methods to measure IOp because the estimation is based solely on the observed marginal distribution of *advantages* (test scores) given by the vector $y = \{y_1, \ldots, y_S\}$ and on the joint distribution of *advantages* and *circumstances* over the sample population $\{y, C_n\}$. Therefore, I follow the measurement approach of Ferreira and Gignoux (2013) which has fewer requirements for data availability than a non-parametric approach. The reason is that the more precisely one tries to design the partition, the smaller *cells* become. Thus, large datasets (best with panel structure) are necessary to conduct a useful non-parametric *within-tranche inequality* decomposition (Checchi & Peragine, 2010).

Consequently, this paper adopts a parametric, *ex-ante* estimation approach to derive IEOp measures. I model test scores (y) as a function of *circumstances* (C) and *efforts* (E), as $y = f(C,E)$. *Efforts* can also depend on *circumstances*, i.e. $E = E(C)$, which implies $y = f(C, E(C))$. Within this framework, innate ability, for instance, is considered an unobserved *circumstance* factor that may influence test scores directly through cognitive skills, but also indirectly via its impact on work ethic and other characteristics associated with *efforts*. However, *efforts* cannot vice versa change other relevant *circumstances*, such as gender or parental education.[14] Moreover, as PISA evaluates students in the ninth grade, they are on average about 15 years old. Hufe et al. (2017) argue that choices made before an age of maturity (16) are likely beyond an individual's control. Thus, it is plausible to assume that tested students are (if at all) only partially responsible for their choices, and most unobserved factors would be *circumstances*. In summary, my model of measuring IEOp considers the role of *circumstances*, *efforts* and their interplay.

[13] One distinguishes between an *ex-ante* and *ex-post* approach. This refers to how one evaluates IOp, thus, to which normative welfare criterion is chosen. Before *effort* is realized (*ex-ante*), following van de Gaer's "mins of means" criterion, EOp is achieved equalizing mean outcomes across *types*. IOp is measured as *between-types inequality* satisfying *ex-ante* compensation. After *effort* is realized (*ex-post*), following Roemer's "means of min" criterion, EOp is achieved eliminating inequality within *tranches* satisfying *ex-post* compensation. Fleurbaey and Peragine (2013) show that *ex-post* and *ex-ante* compensation are incompatible. But if *efforts* are distributed independently from *circumstances*, *ex-post* and *ex-ante* EOp will be similar (Ramos & Van de gaer, 2016, propos. II).

[14] See Appendix A.5.6 for a discussion of how the concept of ability is considered in the context of measuring IEOp.

Following Ferreira and Gignoux (2013), a linear functional form is used:

$$y_i = C_i'\beta + E_i'\gamma + e_i \tag{1}$$

$$with \;\; E_i = C_i'\delta + u_i \tag{2}$$

C_i is a vector capturing *circumstances* variables and E_i is the unobserved vector of m *efforts* per student i. However, the aim being to estimate the full effect of *circumstances* on scores, i.e. both the direct and indirect effect on scores (via their impact on *efforts*), I estimate the reduced form model:

$$y_i = C_i'(\beta + \gamma\delta) + (e_i + u_i'\gamma) \tag{3}$$

$$i.e.: \quad y_i = C_i'\rho + z_i \;, \quad where \quad \rho = (\beta + \gamma\delta) \;\; and \;\; z_i = (e_i + \gamma u_i) \tag{4}$$

The residual, z_i, includes both unobserved *efforts* and unobserved *circumstances*. But at this point, the aim is to estimate the mean score outcome of each *type* conditional on *circumstances*:

$$\widehat{y_i} = C_i'\widehat{\rho} \tag{5}$$

This will create a new, simulated distribution of scores, $\widehat{y} = \{\widehat{y_1}, \ldots, \widehat{y_S}\}$. Thus, every student is assigned the value of their opportunity set (which in a linear regression corresponds to the expected score conditional on *circumstances*). This linear model can be estimated using an Ordinary Least Squares (OLS) regression providing the vector of predicted test scores (the *smoothed* distribution).

Having assigned each individual the value of their opportunity set, the second step, the *Measurement Phase*, then involves calculating inequality in this new distribution, using a particular inequality index, I(.). To estimate IEOp, one would estimate the following ratio:

$$\widehat{\theta}_{IEOp} = \frac{I(\widehat{y_i})}{I(y_i)} = \frac{I(C_i'\widehat{\rho})}{I(y_i)} \tag{6}$$

i.e. the ratio between inequality in *circumstances* (the simulated distribution) and total inequality (actual distribution of scores). Thus, instead of using an absolute measure, I use a relative measure of IEOp. This is also suited best to evaluate the reform effect when comparing treatment and control groups over time because the relative change is of primary interest and can be interpreted most intuitively. What remains is the choice of an appropriate inequality index I(.). The literature on IOp in income has used the Mean Log Deviation (MLD) index due to its desirable properties, e.g. path independence. However, Ferreira and Gignoux (2013) show that the MLD is not appropriate for measuring inequality in PISA data. The reason is that it is not ordinally invariant to the standardization of PISA test scores. Instead, the authors prove that the variance is the most appropriate inequality index for IEOp. Being an absolute measure of inequality itself, it is ordinally invariant in the test score standardization and satisfies the most important axioms to be qualified as meaningful inequality measure, i.e. it satisfies (i) symmetry, (ii) continuity, and (iii) the transfer principle.

Overall, the variance satisfies requirements for the proposed Inequality of Educational Opportunity (IEOp) measure and can be calculated as:

$$\widehat{\theta}_{IEOp} = \frac{variance(\widehat{y})}{variance(y)} \tag{7}$$

This measure is attractive for various reasons. First, it is the coefficient of determination (R^2) of an OLS regression of test scores on *circumstances C* variables which eases measurement procedures.[15] Second, as shown in Ferreira and Gignoux (2011), the R^2 results in a meaningful summary statistic, the lower bound of the true IEOp. Since the object of interest is the joint effect of all *circumstances* on educational outcomes as measured by test scores, it is necessary to understand what percentage of the variation in scores, y, is causally explained by the overall effect of *circumstances* (both directly and indirectly, via *efforts*). *Efforts* are treated as generally unobserved omitted *circumstances* variables. If we observed them, they would only lead to a finer partitioning of $[Y^i_{nm}]$, which would further increase the IEOp measure. Therefore, the R^2 measure, $\widehat{\theta}_{IEOp}$ in Equation (7), is a valid lower bound estimate of the joint effect of all *circumstances* on educational achievement. In other words, it is the lower bound of the share of overall inequality in educational achievement that can be explained by predetermined *circumstances* (a lower-bound estimate of *ex-ante* IEOp).[16] Third, $\widehat{\theta}_{IEOp}$ is a relative measure of IEOp that is cardinally invariant to the standardization of test scores. One can decompose the IEOp measure into components for each variable in the *circumstances* vector which corresponds to a *Shapely-Shorrocks* decomposition.

3.3 Control Variables: Measuring Circumstances

Regarding the selection of relevant control variables, this study follows the most common approaches in the literature (e.g. Ferreira and Gignoux (2013)). The control variables represent *circumstances* which a student cannot influence, but which can determine the dependent variable of interest, cognitive skills, as measured by test scores. Moreover, applying new machine learning methods (Brunori et al., 2018), such as regression tree and random forest algorithms, the data confirm that my choice of control variables is sensible with respect to detecting relevant groups of *circumstances* (see Appendix A.5.2). Control variables can be divided into student-level *circumstances*, such as personal characteristics, and socio-economic family background variables, such as parental household characteristics. Table 1 provides an overview of the main control variables. Students are on average 15.43 years old. The share of female students is slightly greater than 50%. This reflects the steadily increasing female participation in *Gymnasium* (Prenzel et al., 2013). The variable *migration background* indicates that about 16.8% of students have at least one foreign-born parent. But the variable *language spoken at home* improves the extent to which one can control for the student's migration background. As depending on the level of parental integration, one can expect that not all students with migration traits speak a language other than German at home. Evidently, less than half of the number of students with foreign traits indicated that they speak a different language than German when talking to family members. I classify all individual characteristics (*gender*, *age*, *migration background*) as *circumstances*.

Another set of control variables involves socio-economic family background variables. An important *circumstance* is a student's parental education background which serves as an indicator for potential support opportunities available to the student. To measure parental education, I rely on the International Standard Classification of Education (ISCED) index. It indicates whether at least one parent has achieved

[15] The only caveat is that this model cannot estimate the effect of individual *circumstances*. As elements of $\hat{\rho}$ may be biased due to omitted variables, one cannot interpret them as causal effect of certain *circumstances* on scores.

[16] Niehues and Peichl (2014) outline how an upper-bound can be estimated in order to find boundaries for IOp estimates. But this method has not yet been widely applied because of data requirements (e.g. the need for panel data).

an academic degree, ISCED level 5 or 6, in which case they constitute an *academic household*. Table 1 shows that about 60% of students live in such households. As indicator for the socio-economic status (SES) environment in which a student grows up, I take the International Socio-Economic Index of Occupational Status (ISEI).[17]

[17]The International Standard Classification of Occupation (ISCO) serves as an alternative indicator for parental SES (Ganzeboom et al., 1992). It consists of parents' occupational data obtained from questionnaires, the responses to which were coded into ISCO codes. But it is not available for all PISA datasets, in contrast to the mapping of ISCO into ISEI indexes.

Table 1: Descriptive Statistics: Control Variables for *Circumstances*

Time Period (2003-2012)	**Mean**	**SD**	**Min-Max**	**Missings (SD)**
Individual Characteristics				
Female	0.5289	0.4989	[0-1]	0
Age in years	15.43	0.49	[13,75-17,25]	0
Language spoken at home *(Base: German)*	0.0552	0.2285	[0-1]	0.0060 (0.0774)
Migration background *(Base: German)*	0.1679	0.3738	[0-1]	0.0060 (0.0774)
Parental Characteristics				
Parental Education:				
(highest ISCED level)				
# ISCED-level (5-6):	0.6285	0.4832	[0-1]	
# ISCED-level (3-4) *(Base cat.)*:	0.2812	0.4495	[0-1]	0.0371 (0.1890)
# ISCED-level (1-2):	0.0532	0.2244	[0-1]	
Socio-Economic Status				
Number of books in a household:				
# more than 500:	0.2029	0.4022	[0-1]	
# 101-500 *(Base cat.)*:	0.4703	0.4991	[0-1]	0.0497 (0.2174)
# 11-100:	0.2579	0.4375	[0-1]	
# max. 10:	0.0193	0.1375	[0-1]	
Highest-ISEI-level of a job in the family	57.1536	17.2042	[0-90]	0.0177 (0.1317)
Family Characteristics				
Single parent households *(Base cat.: No)*	0.1317	0.3382	[0-1]	0.0808 (0.2726)
Father - employment status				
# full-time (FT) *(Base cat.)*:	0.8120	0.3907	[0-1]	
# part-time (PT):	0.0584	0.2345	[0-1]	0.0728 (0.2598)
# unemployed (UE):	0.0251	0.1564	[0-1]	
# out-of-labor force (OLF):	0.0318	0.1753	[0-1]	
Mother - employment status				
# full-time (FT) *(Base cat.)*:	0.2972	0.4570	[0-1]	
# part-time (PT):	0.4379	0.4961	[0-1]	0.0603 (0.2381)
# unemployed (UE):	0.0452	0.2078	[0-1]	
# out-of-labor force (OLF):	0.1593	0.3660	[0-1]	
Number of students	13,756	G-8 reform dummy: 0.4573 (0.4982)		

Notes: This table reports summary statistics for the sample of ninth graders in *Gymnasium* pooling the data for main period studied (PISA 2003, 2006, 2009, and 2012), weighted by the sampling weights provided in the PISA dataset (compare Appendix A.4.1). In the comments column, the amount of missing observations is provided and standard deviations are reported in parentheses. For categorical control variables, the base category is italicized. Finally, the number of observations and the G-8 reform dummy share is provided.

Higher ISEI scores correspond to higher levels of parental occupational status on a scale from zero to 90. Similarly, I use the *number of books at home* as a further control variable for socio-economic background. This variable is generated in all PISA studies and has been shown to be a good proxy for the family SES because household income is highly correlated with the amount of books in the household. It is plausible to assume that, at the age of 15, students are still financially dependent on their parents. Moreover, access to culture is mostly influenced by the opportunities offered in the household in which a child grows up. Thus, it is generally accepted that for students of age 15 the *number of books* variable represents *circumstances* that control for family SES. I take the range of 101-500 books as a base category for this variable because about 50% of students in the sample live in such a household.

As control for family structure characteristics, I consider whether a student lives in a single parent household which serves as an indicator for whether a student has grown up in a more stressful environment. About 13% of all students are raised under such *circumstances*. In addition, I also consider *employment status* dummies for both mother and father. By determining the time availability and family structure, aspects that influence the environment in which a student can study are considered. In the sample, most fathers work full-time (FT), whereas the largest share of all mothers is part-time employed (PT) (about 44%). This is consistent with the predominant family model in Germany during the 2000s consisting of the father as main bread-winner.

4 Empirical Strategy

Estimation proceeds in two steps. First, appropriate measures of IEOp need to be estimated given the available outcome and control variables in the data. Second, the quasi-experimental variation of the G-8 reform allows to identify the effect of increased learning intensity on IEOp by using a Difference-in-Differences (DiD) strategy based on forming reasonable treatment and control groups.

4.1 Estimating Inequality of Educational Opportunity (IEOp)

In a first step, IEOp is measured using $\widehat{\theta}_{IOP}$, as defined in Equation (7) in Section 3.2. This measure requires estimating the coefficient of determination (R^2) from an OLS regression of PISA test scores on the different *circumstances* variables that are listed in the previous section. The following regression model is estimated separately by federal states which form the respective treatment or control groups, and by PISA test wave:

$$\begin{aligned} Y_{ist} = \beta_0 + \beta_1(Individual\ Characteristics)_{ist} + \beta_2(Parental\ Characteristics)_{ist} \\ + \beta_3(Socio{-}Economic\ Status)_{ist} + \beta_4(Family\ Characteristics)_{ist} + FE(school)_{ist} + \varepsilon_{ist} \end{aligned} \quad (8)$$

where $Y_{ist} = \{stdpvread_{ist}; stdpvmath_{ist}; stdpvscie_{ist}\}$ are test scores of student i in state s at time t in one of three PISA domains. To ease the interpretation of β coefficients, I standardize scores for the effects to be measured as percentages of an international standard deviation in the PISA test.[18]

[18] Appendix A.4.1 provides details on the test metric. Until Section 5.2, I focus on the period (2003-2012) with the reform time set to take effect between 2006-2009, as defined in Section 4.2 (Appendix A.5.4). The regression model can also be estimated separately by treatment/control groups only twice for the pooled pre-reform ((2000-)2003-2006) and post-reform (2009-2012) samples.

This baseline regression model needs to be adjusted to take the following two issues into account. First, to allow for the extrapolation of findings to Germany's entire high school student population, the notion of external validity has to be considered (B. D. Meyer, 1995; Bertrand et al., 2004). This requires the data sample to be as representative as possible with respect to the student population in the ninth grade of *Gymnasium* in the time period under investigation (mainly 2003 to 2012). Thus, the model is estimated using a Weighted Least Squares (WLS) regression with the population weights provided in the data.[19] Second, the sampling strategy may induce some correlation among observations of the same unit (state/school). Therefore, I adjust regressions by calculating standard errors based on available replication weights in the PISA data and allow for clustering at the level of federal states, the level at which the reform has been implemented. Following the OECD guidelines in Appendix A.5.1, I explain how to estimate standard errors for the PISA data used.

As explained in Section 3.3, the control variables that measure *circumstances* in Equation (8) fall into four categories: Individual Characteristics (IC), Parental Characteristics (PC), Socio-Economic Status (SES), and Family Characteristics (FC) (Appendix A.5.3). Individual Characteristics include the *circumstances* variables *age*, *gender*, and *migration background*. As students were sampled based on attending the ninth grade, by controlling for *age*, differences in school entrance age (e.g. due to maturity) are taken into account. Controlling for *gender* considers the existence of any subject-specific differences in academic test score performance between male and female students (Niederle & Vesterlund, 2010). *Migration background* has also been shown to be important in explaining the academic achievements of students in Germany (Klieme et al., 2011). On average, having a migration background is negatively correlated with performance due to, for instance, its implications on non-cognitive skills such as self-esteem.

Socio-Economic family background control variables include Parental Characteristics such as *parental education* levels, SES indicators such as the *number of books in the household*, and Family Characteristics such as *family structure*. A more academically stimulating environment tends to have a positive impact on cognitive skill formation. In that regard, *parental education* can be assumed to constitute *circumstances* that capture investments into a student's early childhood. Similarly, a favorable SES as measured by higher ISEI index values and/or more books available in a household should have a positive impact on a student's test scores. Higher SES of the family in which a student grows up could be an indicator for better and easier access to support for dealing with school-related work. Otherwise, growing up with a single parent or unemployed parents might have a negative effect on test scores because such family conditions are associated with adverse factors for skill formation or limited access to out-of-school support opportunities.

In addition to control variables at the level of student i, the model in Equation (8) includes fixed effects (FEs) at the school level. First, adding school fixed effects allows me to capture quality differences among schools which can also exist within a federal state and to control for other school-level *circumstances*. Second, applying school fixed effects allows to control for characteristics both on the school and state

[19] Baumert and Prenzel (2008) discuss the PISA sampling strategy and the generation of population weights. They argue that, for the PISA-E data, certain student groups might have been over- or underrepresented, and that provided weights can be used to correct for this. These arguments also apply to the PISA-I data. Applying these weights in regressions allows deriving estimates which are representative of the German student population as the weights reflect the importance of each tested student given the population.

level because federal states oversee school policy. Moreover, as the PISA test is not conducted in the same schools over the years, school fixed effects are wave-specific. Thus, they capture year fixed effects when pooling before and after reform period. As a robustness check, a pooled version of Equation (8) is conducted using only fixed effects (FEs) at the state level. Then, state FEs consider time-invariant differences in the outcome variables between federal states due to, for instance, distinct political preferences for school policies neglecting differences between schools. The federal state in which a student attends secondary school represents a *circumstance* variable beyond a student's control because parents decide on where to reside. Although in theory, students may have some influence over which school they attend, their control is likely very limited at age ten. In fact, estimation results do not change much using either only federal state or only school FEs (which shows concerns on potential sorting at the school level not to be relevant). Consequently, it is sufficient to control only for school FEs in the main estimation specifications.

4.2 Definition of Treatment and Control Groups

The G-8 reform and its implementation at different points in time at the federal state level can be exploited as a quasi-experiment to identify the effect of increased learning intensity on a measure of IEOp. This requires categorizing the 16 federal states into treatment and control groups for each PISA test wave. Table 2 shows how useful treatment and control groups can be formed, based on the implementation of the reform and the timing of this process across federal states, and in this subsection, I explain which treatment/control group setting I consider to be the main specification for my DiD estimation approach. For seven out of fourteen states in which a reform took place, the introduction of the G-8 reform occurs between 2006 and 2009. Therefore, PISA 2009 is the first post-treatment wave of ninth graders tested in these states, and regression models including the 2012 wave capture the "medium-term" effect of the reform. Thus, I define the model covering period 2003 to 2012 as the *Model Base* (for an overview of treatment/control groups, see Appendix A.5.4).

Baseline Model Here, the reform takes effect in between 2006 and 2009. Table 2 shows that in this baseline model seven federal states can be classified as treatment group in which tested ninth graders were only in the G-8 model from 2009 onwards. These states belong to the *Treatment Group* **T2** which includes Baden-Württemberg (BW), Bavaria (BV), Lower Saxony (LS), Bremen (BR), Hamburg (HB), Berlin (BE), and Brandenburg (BB). However, the East German federal states are still likely to be different from the West German states. For instance, many teachers in East Germany were still educated in the former German Democratic Republic (GDR). Hence, for the main results, I focus on West Germany only, which means that the main *Treatment Group* **T** consists of BW, BV, LS, BR, and HB. Finally, excluding the city states of HB and BR, the most homogeneous *Treatment Group* **T1** consists of the three territorial West German states BW, BV, and LS. Together with **T2**, **T1** is used for robustness checks.

The control group in the main specification, *Control Group* **C**, consists of two territorial states in West Germany: Rhineland-Palatinate (RP) and Schleswig-Holstein (SH). These two states did not move to a G-8 model over the considered time period; that is, they always maintained a G-9 model. A second control group is made up of the two East German states of Saxony (SN) and Thuringia (TH). These two states had been following a G-8 model since 1949, when the former GDR was founded, and chose to maintain

Table 2: "G-8 reform" Treatment/Control Group Allocation of PISA Cohorts per State

Federal State	Reform Enaction	Double Cohort	Treated grade	PISA cohorts affected					(if) Treatment cohort/grade affected		
				2000	2003	2006	2009	2012	2006	2009	2012
Bavaria (BV)	2004/2005	2010/2011	6	first cohort treated in 6^{th} grade was not in 9th grade in a PISA test year							
	2004/2005	2011/2012	**5**	**C**	**C**	**C**	**T**(1)	**T**(1)	-	**1st** cohort	**4th** cohort
Lower Saxony (LS)	2004/2005	2010/2012	6	first cohort treated in 6^{th} grade was not in 9th grade in a PISA test year							
	2004/2005	2011/2013	**5**	**C**	**C**	**C**	**T**(1)	**T**(1)	-	**1st** cohort	**4th** cohort
Baden-Württemberg (BW)	2004/2005	2011/2012	**5**	**C**	**C**	**C**	**T**(1)	**T**(1)	-	**1st** cohort	**4th** cohort
Hamburg (HB)	2002/2003	2009/2010	**5**	**C**	**C**	**C**	**T**	**T**	-	**3rd** cohort	**6th** cohort
Bremen (BR)	2004/2005	2011/2012	**5**	**C**	**C**	**C**	**T**	**T**	-	**1st** cohort	**4th** cohort
Berlin (BE)	2006/2007	2011/2012	**7**	**C**	**C**	**C**	**T**2	**T**2	-	**1st** cohort	**4th** cohort
Brandenburg (BB)	2006/2007	2011/2012	**7**	**C**	**C**	**C**	**T**2	**T**2	-	**1st** cohort	**4th** cohort
Rhineland-Palatinate (RP)	2008/2009	2015/2016	**5**	**C**	**C**	**C**	**C**	**C**	-	-	-
Schleswig-Holstein (SH)	2008/2009	2015/2016	**5**	**C**	**C**	**C**	**C**	**C**	-	-	-
North Rhine-Westphalia (NRW)	2005/2006	2012/2013	**5**	**C**1	**C**1	**C**1	**C**1	**T**	-	-	**3rd** cohort
Saxony (SN)	since 1949		**5**	**Ch**	**Ch**	**Ch**	**Ch**	**Ch**	*hypoth. control-group: always treated*		
Thuringia (TH)	since 1949		**5**	**Ch**	**Ch**	**Ch**	**Ch**	**Ch**	*hypoth. control-group: always treated*		
Saarland (SL)	2001/2002	2009/2010	**5**	**C**	**C**	**T**	**T**	**T**	**1st** cohort	**4th** cohort	**7th** cohort
Saxony-Anhalt (ST)	2003/2004	2006/2007	9								
		2007/2008	8						**1st** cohort		
		2008/2009	**7**	-	-	**T**	-	-	7^{th} *graders*		
		2009/2010	6							**2nd** cohort	**5th** cohort
		2010/2011	5	**C**	**C**	-	**T**	**T**		5^{th} *graders*	5^{th} *graders*
Mecklenburg-West Pomerania (MWP)	2004/2005	2007/2008	9								
		2008/2009	**8**	-	-	**T**	-	-	**1st** cohort		
		2009/2010	7						8^{th} *graders*		
		2010/2011	6							**1st** cohort	**4th** cohort
		2011/2012	5	**C**	**C**	-	**T**	**T**		5^{th} *graders*	5^{th} *graders*
Hesse (H)[a]	2004/05	2011/2012	5	**C**	**C**	**C**	**T**	**T**	-	($\leq 10\%$)	4^{th} cohort
	2005/06	2012/2013	**5**	**C**	**C**	**C**	**C**	**T**	-	**1st** cohort	**3rd** cohort
	2006/07	2013/2014	**5**	**C**	**C**	**C**	**C**	**T**	-	-	**2nd** cohort

[a] Hesse (H) introduced the reform gradually across three school years (Figure 1 and Table A.1), thus it is neither treatment nor control group.

Notes: In this table, the treatment/control groups are highlighted by rectangular boxes.

For *Model Base* and *Model Robust*:

- Treatment **T** ≡ red box; **T1** ≡ magenta (inner) box and **T2** ≡ red + violet box
- Control Group (**C**) ≡ blue rectangle; **C1** ≡ blue + green rectangle.
- Moreover, TH and S form a hypothetical *Control Group (Ch)* (always G-8 model). Ch and C form the never-taker *Control Group (C-NT)*.

Note: An overview of treatment/control groups is given in Appendix A.5.4.

their secondary school system after reunification. They form a *hypothetical Control Group* **Ch** that could be interpreted as the counter-factual of a permanent G-8 model. Finally, one can form a *Never-Takers Control Group* **C-NT** consisting of the four states that never changed the length of *Gymnasium*: RP, SH, SN, and TH.

The most comparable setting for the baseline model consists of the *Treatment Group* **T** and *Control Group* **C** as it focuses on West German federal states that are very similar in relevant characteristics. Thereby, this setting still accounts for 40 out of 80.6 million people, i.e. 50% of the German population. Hence, it will serve as the main specification for the *Model Base*.[20] Focusing on a treatment that affects students in grade nine from 2009 onwards, five federal states belong neither to treatment nor control groups. In the first West German state that implemented the reform, Saarland (SL), ninth graders were already in a G-8 model by 2006. The same is true for the two East German states of Saxony-Anhalt (ST) and Mecklenburg-West Pomerania (MWP). Moreover, in both ST and MWP, the reform affected students from ninth grade onward, whereas in most other states, students were affected from fifth grade onward. In Hesse (H)[21] and North Rhine-Westphalia (NRW), ninth graders were only taught in a G-8 model since 2010, after the 2006-2009 window.

Robustness Model The *Model Robust* covers the time period 2003 to 2009 and thus considers the effect in response to the reform that is visible in 2009. This effect will be denoted the "short-term" effect of the reform. The treatment groups remain identical to those in the medium-term models (**T/T1/T2**) because only the year 2012 will be dropped in the short-term models with the reform time still set between 2006 and 2009. This also applies to the *Control Group* **C** consisting of RP and SH and to the *Never-Takers Control Group* **C-NT** including additionally SN and TH. Now, North Rhine-Westphalia (NRW) as federal state with the largest population in Germany can be added to the *Control Group* **C**: In the *Model Robust*, ninth graders in NRW were taught in a G-9 model over the whole time period (2000)/2003 until 2009. This creates *Control Group* **C1** consisting of RP, SH, and NRW. The most comparable setting for the robustness models consists of the *Treatment Group* **T** and *Control Groups* **C** or **C1**. With the latter group, I account for 57.6 out of 80.6 million people, i.e. 70% of the German population. Hence, there are two main control groups for the *Model Robust*.

4.3 Difference-in-Differences Estimation Strategy

The second step of the empirical strategy in this paper is a Difference-in-Differences (DiD) estimation. The gradual implementation of the G-8 reform across federal states allows estimating the reform-induced effect of increased learning intensity on IEOp by exploiting the differences between comparable treatment and control groups. For example, in the main specification of *Model Base*, there are five states in the treatment group (Baden-Württemberg, Bavaria, Lower Saxony, Bremen, Hamburg) and two states in the control group (Rhineland-Palatinate, Schleswig-Holstein). Moreover, the pre-reform years cover 2003-2006 ($before$), and 2009-2012 are the post-reform years ($after$). Then, the DiD strategy is implemented via the regression model:

[20] However, in Section 5.3, I also conduct robustness checks using T1, T2 and C-NT (Figure A.3 in Appendix A.1).

[21] Hesse (H) is the only federal state that did not implement the reform uniformly for *Gymnasium* at the start of one school year, but successively over three years as shown in Table 2. Thus, it is not possible to classify Hesse (H) either as treatment or control state in 2009 (without further assumptions) and it has to be excluded from estimations.

$$R^2_{st} = \delta_0 + \delta_1(TreatG8_{st} = after_t \times Treat_s) + \gamma_t \times after_t + \xi_s \times Treat_s (+ \alpha X_{st}) + \varepsilon_{st} \quad (9)$$

where $R^2_{st} = \{R^2(read)_{st}; R^2(maths)_{st}; R^2(science)_{st}\}$ is the estimated coefficient of determination (R^2) from Equation (8) associated with student i in state s in test year t that measures IEOp in the three PISA domains. *Treat* captures the *Treatment Group*-specific effect and $after$ the time trend. δ_1 is the coefficient of the interaction term, being 1 if a student attends a *Gymnasium* in a treatment state after the implementation of the G-8 model: it measures the causal effect of increased learning intensity on IEOp. δ_0 is a constant (*before* control mean), ε_{st} is the regression error term. Standard errors are calculated using replication weights following the method as explained in Appendix A.5.1 and are clustered at the federal state level. X_{st} is a vector of potential state-level variables. It is used to address concerns about differential implementation effects (e.g. school policies) on the level of federal states imposing the reform. For robustness checks, I adjust the regression procedure by including federal-state fixed effects capturing any effects at the highest level of variation that is not captured by the DiD group specific means in Equation (9). However, when the DiD approach is internally valid, results remain robust and the simple DiD specification (without X_{st}) is sufficient.

4.4 Selecting Appropriate Treatment/Control Group Settings

Internal Validity German federal states share a similar legislative and economic framework, and common qualification standards are coordinated by the SC. Thus, exploiting variation in the implementation process of the reform across states is more effective than relying on cross-national variation (Wössmann, 2010).

Next, one should consider whether the reform effect is driven not only by the explanatory variable of interest (increased learning intensity), but by other non-random factors in response to the reform. One concern with the DiD strategy might be that potentially affected students move with their families to a state that has not yet implemented the G-8 reform. If such reactions had occurred in a treatment group before the reform had been implemented, the population's composition across treatment and control groups might have changed in a way that would bias estimation results.

However, such anticipatory behavior is very unlikely. First, options for moving between federal states to avoid the G-8 reform were limited. The implementation across all federal states was fast: Half of all reform states started the transition into shortened duration of the *Gymnasium* within three school years (2003/2004 until 2005/2006). There is no systematic pattern regarding the timing and implementation of the G-8 reform and the geographical location of reforming federal states.[22] Second, direct and indirect moving costs, including bureaucratic hurdles, have been shown to be reasons why only a few families with children of school age move to another federal state in Germany (Bundeszentrale für politische Bildung, 2008). Families tend to move more between municipalities than states. Third, strategic considerations concerning the competition for access to study programs also support the assumption that bias due to movement between states is unlikely. As a result of the reform, it was obvious that several *double cohorts* would graduate in between 2009 and 2016. This temporary increase in the number of applicants for university studies could inversely affect the probability of students to quickly enter a study program of their choice. Hence, G-8 model students could at least insure themselves against the risk of having to take

[22] The geographical maps in Figure A.2 in Appendix A.1 reveal the quick spread of the treatment across states.

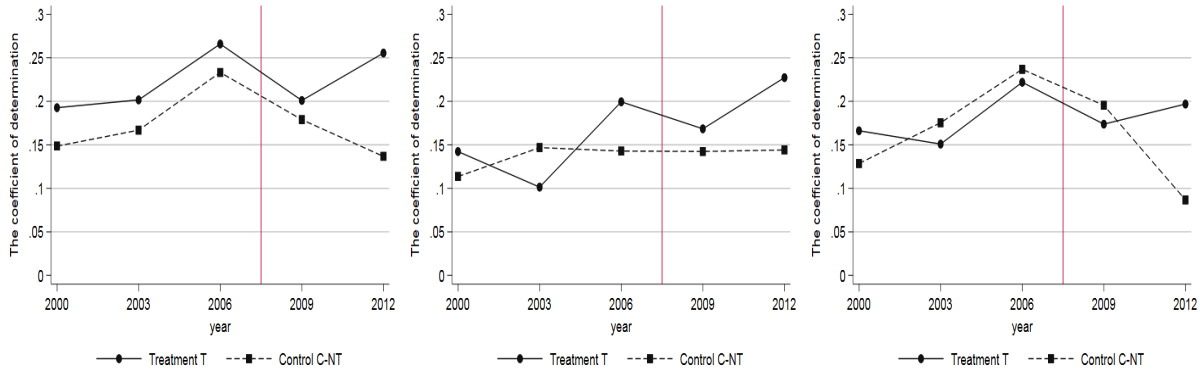
a gap year as their 14^{th} year of education. Instead of spending 13 years in school and having to wait one additional year before entering the study program of their choice, having 12 years of schooling before enrolling at a university means that even after one gap year, G-8 model students could "save" one year compared to G-9 model students completing a gap year.

By focusing on a setting in which treatment states implemented the reform in school year 2004/2005, the quasi-experimental design is also unlikely to suffer from estimation bias due to non-random political reasons for introducing the G-8 reform slightly earlier or later among federal states. Appendix A.5.5 shows that treatment/control groups are similar regarding the stability of state governments in charge of school policy: Political preferences remain stable over the analysis period. Moreover, no systematic change in the transition flows between secondary school tracks is observed due to the G-8 reform (Huebener & Marcus, 2017).

Finally, the internal validity of a DiD estimation requires the common time trend assumption to hold: without the reform, both treatment and control group would have shown a parallel time trend. This can be confirmed by examining the pre-reform trends in terms of the estimated IEOp measure for treatment and control groups in Figure 2. Moreover, I conduct placebo tests (Bertrand et al., 2004) as robustness checks in Section 5.3.

Treatment/Control Group Comparison Due to the quasi-experimental design of the G-8 reform, estimating the effect of the reform on IEOp should not be biased by any selection of students based on pre-reform characteristics. As the identification strategy relies on comparing the change in IEOp for ninth graders attending *Gymnasium* across treatment and control groups before and after the reform, many significant observable pre-reform differences in the control variables might weaken the empirical strategy.

Figure 2: Robustness - DiD Graphs of IEOp measure for enlarged Treatment/Control Groups



(a) IEOp measure based on maths **(b)** IEOp measure based on reading **(c)** IEOp measure based on science

Notes: This figure shows the DiD graphs for IEOp measures based on all three test domains for all available PISA years (PISA 2000 is included for robustness reasons). It confirms that the parallel trend assumption holds. Five (Treatment Group T) federal states are compared to the never-changing control group (C-NT) consisting of four states. Compare also Figure A.6 and A.7 for other specifications showing that trends are not sensitive to alternative compositions of the treatment group and Figure A.5 in Appendix A.1. As discussed in Section 3, the data used for the main regressions cover the time frame 2003 to 2012.
Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009, and 2012.

Following Imbens and Wooldridge (2009), Table A.4 shows standardized means comparison tests for the control variable sets (Table 1) concerning all treatment groups and the main control group **C**.

For the baseline model, *Model Base*, the G-8 reform takes effect between 2006 and 2009. Hence, PISA waves 2003 and 2006 constitute the pre-treatment period. Table A.4 shows that treatment and control groups have similar characteristics in terms of the main *circumstances* variables used for the analysis. Moreover, apart from small differences in the level of *circumstances* variables, the pre-reform comparison of groups **T** and **C** are robust. This supports the internal validity of the strategy because the main treatment and control groups consisting of West German states turn out to be comparable. Using smaller or enlarged treatment groups the pre-reform comparison tests are still robust in combination with the standard control group (**C**).[23]

In summary, the pre-reform sample means comparison test for the main control variable set (Table A.4) suggests that the DiD estimation approach outlined in Section 4.3 is internally valid. This is true at least for *Model Base* to compare *Treatment Groups* **T/T1/T2** versus *Control Group* **C** and for *Model Robust* to compare *Treatment Group* **T** versus *Control Group* **C1** (see Table A.5).

5 Results and Discussion

When presenting the results for the outcome variables, PISA test scores in each of the three domains, the respective five plausible values are standardized based on the distribution of test scores across the sample of students attending the ninth grade of *Gymnasium* that are taken from the representative *grade-based* PISA data sets (Section 3.1).[24] Section 5.1 explains the first-step, Section 5.2 the second-step results for the baseline model specifications (Section 4.2). Section 5.3 provides robustness checks with extended treatment and control group settings, while Section 5.4 rationalizes the results.

5.1 First-Step Result: Inequality of Educational Opportunity Measure

The first step of analyzing the distributional effects of increased learning intensity involves deriving the main outcome variable, the measure of IEOp as share in the standardized PISA test score variance that can only be attributed to observed *circumstances* (Equation (8) in Section 4.1). All six sets of control variables that capture *circumstances* are jointly used to derive this IEOp measure.[25] Its standard errors are obtained by using replication weights and clustering on the highest level on which the reform was implemented (Bertrand et al., 2004), the federal state level. Finally, population weights consider the stratified data structure and representativeness of each observation (Appendix A.5.1).

When estimating IEOp, it is useful to check how *circumstances* variables directly affect cognitive skills as measured by test scores. Detailed regression output per test domain is provided for the main specification

[23] This supports the internal validity of the estimation strategy: see pre-trend graph in Figure A.7 in Appendix A.1.

[24] For the remainder of this paper, I restrict the presentation of *first-step estimation* results to test scores that are standardized with respect to the pooled sample of all students in *Gymnasium* that are part of the representative *grade-based* PISA test cohort in any of the test years that form the sample (2003, 2006, 2009, 2012 in *Model Base*) (*stdpvsubject3*): This allows me to interpret the coefficients relative to the average student performance over the sample period.

[25] In Section 5.3 for robustness check purposes, for all main specifications and each test domain, all results are shown adding step-by-step control variables (covering *circumstances*): from (i) and (ii) constituting control set (I) until (VI) encompassing controls (i), (ii), (iii), (iv), (v), (vi), (vii). See Appendix A.5.1 for details on computing standard errors.

Model Base: **T** versus **C** (Table A.6, A.7, A.8).[26] The following patterns can be observed concerning how the *circumstances* variables (as defined in Section 3.3) affect test scores. The only control variable changing the direction of its effect on achievement scores depending on the test domain is *gender*. Being female decreases a student's achievement in the PISA mathematics test by 45-65% and in the science test by 30-50% in terms of an international standard deviation (SD). The effect size slightly declines in the post-reform period across both treatment and control group. However, female students increase their reading performance by up to 40% in terms of one international SD. This is consistent with the literature on gender-specific achievement differences in educational test outcomes (Niederle & Vesterlund, 2010).

All the other control variable estimates are robust in their signs independent of the test domain. As expected, the *age* effect is negative. Those who started school at an older age or had to repeat a grade before entering the ninth grade will be older compared to their peers due to factors correlated with below-average performance in test scores. Similarly, having a *migration background* is associated with performing worse in all three testing domains. Additionally controlling for whether a *foreign language is spoken at home*, the negative effect shrinks as expected. Thus, the degree to which a migrant student experiences integration to the host country's standards on a daily family life level, seems to be key for test scores, in particular for the domain of reading.

Regarding the socio-economic status (SES) of the household in which a student grows up, a higher *amount of books* than the base category (101-500) is positively correlated with test scores. Likewise, the higher the ISEI index of a parental job in the family, the higher is the positive effect on scores.[27] Thus, the SES control variables tend to match the literature suggesting that higher family SES correlates with beneficial conditions for early childhood development. *Parental education* is also indicative for academic support opportunities, and indeed a positive impact on test scores for both mathematics and science can be found for the variable indicating that a student grew up in an *academic household* (at least one parent with ISCED level 5-6). The effect is less important for reading. As mathematics and science are subjects likely requiring more specific and targeted knowledge from parents for them to be able to support their children, this may explain the difference.[28] But Parental Characteristics have less effect on scores once individual *circumstances* are considered. Finally, *family structure* and *employment status* show no clear patterns.

In summary, *first-step* regressions demonstrate that in the medium-term, most of the *circumstances* variables affect the PISA test scores in the expected directions. The fact that these patterns are consistent over varying time horizons and across PISA data sets confirms that the chosen *circumstances* variables were appropriately selected (compare Appendix A.5.2). Furthermore, the explanatory power of these *first-step* regressions remains in a range of 15-35% across the different specifications. Thereby, IEOp

[26] Table A.6 shows the *first-step* results for reading test scores, Table A.7 for mathematics test scores and Table A.8 provides the corresponding output for science test scores. In each table, the columns (1) to (2) refer to *Control Group* **C**, columns (3) and (4) to *Treatment Group* **T**. Within both *Groups*, the first column refers to the "*Before*" reform period (2003-2006), the second even numbered one repeats regressions using only "*After*" reform (2009-2012) data.

[27] With the average family's highest job ISEI index being 58, an effect on test scores of 0.001 translates into 5.8% of a PISA international test standard deviation. See also Section 3.3 and Appendix A.5.3 for further explanations.

[28] Furthermore, highly educated parents might be more aware of the greater importance of numeracy skills for labor market outcomes. However, the effects of growing up in an *academic household* are rather insignificantly positive, whereas those of growing up in less educated families are rather significantly negative for test scores.

measures tend to be higher when measured with respect to mathematical and scientific skills (20-35%) than reading literary (15-25%). Consequently, the level of the IEOp measure found in this paper can be categorized as a lower bound within the range of few available IEOp estimates for European countries. For instance, Ferreira and Gignoux (2013) find that about 35% of test score variation in PISA-I 2006 can be attributed to *circumstances* for the case of Germany, and Carneiro (2008) finds that IEOp amounts to about 40% in the case of Portugal.

5.2 Main Results: The Effect of Increased Learning Intensity on IEOp

In this section, I switch to the *second-step* of the estimation approach, the DiD framework. The IEOp measure derived above by the *first-step* regressions is the share of total variance in test scores which is accounted for by the student's predetermined *circumstances* variables.

Baseline Model Results Starting with the main treatment and control group specification, the *Model Base* results are shown in Table 3. The top panel outlines DiD estimates for reading, the middle panel for mathematics and the bottom panel for science test scores. IEOp is calculated with school fixed effects. The DiD table illustrates that the change in IEOp as measured by the R^2 in the *first-step* estimation exhibits a common pattern across all three test domains - IEOp has increased due to the G-8 reform. That is, the share of inequality in test scores that can be attributed to *circumstances* has risen. With the estimate being a lower bound of the true IEOp the results can be interpreted as follows. At least about 10% of the variation in reading test scores can be additionally attributed to *circumstances* beyond the control of a ninth-grade student. For mathematics, at least about 14% and for science at least about 18% of the test score variation can be additionally considered to constitute IEOp. These results are statistically significant, with standard errors computed as explained in Appendix A.5.1. Thus, given initial values of 20-30% in IEOp, DiD estimates would correspond to a relative increase in IEOp of at least 25% in response to the rise in learning intensity induced by the G-8 reform. Hence, the increase in IEOp is economically significant. The effects are stronger when IEOp is measured with respect to science or mathematics than with respect to reading test scores.

Going into further detail, one notes that IEOp seems to have considerably decreased in the time period after the reform for Control Group **C**. Instead, for Treatment Group **T**, the level of IEOp appears to have remained practically static across all three domains. In this setting, the increase in learning intensity appears to have maintained the role of *circumstances* in treated states, while IEOp tends to have decreased without shorter school duration. The *Model Base* takes a medium-term perspective as not only the first affected cohorts are taken into account, but data up to 2012 are considered, when the reform had already been fully enacted. By 2012, in most federal states, the *double cohort* had already graduated or was about to graduate (Figure 1).

Robustness Model Results To learn about the robustness of the effects, it is useful to see how results change for the main treatment and control group specification when conducting the same two-step estimation procedure for the *Model Robust* covering only years 2003 until 2009. Therefore, the left-hand panels in Table A.9 in Appendix A.2 show the short-term effects of increased learning intensity on IEOp focusing mainly on the first student cohorts treated by the G-8 reform for *Treatment Group* **T** versus *Control Group* **C**. The DiD estimates remain positive across all test domains. However, the increase in

IEOp only reaches levels that rest within a range of about 5-10% of the variance in educational test scores that can be additionally attributed to *circumstances*. However, results are no longer statistically significant at the 5% level. Thus, the relative deterioration in IEOp is lower in the short term - if different from zero at all - compared to its significant size in the medium term (Table 3). Otherwise, the underlying patterns of the reform effect also remain robust in the short term. Educational acceleration tends to inhibit students in the treatment group from experiencing any improvements in IEOp. Instead, ninth graders in the control group experience less IEOp as *circumstances* lose explanatory power for academic achievement. To understand how the G-8 reform changed educational opportunities in *Gymnasium*, it is useful to expand the robustness model to consider treatment and control group specifications that bear even more external

Table 3: Main Results for T vs. C

Subject	**IEOp measured as R^2**			**IEOp measured as R^2 adj.**		
Reading	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before	0.242	0.180	-0.062	0.172	0.154	-0.018
	(0.057)	(0.031)	(0.065)	(0.062)	(0.032)	(0.070)
After	0.162	0.213	0.051	0.114	0.192	0.078
	(0.034)	(0.020)	(0.039)	(0.036)	(0.021)	(0.041)
Change in R^2	-0.080	0.033	**0.113**	-0.058	0.037	**0.096**
	(0.066)	(0.037)	(0.076)	(0.072)	(0.038)	(0.081)
Mathematics	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before	0.353	0.267	-0.086	0.294	0.245	-0.049
	(0.060)	(0.033)	(0.068)	(0.065)	(0.034)	(0.073)
After	0.190	0.249	0.060	0.143	0.229	0.086
	(0.040)	(0.027)	(0.048)	(0.043)	(0.027)	(0.051)
Change in R^2	-0.164	-0.018	**0.146**	-0.151	-0.015	**0.136**
	(0.072)	(0.042)	(0.083)	(0.078)	(0.043)	(0.089)
Science	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before	0.363	0.215	-0.148	0.304	0.190	-0.114
	(0.052)	(0.025)	(0.058)	(0.057)	(0.026)	(0.063)
After	0.173	0.210	0.037	0.125	0.188	0.063
	(0.048)	(0.023)	(0.053)	(0.051)	(0.023)	(0.056)
Change in R^2	-0.190	-0.005	**0.185**	-0.179	-0.002	**0.177**
	(0.071)	(0.034)	(0.079)	(0.077)	(0.035)	(0.084)

Notes: Table entries are R^2 measures of IEOp (Equation (7)). Robust standard errors are in parentheses and were calculated using replication weights following the method as explained in Appendix A.5.1, clustering at the federal state level. DiD results are estimated according to Equation (9) taking into account population weights. Positive changes in R^2 indicate increasing IEOp or decreasing EEOp and vice versa for negative changes.
Background variables used to derive R^2:
(i) individual characteristics (IC) I: *age* and *gender*
(ii) individual characteristics (IC) II: *language spoken at home*; *migration background* (based on (parental) birth place)
(iii) parental characteristics (PC): *highest parents' qualification (ISCED-level 1-2/ISCED-level 3-4/ISCED-level 5-6)*
(iv) socio-economic status (SES) I: *number of books in household (max. 11, 11-100, 101-500, more than 500)*
(v) socio-economic status (SES) II : *highest ISEI-level-index [0-90] of job in the family*
(vi) family characteristics (FC) I: *family structure - growing up in single parent household?*
(vii) family characteristics (FC) II: *mother/father: working part-time (PT) - unemployed (UE) - out of labor force (OLF)*
Compare: The first-step regressions of the setting: treatment group **T** vs. control group **C** are provided in Table A.6, A.7 and A.8 in Appendix A.2. For graphical evidence on the DiD treatment/control groups, see Figure A.6 and A.7 in Appendix A.1.
Source: Author's calculations based on PISA 2003, 2006, 2009, 2012 (compare Section 3.1).

validity for the German school system. With **C1** about 70% of the German high school student population can be considered in the short-term reform analysis. DiD results for this extended treatment and control group specification are shown in the right panel of Table A.9. There appears to be no effect on IEOp across all three test domains in response to the G-8 reform. However, the IEOp measures still range between 15 to 25%, their magnitude increasing from reading to mathematics to science. Students in both treatment and control group experience similar rise in IEOp, such that in total the DiD effect is canceled out. The DiD estimation findings on the effect of the G-8 reform are similar across **T/C** and **T/C1** specifications: There is no statistically significant short-term effect of the reform-induced increase in learning intensity on IEOp.

In summary, the impact of the reform on IEOp is robust for the alternative specification. Focusing on *Model Robust* (2003-2009), increased learning intensity does not affect IEOp, that is, unfair inequality in terms of how much in the cognitive test score variation can be explained by *circumstances* beyond a student's control (right panel in Table A.9 in Appendix A.2). Narrowing the control group to include only federal states that did not plan to shorten the duration of their G-9 model *Gymnasium*, a considerable increase in IEOp of about 5-10% in terms of additional explanatory power is observable also in *Model Robust* setting, but results are barely statistically significant (left panel in Table A.9). However, taking a medium-term perspective on the G-8 reform (*Model Base* (2003-2012)) shows that the reform-induced increase in learning intensity causally increases the IEOp measures (Table 3). The observed rise in inequality of opportunity is statistically significant and covers at least 25% of the general IEOp measure estimated for students attending German secondary schools. Results reveal that for students in *Gymnasium*, the lower bound levels of IEOp correspond to about 17-35% of the variance in educational outcomes that can be attributed to the role of *circumstances* only.[29] Thus, the main results show that increased learning intensity aggravates IEOp. The effects are stronger when measured for mathematics and science than for reading.

5.3 Robustness Checks

Placebo Test To evaluate the plausibility of the quasi-experimental identification strategy that allows a causal interpretation of the effects of the G-8 reform-induced increase in learning intensity on IEOp, it is important to conduct Placebo Tests (Bertrand et al., 2004). Setting the reform to artificially take effect between 2003 and 2006, no statistically significant effects can be detected for any of the main treatment and control group specifications (**T** vs. **C** in Table A.10 in Appendix A.2). In addition to the pre-reform comparison test (Section 4.4), this finding supports the internal validity of the estimation strategy, in particular that the common time trend assumption holds. This can also be seen from examining the pre-reform trends in terms of the estimated IEOp measure for the main treatment and control groups in Figure 2 in Section 4.4. Thus, Placebo Tests confirm the plausibility for interpreting the main estimation results as causal effects of the reform on IEOp.

[29] To check whether this increase in IEOp is long-lasting, one should consider longer time periods, data which are not yet available. However, once shifting attention to cohorts long after the first treated ones, potential new curricular reforms undertaken in response to the initial G-8 reform (Table A.1) should be taken into account. Instead, it is plausible to assume that medium-term effects on IEOp as defined in this paper are long-lasting given the literature on the persistence of education on lifetime outcomes (Deming, 2009).

Moreover, multi-level regressions confirm that school level *circumstances* are indeed already considered by school fixed effects. Furthermore, using school fixed effects or only federal state effects to measure IEOp does not change DiD results (Table A.11 in Appendix A.2). This indicates that sorting based on schools is not a concern, which also corroborates the internal validity of the empirical strategy taken.

To further investigate the robustness of my main results, I focus on three margins of interest. First, I analyze how findings change depending on which of the available six control variable sets are included in the *first-step* regression for deriving the IEOp measure. Second, I focus on how DiD results change when extending or reducing the treatment group. Third, I show how results change for enlarged control groups consisting of states that never changed their academic track.[30]

Varying the Control Set of Circumstances Variables To understand how robust DiD results remain when changing the amount of control variables chosen to cover predetermined *circumstances*, I analyze how *adjusted* R^2 measures of IEOp behave in particular. The *adjusted* R^2 can help detect which *Control set*[31] combination appears to have most explanatory power among the available *circumstances* variables (Table 1).

Looking across the DiD result tables, including as *circumstances* variables Individual Characteristics (IC), Parental Characteristics (PC) and Socio-Economic Status (SES) may be optimal among the six control variable sets. However, the analysis across different sets reveals that, for each test domain, the final reform estimate of increased learning intensity on IEOp does not change much across *Control sets* 3 to 6 (see Table A.12). This also provides support for the empirical strategy taken to derive the main results: Using all six variable sets in the *first-step* regression. In fact, this approach renders estimates that correspond to the highest *adjusted* R^2 generating *Control set* combination. Moreover, regression patterns stay robust in size and direction independent of which set is used to derive IEOp. This is evidence for the quasi-experimental design assumption that assignment to treatment occurred without selection on observables, but randomly.

Extending Treatment Groups Next, it is useful to repeat the estimations with extended treatment groups to investigate the potential external validity of the main results. Therefore, all main regressions (Section 5.2) are rerun with *Treatment Group* **T1** excluding the two West German city states Hamburg and Bremen, and for *Treatment Group* **T2**, which is **T** plus Berlin and Brandenburg. When the treatment group gets larger, on average DiD reform effects become smaller, for instance, in the regression settings with *Control Group* **C** (Table A.12), the increasing effect on IEOp declines as we move from **T** to **T2** consistently within each test domain and across all *Control sets* . In summary, despite their increasingly heterogeneous composition, the main results in terms of direction and size are reconfirmed. This supports the potential external validity of the results based on the carefully chosen ***T/C** Group* specification in the previous section. Thus, focusing on the *Treatment Group* **T** does not mean that results do not carry implications which are likely to be valid for the entire German secondary school system.

[30] The main output tables for robustness checks are shown in Appendix A.2: Table A.12 to A.13. All tables are structured in the same way to provide an overview of DiD estimation results of increased learning intensity as induced by the G-8 reform on IEOp.

[31] *Control set 1* provides results based on deriving the IEOp measure including only Individual Characteristics (IC) as control variables (that is (i) and (ii) in Appendix A.5.3). Subsequently, additional control variables are added until in set 6 all available *circumstances* are applied together in the *first-step* regression.

Extending Control Groups As mentioned in Section 4.2, one can also compare treatment groups with states that always maintained the same length for *Gymnasium*. When using *Never-Taker Control Group* **C-NT** , the DiD results in all specifications show a smaller increase in IEOp. The results for this specification can be seen in Table A.13. If one takes the complementary part of **C-NT**, that is, the hypothetical control group consisting of Saxony and Thuringia, the effects are rather slightly negative, but barely significant.

5.4 Discussion and Interpretation of Results - Potential Mechanisms

To begin with, the key concept of IEOp in this paper is closely related to the issue of social mobility. Estimating $\widehat{\theta}_{IEOp}$ can be regarded as isomorphic to measuring intergenerational persistence of IEOp. For the latter, following Galton, one usually regresses a child's (y_{it}) on parental outcomes ($y_{i,t-1}$):

$$y_{it} = \beta y_{i,t-1} + \varepsilon_{it}, \tag{10}$$

with β as measure of persistence. If one used family background variables instead of parental outcome variables for ($y_{i,t-1}$), then the R^2 measure of immobility (Equation (10)) would be similar to $\widehat{\theta}_{IOP}$ (Equation (7)) as long as the *circumstances* vector contains mostly family background variables. Thus, $\widehat{\theta}_{IEOp}$ is connected to measures of intergenerational educational immobility, which are used to measure social (im)mobility (as β).

In analogy, this is related to the findings that childhood wealth can serve as a proxy for *circumstances* explaining future wealth inequality (Boserup, Kopczuk, & Kreiner, 2018). Moreover, intergenerational income elasticity and the Gini coefficient of income have been shown to be highly correlated (*Great Gatsby Curve*) which points to a link between IEOp and intergenerational social mobility (Black & Devereux, 2011). The connection between both concepts can be characterized by two adjoint forces, *upward* and *downward* social mobility. A decrease in IEOp would be indicative for improved *upward* mobility, as it means that *circumstances*, such as the SES of the family in which one grows up, became less important for a student's academic performance. Therefore, if lower IEOp translates into providing more equalizing learning conditions such that ability, but in particular *efforts*, are rewarded, extending EEOp would be welfare enhancing in a society with meritocratic preferences. While decreasing IEOp may lead to social *upward* mobility for high-performing students from disadvantaged backgrounds, it may also lead to social *downward* mobility for students with beneficial *circumstances* who lack talent and/or *efforts* to maintain their position as soon as *circumstances* were less important for a student's educational outcome.

Table 4 presents the results for a DiD based on standardized scores for the main setting 2003-2012. Overall, I find no significant effect of the reform on scores.[32] Again, the effect size for mathematics/science is stronger than for reading. The DiD with SES interaction terms delivers highly significant positive

[32]Please note that a direct comparison of Huebener et al. (2017), Andrietti and Su (2019) and this work should be done with caveat. This can be attributed to at least two aspects that distinguish my study from the other two authors: First, as explained in Section 4.2, I focus on a much finer setting concerning the division of states into treatment and control groups in order to only compare states to each other that experienced a similar treatment effect size. Including as many federal states as possible for the DiD (like Huebener et al. (2017); Andrietti and Su (2019)) implies mixing heterogeneous treatment effects (compare Table 2). Note that my results remain robust when extending the setting to more states (see Section 5.3), however, I refrain from bunching together all states because then very different treatment effects are mixed making the interpretation of results difficult. Second, the data used in this work only encompasses PISA-I data sets, as they can be combined much more safely than mixing PISA-I and PISA-E because the associated population weights are not really comparable across both types of datasets (Section 3.1).

coefficients indicating that students from a higher social background have improved more, hinting to an increase in IEOp. Intuitively, the starker distinction between low and high SES based on the first compared to the fourth quartile instead of using the median as cut-off leads to stronger results (column (ii) vs. (iii) in Table 4). It is worth noting that SES status appears more important than growing up in an academic household. As with the DiD on IEOp (measured by R^2 or $adjusted\ R^2$), results are weakest for reading, stronger for science, and most striking for mathematics, which confirms the findings of Section 5.2.

Table 4: Score-DiD with Interaction Terms

Subject	(i) Basic	(ii) SES Median	(iii) SES Quartiles	(iv) Academic
Reading				
Treated*post	-0.174	-0.214	-0.137	-0.198
	(0.290)	(0.291)	(0.319)	(0.288)
Treated*post*SES	-	0.072***	0.157***	-
		(0.0274)	(0.0389)	
Treated*post*academic	-	-	-	0.0335*
				(0.0195)
R^2	0.130	0.131	0.148	0.132
Mathematics				
Treated*post	-0.300	-0.382+	-0.322	-0.393+
	(0.262)	(0.264)	(0.398)	(0.261)
Treated*post*SES	-	0.148***	0.284***	-
		(0.0315)	(0.0386)	
Treated*post*academic	-	-	-	0.127***
				(0.0218)
R^2	0.167	0.172	0.199	0.173
Science				
Treated*post	-0.309	-0.384*	-0.382*	-0.402*
	(0.230)	(0.231)	(0.220)	(0.228)
Treated*post*SES	-	0.135***	0.261***	-
		(0.0306)	(0.0414)	
Treated*post*academic	-		-	0.127***
				(0.0214)
R^2	0.138	0.142	0.162	0.143
Observations	6,649	6,630	3,208	6,483

Notes: Robust standard errors are shown in parentheses and significance levels are indicated by: *** 1%, ** 5%, * 10%, + 15%. Table entries show the results of a DiD of the reform's effect on standardized test scores in all three test domains. Columns *(ii-iv)* show the results of a DiD, as well as its interaction with background variables. Column *(ii)* shows a distinction between high and low SES using the median of the highest ISEI in the familiy, whereas column *(iii)* displays results where the first quartile according to ISEI is assigned low SES and the fourth quartile is assigned high SES. Column *(iv)* displays interaction results for academic, a dummy variable taking value 1 if mother or father achieved a university degree and 0 else. All regressions have been conducted with *school* fixed effects and using control set $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$.
Source: Author's calculations based on PISA 2003, 2006, 2009, and 2012.

Table 5: Tuition DiD-Results

	(i) Basic	(ii) SES Median	(iii) SES Quartiles	(iv) Academic
Treated*post	-0.103	-0.113	-0.0715	-0.114
	(0.121)	(0.122)	(0.192)	(0.122)
Treated*post*SES	-	0.0172	0.0454*	-
		(0.0175)	(0.0261)	
Treated*post*academic	-	-	-	0.0149
				(0.0182)
Observations	5,852	5,843	2,821	5,781
R-squared	0.269	0.269	0.297	0.272

Notes: Robust standard errors are shown in parentheses and significance levels are indicated by: *** 1%, ** 5%, * 10%, + 15%. Table entries show the results of a DiD on private tuition. Columns *(ii-iv)* show the results of a DiD and its interaction with background variables. Column *(ii)* shows a distinction between high and low SES using the median of the highest ISEI in the familiy, whereas column*(iii)* displays results where the first quartile according to ISEI is assigned low SES and the fourth high SES. Column *(iv)* displays interaction results for academic, a dummy variable taking value 1 if mother or father achieved a university degree and 0 else. All regressions have been conducted with *school* fixed effects and using control set $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$.
Source: Author's calculations based on PISA 2003, 2006, 2009, and 2012.

Table 5 depicts the same DiD regression as above but now using extra tuition as the dependent variable. The graphs presented in Figure A.8 in Appendix A.1 justify the assumption of common pre-trends for private tuition, thus giving leverage to this regression. Whereas the basic DiD delivers a slightly negative though insignificant effect of the reform on private tuition, the interaction-term coefficients all carry a positive sign. Admittedly, only the interaction with SES based on highest and lowest quartile is significant at the ten p-value percentage level. Nonetheless, all of them point into the same direction: Families of higher socio-economic background react to the reform by actively providing their children with extra tuition. This suggests that disparities in private tuition between different social backgrounds are among the main drivers of the rise in IEOp induced by increasing learning intensity due to the G-8 reform.

Returning to the G-8 reform, one can provide the following explanation for the observed findings. First, the fact that increased learning intensity had only a limited impact on IEOp in the short run may be indicative for the reform heterogeneously promoting both *downward* mobility among students with advantageous *circumstances* and *upward* mobility among those with disadvantaged *circumstances* who - having managed to enter the *Gymnasium* - may have already undergone a harder selection process.[33] As the implementation process of the reform suggests, the reform-induced increase in learning intensity affected students and their parents by surprise in a manner that they could not adapt to immediately. For instance, being the first one confronted with the newly intensified system, it is harder to adapt as one cannot easily rely on the experiences of older students as was the case for later cohorts in the new G-8 model. This may explain why IEOp increased only moderately or not at all in the short term. Thus, in the initial reform period, the lag with which favorable *circumstances* adapt to help a student implies that *downward* rather than *upward* mobility forces were more relevant for the first affected student cohorts.

[33]The high correlation of parental education and a student's probability of entering the *Gymnasium* has been shown (e.g. Klieme et al. (2011)) to be persistent in the German school system at least over the last two decades.

Second, in the medium term, after favorable *circumstances* had time to adapt and provide support to the associated students, both *upward* mobility and *downward* mobility would be lessened. For instance, parents are more likely to be aware and prepared to deal with the increased requirements of a G-8 model and new forms of additional professional tuition services may become available in response to the reform based on the experiences of the first affected cohorts. Consequently, favorable *circumstances* may then allow students quicker, easier, and better access to a support system helping them deal with the higher learning intensity. Then, increased IEOp associated with lower *upward* rather than higher *downward* mobility may be expected in the medium term after the G-8 reform was enacted. Descriptive evidence on the evolution of additional, paid tuition for students attending a *Gymnasium* available from PISA questionnaires supports the explanation given above (cf. Figure A.8). There has been a rise in extra tuition following the reform, with this effect being stronger in the treatment compared to the control group.

Additionally, the increase in extra tuition has been more pronounced for students from more privileged family environments (*circumstances*), such as those living in academic households (Table 4). This trend is confirmed by Klemm and Hollenbach-Biele (2006), who analyzed the evolution of private tuition in the same time period using representative survey data for Germany. Moreover, looking at the medium-term effect evidence (Table 3), DiD estimates of the effect of increased learning intensity on IEOp reveal some subject-related patterns. The level of IEOp is consistently higher for mathematics/science than reading across all treatment and control group specifications. This observation can be interpreted as evidence in favor of the existence of heterogeneous subject-dependent curricular flexibilities. In fact, reading skills comprise more general competencies that are not only learnt in language-related courses at school, but also in other classes and everyday life, reading often being a necessary prerequisite to comprehend, learn, or interact with other people.

Consequently, variations in learning intensity might have less influence on reading skills. In contrast, mathematics/science can be regarded as requiring more specific skills accumulated through taught courses at school rather than learnt indirectly, for instance in everyday life. Thus, for the complementary skill set required by mathematics/science, it seems plausible that positive *circumstances* such as growing up in an academic household are relatively more important. In that context, the fact that the impact of the reform with respect to reading skills is less pronounced could be interesting for another reason. On one hand, it might raise the question of whether to improve reading skills, current curricula and teaching methods need to be adjusted. On the other hand, it could also only indicate that the reading practice from additional teaching only balances out the negative impact of increased intensity on the actual learning process - which would be another potential part of the explanation for why IEOp levels for the domain of reading may be less pronounced than in the other domains. However, given the broad definition of learning intensity, this may still be compatible with findings that the G-8 reform itself had small positive effects on mathematics/science test scores in contrast to reading test scores (Camarero Garcia, 2012; Andrietti & Su, 2019; Huebener et al., 2017; Büttner & Thomsen, 2015). Furthermore, Dahmann (2017) shows that cognitive skills measured by IQ proxies did not causally change due to the reform, but gender-specific differences were reinforced. The fact that there appear to be no SES-specific differences in IQs supports my findings: The observed overall increase in IEOp seems to be mainly driven by heterogeneity in parental support opportunities in dealing with the higher learning intensity

and cannot be simply explained by potential differences in ability. Finally, as the reform did not adjust teaching-related quality factors for the first affected cohorts, the findings might be regarded to be merely a lower bound for the effects of increased learning intensity on performance, in particular as the variance of test scores did not change much.

In summary, even though it is beyond the scope of this article to precisely detect all underlying channels and mechanisms explaining how IEOp may be changed and all implications for its translation into both *upward* and *downward* mobility, this paper does reveal one mechanism of how IEOp can be causally changed through an educational reform. That is, by increasing learning intensity.

6 Conclusion

The goal of this paper has been to shed light onto how Inequality of Educational Opportunity (IEOp) may be shaped by the recent trend of accelerating and intensifying the educational process. This is important to understand the role of learning intensity as one policy channel influencing educational opportunities and thus social mobility. Beyond that, the understanding of how institutions affect IEOp is still limited (Ramos & Van de gaer, 2016). To approach an answer to these questions, I focus on the academic track of the German secondary school system, the *Gymnasium* and exploit the shortening of school duration from nine to eight years as a quasi-experiment that exogenously increased learning intensity. This paper is among the first to combine an evaluation of the G-8 reform with PISA data that are comparable across federal states and over time to analyze how increased learning intensity causally affects IEOp in Germany, contributing to the still limited literature on measuring IOp with respect to educational outcomes by adding new evidence.

The first step of the analysis involves measuring IEOp as share in the variance of standardized PISA test scores that can only be attributed to *circumstances* beyond an individual's control. Interestingly, the estimated IEOp measures correspond to the levels of estimates for inequality of opportunity in income, pointing to the link between IEOp and (intergenerational) social immobility. The innovative approach of employing a machine learning algorithm to evaluate which *circumstances* variables are relevant can provide us with a second layer of data-driven evidence for the credibility of my IEOp measure (Appendix A.5.2). As a second step, I conduct a DiD estimation strategy to derive causal estimates, with treatment and control groups chosen according to the implementation of the G-8 reform across federal states. The results reveal that the reform-induced increase in learning intensity did not affect IEOp in the short term. Instead, in the medium term, IEOp significantly increases for affected student cohorts. These findings can be rationalized by differential compensation possibilities for higher learning intensity depending on parental resources in terms of the capacity to pay for additional tuition, which may also explain the increased use of private tutoring as documented by Hille et al. (2016). This interpretation is also supported by the outcomes of a DiD estimation with interaction terms on PISA-scores which allow distinguishing the effects by students' socio-economic background (Tables 4 and 5). Moreover, results point to the existence of subject-dependent curricular flexibilities, with mathematics/science being more inflexible, that is, more responsive to changes in curricular intensity compared to reading.

This paper also contributes to the literature on evaluating this German school reform which is still controversially debated as it shifts attention in the evaluation of the G-8 reform onto distributional concerns. I show that the G-8 reform can be considered to be a *selective* reform that at least maintains test results, but at the same time increases IEOp, and not to be an *inclusive* reform that at least maintains test results while reducing IEOp (Checchi & van de Werfhorst, 2018). To lower IEOp despite higher learning intensity, whole-day schooling and methods reducing the dependence of educational support on *circumstances* might be a solution (Deckers et al., 2019).[34] Alternatively, to maintain equality of opportunity when reducing school duration without adjusting the support schemes at school, the curriculum may need to be reduced accordingly.

Beyond the narrow context of the G-8 reform, there are two broader issues this paper touches on. First, the interaction of IEOp and social mobility is likely to be pivotal for understanding phenomena such as the high persistence in the observed intergenerational transmission of educational achievement. Generally, it would be interesting to evaluate social mobility in regard of *upward* and *downward* mobility. This component seems to be still neglected, in the sense that focus appears to have shifted onto improving *upward* mobility, while ignoring that this cannot be discussed independently from removing rigidities that potentially limit *downward* mobility. Thus, understanding the effects of compressing education on IEOp and its implications for social mobility are highly relevant.[35] Second, the factor of time compression in the context of education appears to have been largely neglected so far and more research on this topic is needed. Politicians consider changes on the margin of educational intensity, but as the G-8 reform shows, this could involve unintended and underestimated welfare costs. A better understanding of the relationship between schooling duration, intensity, and IEOp would also be important in the context of evaluating the welfare benefits and costs of investments into the educational system. As the costs associated with the misallocation of talents due to a lack of social (educational) mobility may be considerable (Philippis & Rossi, 2019; Boneva & Rauh, 2019), it is economically desirable to achieve more equality of educational opportunities. Therefore, this paper shows that the implementation of an appropriate level of educational intensity should not only depend on efficiency considerations, but also consider the effects on equal access to resources.

Taking stock of this discussion, the paper shows that *circumstances* matter at school with an emphasis on the relevance of variation in learning intensity on IEOp. Future research should aim at understanding further potential mechanisms and channels shaping IEOp (Rothstein, 2019). Furthermore, additional work is needed to establish how IEOp translates into social mobility. This in turn may then permit us to assess the welfare effects of IEOp with respect to its impact on future income and wealth inequality. Finally, the outcomes of this research agenda would allow for the evaluation of new policy recommendations aimed at improving equality of opportunity to tackle challenges surrounding high levels of inequality.

[34] For a discussion concerning the role of the government concerning child investment, see also Black and Rothstein (2019).

[35] Thereby, a new theory of how learning (duration and intensity) and IEOp as well as how IEOp and social mobility are linked together could allow quantifying precisely the role of learning intensity for absolute educational mobility, thus social mobility.

References

Aakvik, A., Salvanes, K. G., & Vaage, K. (2010). Measuring Heterogeneity in the Returns to Education Using an Education Reform. *European Economic Review*, *54*(4), 483–500. doi: 10.1016/j.euroecorev.2009.09.001

Aksoy, T., & Link, C. R. (2000). A Panel Analysis of Student Mathematics Achievement in the US in the 1990s: Does Increasing the Amount of Time in Learning Activities Affect Math Achievement? *Economics of Education Review*, *19*(3), 261–277. doi: 10.1016/S0272-7757(99)00045-X

Almås, I., Cappelen, A. W., Lind, J. T., Sørensen, E. Ø., & Tungodden, B. (2011). Measuring unfair (in)equality. *Journal of Public Economics*, *95*(7-8), 488–499. doi: 10.1016/j.jpubeco.2010.11.002

Andreoli, F., Havnes, T., & Lefranc, A. (2018). Robust Inequality of Opportunity Comparisons: Theory and Application to Early Childhood Policy Evaluation. *The Review of Economics and Statistics*, *98*(2), 1–15. doi: 10.1162/rest_a_00747

Andrietti, V., & Su, X. (2019, sep). The Impact of Schooling Intensity on Student Learning: Evidence from a Quasi-Experiment. *Education Finance and Policy*, *14*(4), 679–701. doi: 10.1162/edfp_a_00263

Angrist, J. D., & Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. doi: 10.2307/2937954

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275. doi: 10.1162/003355304772839588

Black, S. E., & Devereux, P. J. (2011). Recent Developments in Intergenerational Mobility. *Handbook of Labor Economics*, *4*(Part B), 1487–1541. doi: 10.1016/S0169-7218(11)02414-2

Boneva, T., & Rauh, C. (2018). Parental Beliefs About Returns to Educational Investments - The Later the Better? *Journal of the European Economic Association*, *16*(6), 1669–1711. doi: 10.1093/jeea/jvy006

Boneva, T., & Rauh, C. (2019). Socio-Economic Gaps in University Enrollment: The Role of Perceived Pecuniary and Non-Pecuniary Returns. *HCEO Working Paper Series*(2017-080).

Boserup, S. H., Kopczuk, W., & Kreiner, C. T. (2018). Born with a Silver Spoon? Danish Evidence on Wealth Inequality in Childhood. *The Economic Journal*, *128*(612), F514–F544. doi: 10.1111/ecoj.12496

Brunori, P., Hufe, P., & Mahler, D. G. (2018). *The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees*. The World Bank. doi: 10.1596/1813-9450-8349

Brunori, P., Peragine, V., & Serlenga, L. (2012). Fairness in Education: The Italian University Before and After the Reform. *Economics of Education Review*, *31*(5), 764–777. doi: 10.1016/j.econedurev.2012.05.007

Bundeszentrale für politische Bildung. (2008). *Datenreport 2008 - Ein Sozialbericht für die Bundesrepublik Deutschland*. Bonn: Statistisches Bundesamt (Destatis).

Büttner, B., & Thomsen, S. L. (2015). Are We Spending Too Many Years in School? Causal Evidence of the Impact of Shortening Secondary School Duration. *German Economic Review*, *16*(1), 65–86. doi: 10.1111/geer.12038

Camarero Garcia, S. (2012). *Does Shortening Secondary School Duration Affect Student Achievement and Educational Equality? - Evidence from a Natural Experiment in Germany: The G-8 Reform* [Bachelor Thesis, University of St. Gallen].

Cantoni, D., Chen, Y., Yang, D. Y., Yuchtman, N., & Zhang, Y. J. (2017). Curriculum and Ideology. *Journal of Political Economy*, *125*(2), 338–392. doi: 10.1086/690951

Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2010). Responsibility for What? Fairness and Individual Responsibility. *European Economic Review*, *54*(3), 429–441. doi: 10.1016/j.euroecorev.2009.08.005

Carneiro, P. (2008). Equality of Opportunity and Educational Achievement in Portugal. *Portuguese Economic Journal*, *7*(1), 17–41. doi: 10.1007/s10258-007-0023-z

Checchi, D., & Peragine, V. (2010). Inequality of Opportunity in Italy. *The Journal of Economic Inequality*, *8*(4), 429–450. doi: 10.1007/s10888-009-9118-3

Checchi, D., & van de Werfhorst, H. G. (2018). Policies, Skills and Earnings: How Educational Inequality Affects Earnings Inequality. *Socio-Economic Review*, *16*(1), 137–160. doi: 10.1093/ser/mwx008

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2020, aug). Income Segregation and Intergenerational Mobility Across Colleges in the United States*. *The Quarterly Journal of Economics*, *135*(3), 1567–1633. doi: 10.1093/qje/qjaa005

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., & Narang, J. (2017). The Fading American Dream: Trends in Absolute Income Mobility since 1940. *Science*, *356*(6336), 398–406. doi: 10.1126/science.aal4617

Dahmann, S. C. (2017). How Does Education Improve Cognitive Skills? Instructional Time Versus Timing of Instruction. *Labour Economics*, *47*, 35–47. doi: 10.1016/j.labeco.2017.04.008

Deckers, T., Falk, A., Kosse, F., Pinger, P., & Schildberg-Hörisch, H. (2019). Socio-Economic Status and Inequalities in Children's IQ and Economic Preferences. *Journal of Political Economy (forthcoming),*, 1–75.

Deming, D. (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111–134. doi: 10.1257/app.1.3.111

Dustmann, C., Puhani, P. A., & Schönberg, U. (2017). The Long-Term Effects of Early Track Choice. *The Economic Journal*, *127*(603), 1348–1380. doi: 10.1111/ecoj.12419

Eble, A., & Hu, F. (2019, jun). Does Primary School Duration Matter? Evaluating the Consequences of a Large Chinese Policy Experiment. *Economics of Education Review*, *70*(March), 61–74. doi: 10.1016/j.econedurev.2019.03.006

Edmark, K., Frölich, M., & Wondratschek, V. (2014). Sweden's School Choice Reform and Equality of Opportunity. *Labour Economics*, *30*, 129–142. doi: 10.1016/j.labeco.2014.04.008

Ferreira, F. H. G., & Gignoux, J. (2011). The Measurement of Inequality of Opportunity: Theory and an Application to Latin America. *Review of Income and Wealth*, *57*(4), 622–657. doi: 10.1111/j.1475-4991.2011.00467.x

Ferreira, F. H. G., & Gignoux, J. (2013). The Measurement of Educational Inequality: Achievement and Opportunity. *The World Bank Economic Review*, *28*(2), 210–246. doi: 10.1093/wber/lht004

Fleurbaey, M., & Peragine, V. (2013). Ex Ante Versus Ex Post Equality of Opportunity. *Economica*, *80*(317), 118–130. doi: 10.1111/j.1468-0335.2012.00941.x

Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of Opportunity for Educational Achievement in Latin America: Evidence from PISA 2006-–2009. *Economics of Education Review*, *31*(5), 694–708. doi: 10.1016/j.econedurev.2012.05.002

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A Standard International Socio-Economic Index of Occupational Status. *Social Science Research*, *21*, 1–56. doi: 0049-089X/92

Grenet, J. (2013). Is Extending Compulsory Schooling Alone Enough to Raise Earnings? Evidence from French and British Compulsory Schooling Laws. *The Scandinavian Journal of Economics*, *115*(1), 176–210. doi: 10.1111/j.1467-9442.2012.01739.x

Hille, A., Spieß, C. K., & Staneva, M. (2016). More and More Students, Especially Those from Middle-Income Households, are Using Private Tutoring. *DIW Economic Bulletin*(6), 63–71. doi: 10.5684/soep.v30.

Hübner, N., Wagner, W., Hochweber, J., Neumann, M., & Nagengast, B. (2020, jan). Comparing Apples and Oranges: Curricular Intensification Reforms Can Change the Meaning of Students' Grades! *Journal of Educational Psychology*, *112*(1), 204–220. doi: 10.1037/edu0000351

Huebener, M., Kuger, S., & Marcus, J. (2017). Increased Instruction Hours and the Widening Gap in Student Performance. *Labour Economics*, *47*(1561), 15–34. doi: 10.1016/j.labeco.2017.04.007

Huebener, M., & Marcus, J. (2017). Compressing Instruction Time into Fewer Years of Schooling and the Impact on Student Performance. *Economics of Education Review*, *58*, 1–14. doi: 10.1016/j.econedurev.2017.03.003

Hufe, P., Peichl, A., Roemer, J., & Ungerer, M. (2017). Inequality of Income Acquisition: The Role of Childhood Circumstances. *Social Choice and Welfare*, *49*(3-4), 499–544. doi: 10.1007/s00355-017-1044-x

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86. doi: 10.1257/jel.47.1.5

Klemm, K., & Hollenbach-Biele, N. (2006). *Nachhilfeunterricht in Deutschland: Ausmaß-Wirkung-Kosten*. Bertelsmann Stiftung.

KMK. (2016). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II.* Retrieved from: `https://www.bildung.sachsen.de/Sek2.pdf`. (Last access: September 16, 2020)

Krashinsky, H. (2014). How Would One Extra Year of High School Affect Academic Performance in University? Evidence from an Educational Policy Change. *Canadian Journal of Economics*, *47*(1), 70–97. doi: 10.1111/caje.12066

Lavy, V. (2015). Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries. *The Economic Journal*, *125*(588), F397–F424. doi: 10.1111/ecoj.12233

Lefranc, A., & Trannoy, A. (2017). Equality of Opportunity, Moral Hazard and the Timing of Luck. *Social Choice and Welfare*, *49*(3-4), 469–497. doi: 10.1007/s00355-017-1054-8

Machin, S. (2014). Developments in Economics of Education Research. *Labour Economics*, *30*, 13–19. doi: 10.1016/j.labeco.2014.06.003

Marcotte, D. E. (2007). Schooling and Test Scores: A Mother-Natural Experiment. *Economics of Education Review*, *26*(5), 629–640. doi: 10.1016/j.econedurev.2006.08.001

Meyer, B. D. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics*, *13*(2), 151–161. doi: 10.2307/1392369

Meyer, T., & Thomsen, S. L. (2016). How Important is Secondary School Duration for Postsecondary Education Decisions? Evidence from a Natural Experiment. *Journal of Human Capital*, *10*(1), 67–108. doi: 10.1086/684017

Meyer, T., Thomsen, S. L., & Schneider, H. (2018, mar). New Evidence on the Effects of the Shortened School Duration in Germany: An Evaluation of Post-School Education Decisions. *German Economic Review*, *9507*(9507). doi: 10.1111/geer.12162

Niederle, M., & Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives*, *24*(2), 129–144. doi: 10.1257/jep.24.2.129

Niehues, J., & Peichl, A. (2014). Upper Bounds of Inequality of Opportunity: Theory and Evidence for Germany and the US. *Social Choice and Welfare*, *43*(1), 73–99. doi: 10.1007/s00355-013-0770-y

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing. doi: 10.1787/9789264167872-en

OECD. (2014). *Education at a Glance 2014* (Vol. 2012). OECD Publishing. doi: 10.1787/eag-2014-en

Philippis, M. D., & Rossi, F. (2019). Parents, Schools and Human Capital Differences across Countries. *Journal of the European Economic Association (forthcoming)*, 1–43.

Piketty, T., & Zucman, G. (2014). Capital is Back: Wealth-Income Ratios in Rich Countries 1700–2010. *The Quarterly Journal of Economics*, *129*(3), 1255–1310. doi: 10.1093/qje/qju018

Pischke, J.-S. (2007). The Impact of Length of the School Year on Student Performance and Earnings: Evidence From the German Short School Years. *The Economic Journal*, *117*(523), 1216–1242. doi: 10.1111/j.1468-0297.2007.02080.x

Ramos, X., & Van de gaer, D. (2016). Approaches to Inequality of Opportunity: Principles, Measures and Evidence. *Journal of Economic Surveys*, *30*(5), 855–883. doi: 10.1111/joes.12121

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Roemer, J. (1998). *Equality of Opportunity*. Cambridge: Harvard University Press.

Roemer, J., & Trannoy, A. (2015). Equality of Opportunity. In *Handbook of Income Distribution* (Vol. 2, pp. 217–300). Amsterdam, The Netherlands: Elsevier. doi: 10.1016/B978-0-444-59428-0.00005-9

Rothstein, J. (2019). Inequality of Educational Opportunity? Schools as Mediators of the Intergenerational Transmission of Income. *Journal of Labor Economics*, *37*(S1), S85–S123. doi: 10.1086/700888

Sen, A. (1980). Equality of What? *The Tanner Lecture on Human Values*, *I*, 197–220.

Thiel, H., Thomsen, S. L., & Büttner, B. (2014). Variation of Learning Intensity in Late Adolescence and the Effect on Personality Traits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(4), 861–892. doi: 10.1111/rssa.12079

Wössmann, L. (2010). Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries. *Jahrbücher fur Nationalökonomie und Statistik*, *230*(2).

List of Abbreviations

DiD	Difference-in-Difference estimation approach.
EEOp	Equality of Educational Opportunity.
EOp	Equality of Opportunity.
G-8 model	Gymnasium-8 model.
G-8 reform	Gymnasium-8 reform.
G-9 model	Gymnasium-9 model.
IEOp	Inequality of Educational Opportunity.
IOp	Inequality of Opportunity.
IQB	Institut zur Qualitätsentwicklung im Bildungswesen.
ISCED	International Standard Classification of Education.
ISCO	International Standard Classification of Occupation.
ISEI	International Socio-Economic Index of Occupational Status.
OECD	Organization of Economic Co-operation and Development.
PISA	Program for International Student Assessment.
SC	Standing Conference of the Ministers of Education and Cultural Affairs.
SES	Socio-Economic Status.

Glossary

Gymnasium is the academic track of secondary school education in Germany covering both lower and upper secondary level (grades 5–13 or 5–12) and providing an in-depth general education aimed at the general higher education entrance qualification (*Allgemeine Hochschulreife*).

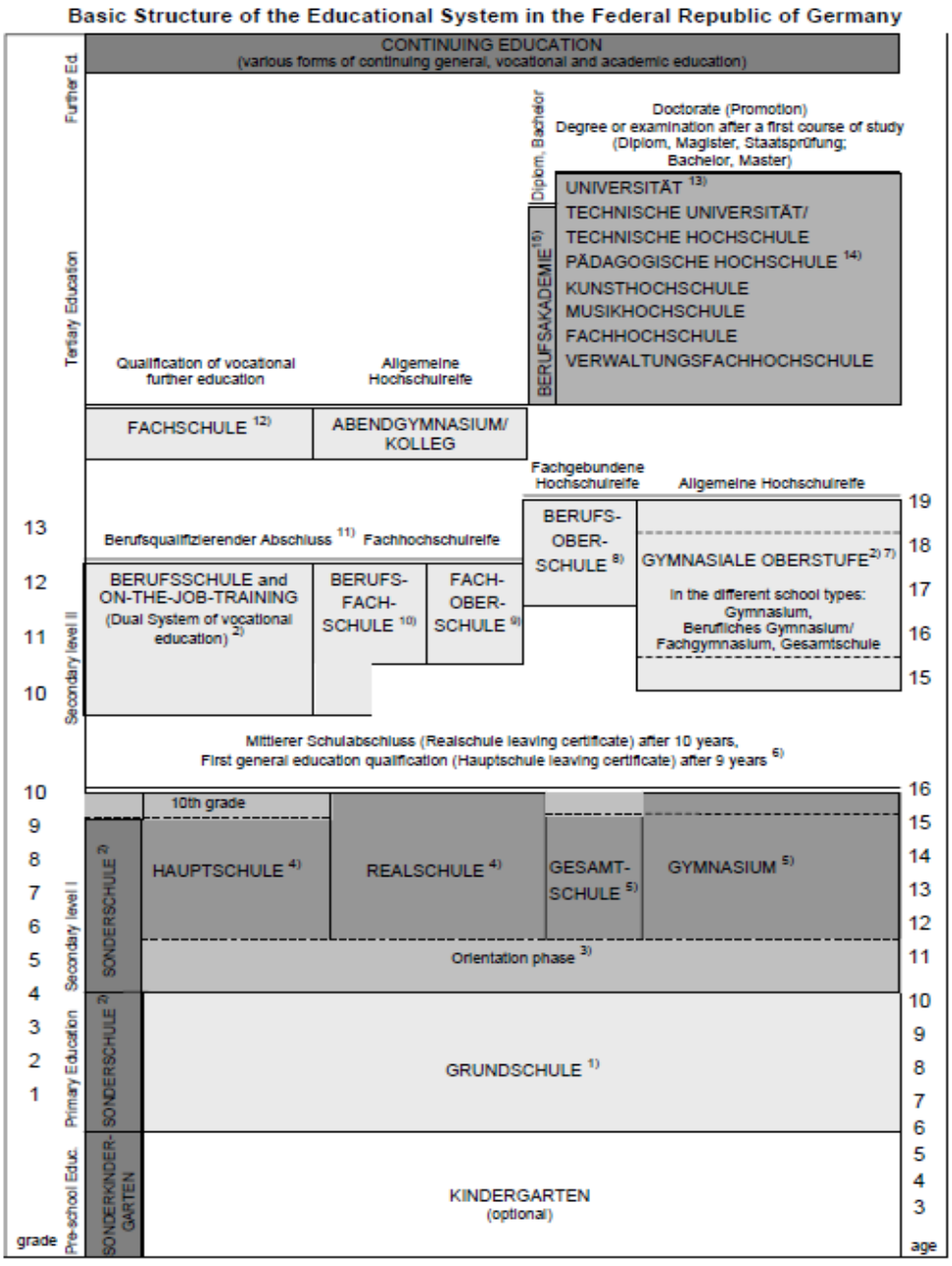
Learning Intensity is the ratio of curricular content covered in a given period of time. In particular, the G-8 reform led to an increase in schooling intensity in such a way that by the end of grade 9 in the post-reform period, students have received about the same amount of instruction, and covered the same curriculum as students that had completed two-thirds of grade 10 in the pre-reform period. Learning Intensity, thus, reflects the amount of content (curriculum) to be studied within a fixed amount of instruction time, whereas school duration (in years) refers to the total amount of instruction required to be eligible for graduation.

Plausible Value Following OECD (2009b) in chapter 6: Instead of directly estimating a student's ability θ, a probability distribution is estimated. Thus, instead of obtaining a point estimate, a range of possible values with an associated probability for each is estimated. Plausible Values are random draws from this (estimated) distribution.

A Appendix

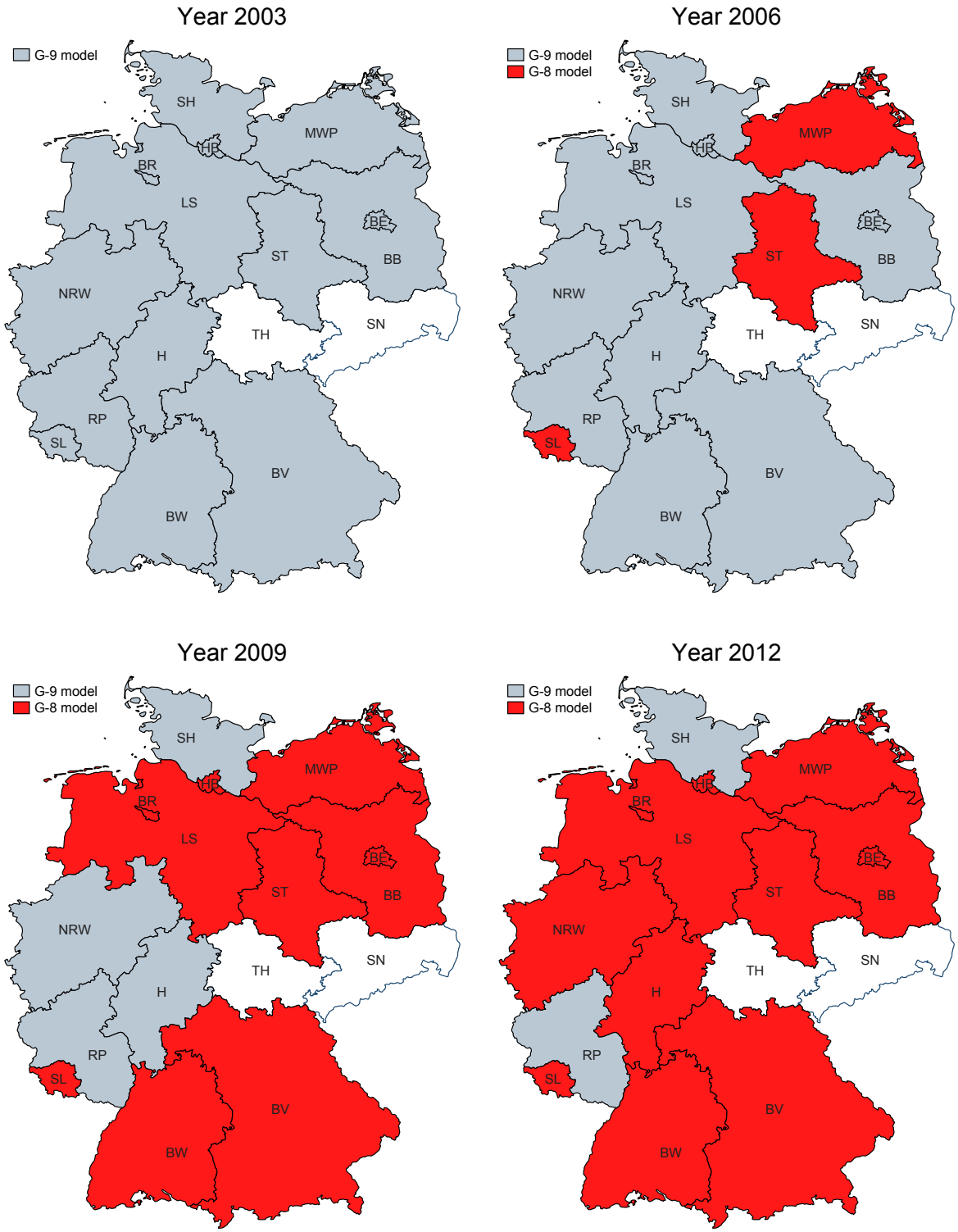
A.1 Supplementary Figures

Figure A.1: Structure of the German Educational System (as explained in Section 2)



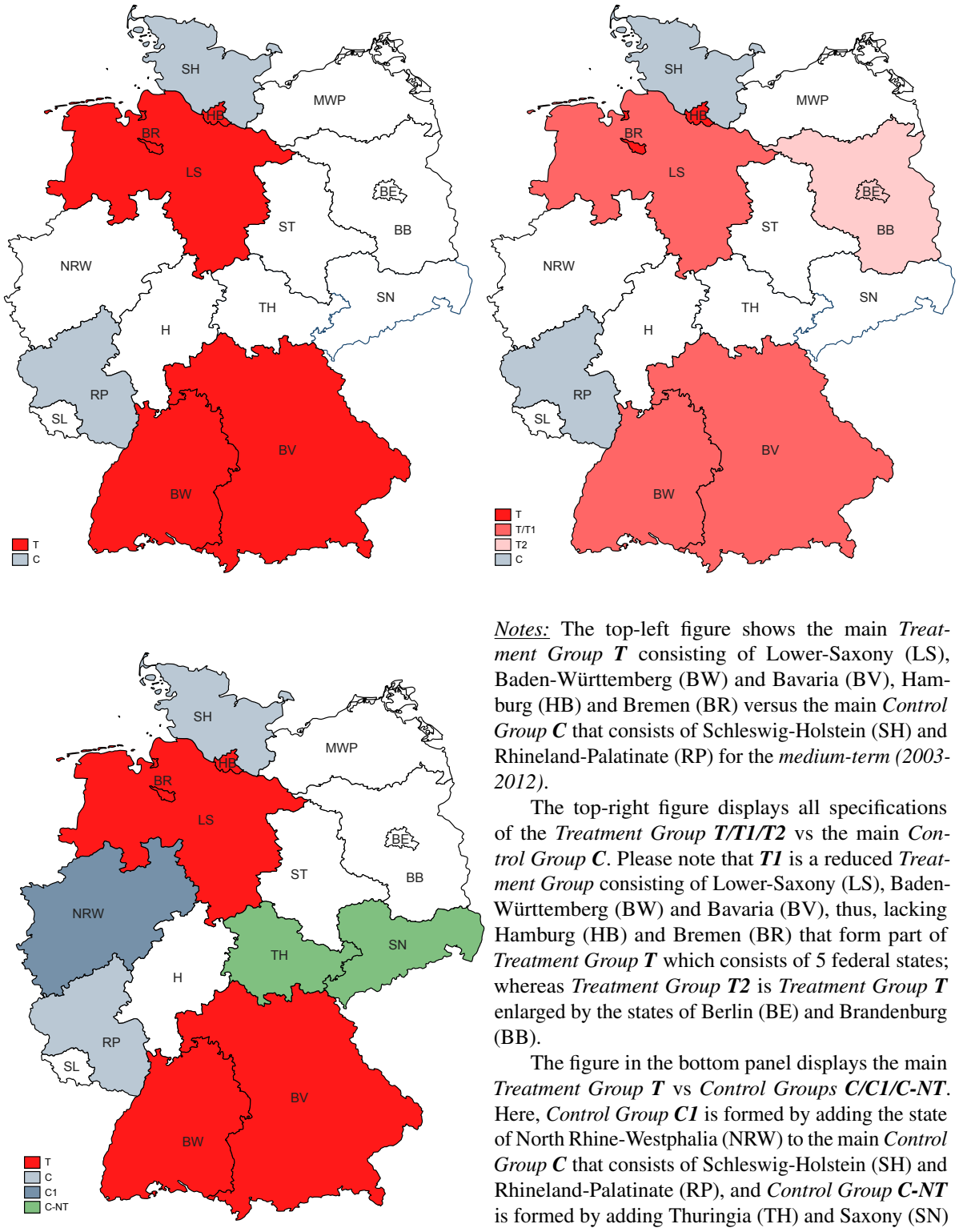
Notes: This figure illustrates the basic structure of the German education system. For the data source and more details, see Standing Conference of Education Ministers (2009): Basic Structure of the Education System in the Federal Republic of Germany.

Figure A.2: Overview of G-8 Reform across Federal States for Students Tested in PISA (2003-2012)



Notes: This figure illustrates whether 9^{th} graders attending a *Gymnasium* tested in a PISA test year (2003, 2006, 2009, 2012) were still taught in a G-9 model (grey/blue) or were already attending a reformed G-8 model (dark grey/red).

Figure A.3: Overview of the Treatment/Control Group Setting

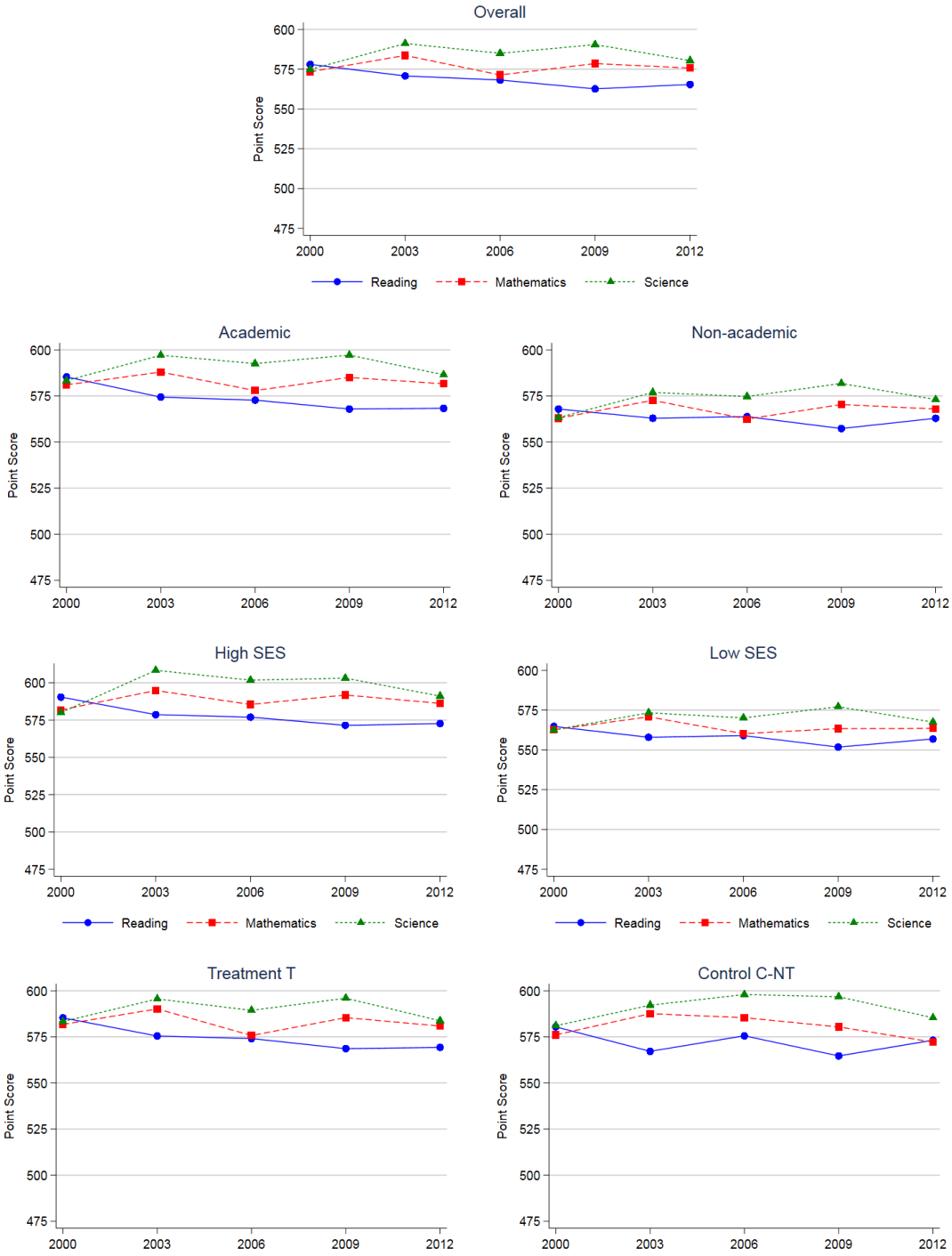


Notes: The top-left figure shows the main *Treatment Group* ***T*** consisting of Lower-Saxony (LS), Baden-Württemberg (BW) and Bavaria (BV), Hamburg (HB) and Bremen (BR) versus the main *Control Group* ***C*** that consists of Schleswig-Holstein (SH) and Rhineland-Palatinate (RP) for the *medium-term (2003-2012)*.

The top-right figure displays all specifications of the *Treatment Group* ***T/T1/T2*** vs the main *Control Group* ***C***. Please note that ***T1*** is a reduced *Treatment Group* consisting of Lower-Saxony (LS), Baden-Württemberg (BW) and Bavaria (BV), thus, lacking Hamburg (HB) and Bremen (BR) that form part of *Treatment Group* ***T*** which consists of 5 federal states; whereas *Treatment Group* ***T2*** is *Treatment Group* ***T*** enlarged by the states of Berlin (BE) and Brandenburg (BB).

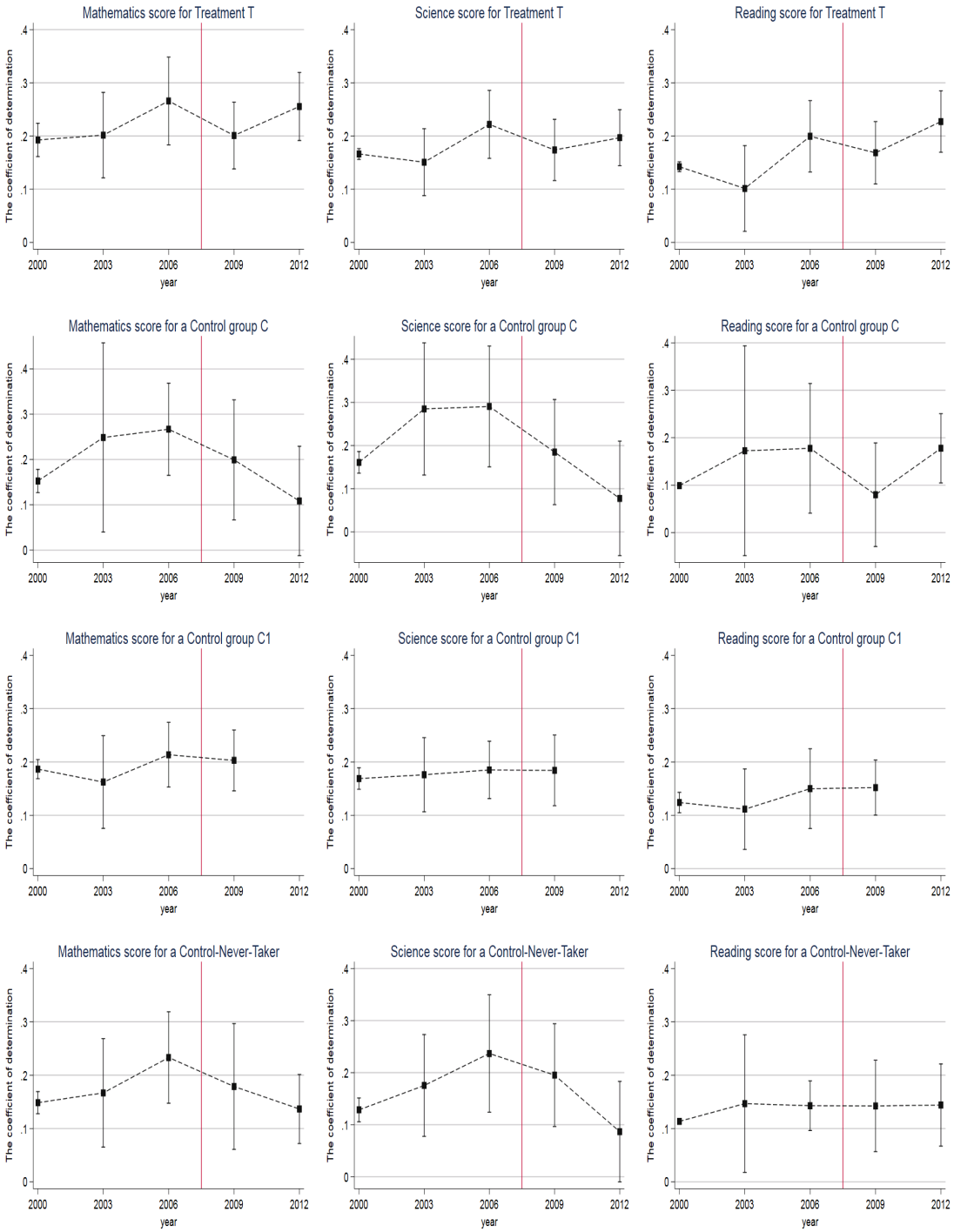
The figure in the bottom panel displays the main *Treatment Group* ***T*** vs *Control Groups* ***C/C1/C-NT***. Here, *Control Group* ***C1*** is formed by adding the state of North Rhine-Westphalia (NRW) to the main *Control Group* ***C*** that consists of Schleswig-Holstein (SH) and Rhineland-Palatinate (RP), and *Control Group* ***C-NT*** is formed by adding Thuringia (TH) and Saxony (SN) to the main *Control Group* ***C***.

Figure A.4: Descriptive Analysis: Mean Test Score by Main Groups (as explained in Section 3.1)



Notes: This figure shows the mean scores for all three PISA test domains. Focusing on students in *Gymnasium*, the scores are above the average of 500 points. On overall, students perform best in science, then mathematics and relatively worst in reading. The grouping into academic vs. non-academic background is based on the binary variable indicating whether at least one parent has a college degree. To distinguish between high and low SES, students have been assigned to quartiles of their highest parental job's ISEI. Being in the first quartile translates into low SES, whereas the fourth quartile into high SES. Students from academic households achieve slightly higher scores than those from non-academic ones. A similar picture derives when distinguishing between high and low SES. Finally, the main *Treatment-Group* **T** and *Control-Group* **C-NT** (Never-Takers), as defined in Section 4.2, have similar test score levels.

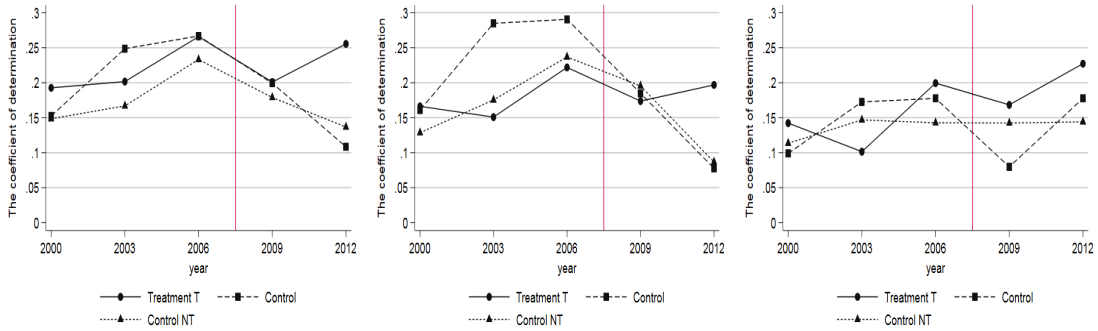
Figure A.5: IEOp Measure for Treatment/Control Groups Over Time (2000-2012)



Notes: This figure shows the IEOp measure ($R^2_{adjusted}$) with 90% confidence intervals over the whole time period. Standard errors to construct confidence intervals are calculated according to Appendix A.5.1. Standard errors for the year 2003 are particularly large due to idiosyncratic weights for that year. PISA 2000 is included for robustness reasons, as explained in Section 3.

Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009, and 2012.

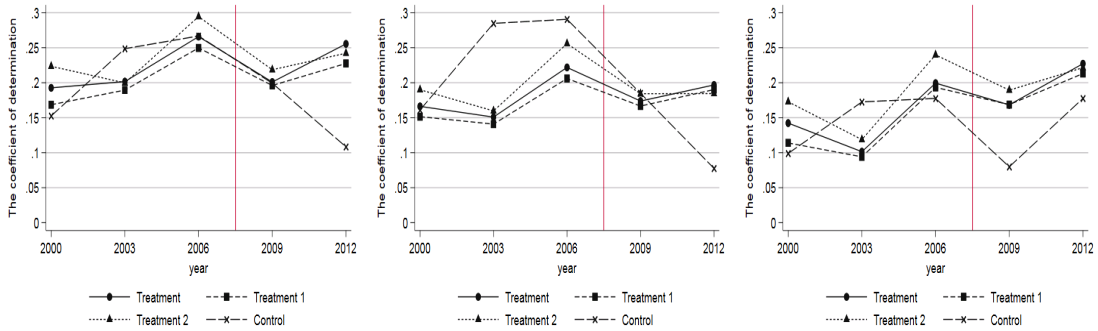
Figure A.6: DiD Graphs of IEOp Measure for Main Treatment/Control Groups



(a) IEOp Measure Based on Maths **(b)** IEOp Measure Based on Science **(c)** IEOp Measure Based on Reading

Notes: This figure shows the DiD graphs for all three test domains confirming the parallel trend assumption. *Treatment* is the main treatment group **T**; *Control* is the main control group **C**; *Control-NT* is the never-changing control group. PISA 2000 is included for robustness reasons, as explained in Section 3.
Compare: These graphs correspond to the main strategy and the main results as explained in Section 5. This figure eintails Figure 2. The treatment and control groups are explained in Figure A.3 as well as in Section 4.2.
Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009, and 2012.

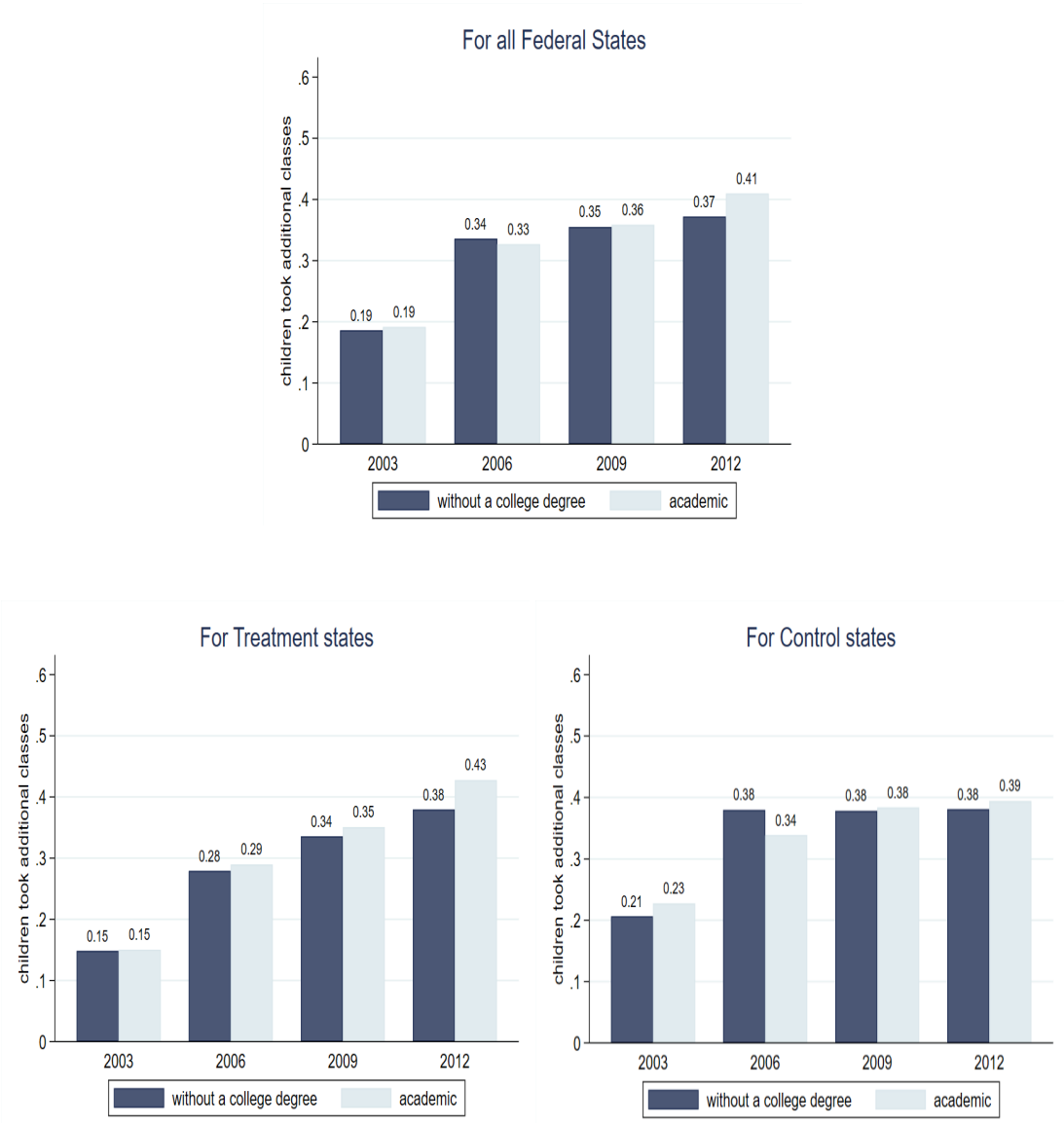
Figure A.7: Robustness - DiD Graphs of IEOp Measure for Enlarged Treatment/Control Groups



(a) IEOp Measure Based on Maths **(b)** IEOp measure based on science **(c)** IEOp Measure Based on Reading

Notes: This figure shows the DiD graphs for all three test domains confirming the parallel trend assumption to hold and being invariable to alternative compositions of the treatment group. *Treatment* is the main treatment group **T** consisting of five federal states, *Treatment 1* is the **T1** Group consisting of three federal states, and *Treatment 2* is the extension Group consisting of seven federal states compared to *Control* which is the main control group **C**.
Compare: These graphs correspond to the main strategy and the main results as explained in Section 5. The treatment and control groups are explained in Figure A.3 as well as in Section 4.2. PISA 2000 is included for robustness reasons, as explained in Section 3.
Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009, and 2012.

Figure A.8: Potential Mechanism: Extra Tuition



Notes: This figure shows the percentage of tested students indicating that they took extra classes beyond official school lessons, mostly referring to paid extra tuition. The dark and light blue bars correspond to students growing up in non-academic households and academic households (at least one parent has a university diploma (ISCED-level 5 or 6)), respectively. The first panel shows an upward trend in the demand for extra tuition between 2003 and 2012 across all federal states. In treatment states, the increase in extra tuition has been stronger for students from academic households in the post-reform period from 2009 to 2012. This indicates a differential adjustment with respect to extra-tuition depending on a student's parental educational background. In contrast within control states, no differential response of students - depending on their parental educational background - can be found in post-reform years (2009 and 2012).
Source: Author's own calculations based on PISA 2003, 2006, 2009, and 2012.

A.2 Supplementary Tables

Table A.1: Overview of "G-8 reform" Across Federal States by Year of *Double Cohort*

Federal state	Type of Federal State			Reform Timeline		Gymnasium		Reversal[b]
	West/East	City/Terr.	Population[a]	Begins	Ends	Type	Grade	yes/no
Saxony (SN)	East	territorial	4,0 mio	-	-	5-12	-	no reform[c]
Thuringia (TH)	East	territorial	2,2 mio	-	-	5-12	-	no reform[c]
Saxony-Anhalt (ST)	East	territorial	2,3 mio	2003/2004	2006/2007	5-12	9^{th}	no
Mecklenburg-West Pomerania (MWP)	East	territorial	1,6 mio	2004/2005	2007/2008	7-12	9^{th}	no
Saarland (SL)	West	territorial	1,0 mio	2001/2002	2008/2009	5-12	5^{th}	no[d]
Hamburg (HB)	West	city state	1,7 mio	2002/2003	2009/2010	5-12	5^{th}	no[e]
Bavaria (BV)[f]	West	territorial	12,5 mio	2004/2005	2010/2011	5-12	5^{th},6^{th}	yes[g]
Lower Saxony (LS)[f]	West	territorial	7,8 mio	2004/2005	2010/2011	5-12	5^{th},6^{th}	yes[h]
Baden-Württemberg (BW)	West	territorial	10,5 mio	2004/2005	2011/2012	5-12	5^{th}	no[i]
Bremen (BR)	West	city state	0,7 mio	2004/2005	2011/2012	5-12	5^{th}	no[j]
Berlin (BE)	West	city state	3,4 mio	2006/2007	2011/2012	7-12	7^{th}	no[k]
Brandenburg (BB)	East	territorial	2,5 mio	2006/2007	2011/2012	7-12	7^{th}	no[k]
North Rhine-Westphalia (NRW)	West	territorial	17,6 mio	2005/2006	2012/2013	5-12	5^{th}	no[l]
Hesse (H)	West	territorial	6,0 mio	varies[m]	varies[m]	5-12	5^{th}	yes[n]
Rhineland-Palatinate (RP)	West	territorial	4,0 mio	2008/2009	2015/2016	5-13	5^{th}	yes[o]
Schleswig-Holstein (SH)	West	territorial	2,8 mio	2008/2009	2015/2016	5-13	5^{th}	yes[p]

[a] Numbers taken from the most recent census in 2011 are valid for the considered time period from 2003 to 2012 (German Federal Statistical Office, 2014, Area and population).
[b] See Secretariat of Standing Conference of Ministers of Education: `https://www.kmk.org/themen/allgemeinbildende-schulen/bildungswege-und-abschluesse/sekundarstufe-ii-gymnasiale-oberstufe-und-abitur.html`
[c] Since 1949, these states have implemented a G-8 model in the GDR and never had a G-9 model.
[d] *Gymnasium* remains in G-8 model, but in comprehensive schools, a G-13 model is possible.
[e] *Gymnasium* remains in G-8 model, whereas the Stadtschule as a comprehensive school offers a G-13 model.
[f] In Bavaria (BV) and Lower Saxony (LS), the 6^{th} and 5^{th} grade were allocated to the G-8 model in the same year, suggesting that educational intensity was stronger for then 6^{th} graders who had to learn the curriculum over 7 instead of 8 years than (for then) 5^{th} graders. Yet, tested 9^{th} graders in 2009 were affected by the reform right from grade 5.
[g] General revision to G-9 model starting with school year 2019/2020 as announced in April 2017
[h] General revision to G-9 model starting with school year 2015/16, but with a voluntary option for the G-8 model
[i] But: since 2012/2013 a state-wide pilot project allows 44 model schools to offer a G-9 model.
[j] But: the so-called Oberschule as comprehensive school offers a G-13 model.
[k] But: integrated comprehensive schools are allowed to offer G-9 (G-13) model.
[l] But: in 2011/2012 there was a pilot project with 13/630 Gymnasien offering a G-9 model.
[m] Successive introduction of the reform in # % of all normal *Gymnasium* (5-12) 2004/2005: 10%; 2005/2006: 60%; 2006/2007: 30% with double cohorts graduating respectively in 2011/2012, 2012/2013 and 2013/2014.
[n] Since 2013/2014: students allowed to choose between G-12 or G-13 model from 5^{th} grade onward.
[o] Always maintained schools with G-9 model: but since 2008/2009 a G-8 model is offered at 19 Gymnasien.
[p] Since 2011/12 schools are allowed by state law to offer a G-9 model (11 of 99 schools), G-8 model or both (4 of 99).

Table A.2: Available Grade-sample based PISA-I Datasets (as explained in Section 3.1)

	Before Reform			**After** Reform	
Dataset	PISA-2000 [a]	**PISA-2003-I**	**PISA-2006-I**	**PISA-2009-I**	**PISA-2012-I**
# of variables	914	1,292	1,095	1,231	1,215
# of students[b]	34,754	8,559	9,577	9,460	9,998
test scores[c]	reading	mathematics	science	reading	mathematics
School-dataset:					
# of variables	470	572	565	534	502
# of schools	1,342	216	226	226	230
Teacher-dataset:[d]					
# of variables	-	653	-	639	257
# of teachers	-	1939	-	2,201	2,084

[a] For the year 2000, there was no specific *grade-based PISA-I-sample* available from the IQB. However, *PISA-2000* (being the *PISA-2000-E* data) is *ninth grade-based* (Baumert et al., 2002). It has a lower number of variables but more observations than the other datasets.

[b] The number of observations for students as included in the PISA datasets (2000, 2003, 2006, 2009, 2012): the data is provided by the IQB and consists of the *grade-based sample* (see also Appendix A.4.2). Note that here the *student*-dataset includes both the original students' questionnaire answers and their parents' responses.

[c] These test score domains have been in focus for the respective PISA test cycle.

[d] For 2000 and 2006, the *teacher*-dataset was not part of the Germany-specific PISA data, as provided by the IQB.

Table A.3: Descriptive Statistics: Outcome Variables and Sample Size (as explained in Section 3.1)

	Before Reform			**After Reform**	
Test Scores of Students in *Gymnasium*	**PISA-2000**	**PISA-2003-I**	**PISA-2006-I**	**PISA-2009-I**	**PISA-2012-I**
Reading Mean	577.92	570.77	568.20	562.65	565.42
Reading SD	55.86	51.98	56.97	55.25	52.81
Reading Median	578.83	572.14	571.50	566.23	567.06
Mathematics Mean	573.65	583.66	571.39	578.53	575.73
Mathematics SD	62.18	57.85	58.48	56.59	58.52
Mathematics Median	572.68	584.70	571.19	580.47	576.19
Science Mean	575.14	591.15	585.01	590.48	580.44
Science SD	67.43	60.20	61.47	58.88	58.61
Science Median	576.35	594.80	587.12	594.68	581.07
# of federal states	16	16	16	16	16
# of schools	409	62	67	68	78
# of students	10,276	3,017	3,356	3,473	3,910

Notes: This table reports summary statistics for the sample of ninth graders attending a *Gymnasium* and is weighted by the sample weights provided in the PISA dataset from the IQB. Note that the average across plausible values can be taken as a metric of individual-level performance (further information on test scores and the weighting procedure is provided in Appendix A.4 and OECD (2012). Mean, standard deviations, and median of the test scores across all federal states and for all academic track schools that are in the German PISA dataset are provided for each test cycle (2000, 2003, 2006, 2009, 2012) as shown in Table A.2.

Table A.4: Pre-Reform Treatment/Control Group Comparison of Control Variable Sets

	T	C	Δ (T-C)	T1	Δ (T1-C)	T2	Δ (T2-C)
Individual Characteristics							
Female	0.533	0.501	0.031	0.537	0.036	0.535	0.033
Age in years	15.495	15.475	0.020	15.491	0.016	15.478	0.003
Language at home not German	0.056	0.043	0.013	0.054	0.010	0.057	0.013
Migration Background	0.190	0.145	0.045**	0.184	0.039*	0.186	0.041*
Parental Characteristics							
Parental Education:							
(highest ISCED level)							
# ISCED-level (5-6):	0.664	0.644	0.019	0.666	0.021	0.670	0.025
# ISCED-level (3-4):	0.290	0.329	-0.040	0.290	-0.039	0.285	-0.045*
# ISCED-level (1-2):	0.046	0.026	0.02*	0.044	0.018	0.045	0.019 *
Socio-Economic Status							
Number of books in household:							
# + 500:	0.233	0.235	-0.003	0.228	-0.008	0.223	-0.012
# 101-500:	0.509	0.520	-0.011	0.513	-0.007	0.504	-0.016
# 11-100:	0.204	0.189	0.015	0.206	0.017	0.215	0.026
# max. 10:	0.054	0.055	-0.001	0.052	-0.003	0.058	0.002
Highest ISEI of parental job	59.427	57.072	2.355***	59.322	2.25**	59.109	2.037 **
Family Characteristics							
Single Parent *(Base cat.: No)*	0.140	0.141	-0.001	0.137	-0.004	0.168	0.027
Father employment status							
# full-time (FT):	0.875	0.866	0.008	0.875	0.008	0.864	-0.002
# part-time (PT):	0.067	0.065	0.002	0.066	0.001	0.067	0.002
# unemployed (UE):	0.026	0.033	-0.007	0.025	-0.008	0.036	0.003
# out-of-labor force (OLF) :	0.032	0.036	-0.003	0.034	-0.002	0.033	-0.003
Mother employment status							
# full-time (FT):	0.220	0.218	0.002	0.220	0.002	0.303	0.084***
# part-time (PT):	0.521	0.513	0.007	0.522	0.008	0.457	-0.057**
# unemployed (UE):	0.061	0.077	-0.016	0.061	-0.015	0.068	-0.008
# out-of-labor force (OLF):	0.198	0.192	0.006	0.197	0.005	0.172	-0.019
Number of students	2,365	347	-	2,175	-	2,999	-

Notes: Robust standard errors are shown in parentheses and significance levels are indicated by: *** 1%, ** 5%, * 10%. This table shows a *two-sample t-test* for comparing the main control variables in the pre-reform period of the main specification between *Treatment Groups* **T/T1/T2** and *Control Group* **C** (see Section 3.3 and Section 4.1). This is for *PISA-I* the respective pooled average of control variables for 2003 and 2006. This table illustrates the checks as described in Section 4.4.
Source: Author's calculations based on PISA 2003 and 2006.

Table A.5: Robust Model: Pre-Reform Treatment/Control Group Comparison of Control Variables

	T	C1	Δ (T-C1)	T1	Δ (T1-C1)	T2	Δ (T2-C1)
Individual characteristics							
Female	0.533	0.549	-0.016	0.537	-0.011	0.535	-0.014
Age in years	15.495	15.468	0.028*	15.491	0.024	15.478	0.011
Language at home not German	0.056	0.056	0.000	0.054	-0.002	0.057	0.001
Migration Background	0.190	0.178	0.012	0.184	0.006	0.186	0.008
Parental characteristics							
Parental Education (highest ISCED level):							
# ISCED-level (5-6):	0.664	0.666	-0.003	0.666	-0.001	0.670	0.003
# ISCED-level (3-4):	0.290	0.296	-0.007	0.290	-0.006	0.285	-0.012
# ISCED-level (1-2):	0.046	0.037	0.009	0.044	0.007	0.045	0.008
Socio-Economic Status							
Number of books in household:							
# + 500:	0.233	0.253	-0.021	0.228	-0.026*	0.223	-0.030**
# 101-500:	0.509	0.496	0.013	0.513	0.018	0.504	0.008
# 11-100:	0.204	0.197	0.007	0.206	0.009	0.215	0.018
# max. 10:	0.054	0.054	0.000	0.052	-0.001	0.058	0.004
Highest ISEI of parental job	59.427	58.818	0.609	59.322	0.503	59.109	0.291
Family Characteristics							
Single Parent *(Base cat.: No)*	0.140	0.150	-0.010	0.137	-0.013	0.168	0.018
Father employment status							
# full-time (FT):	0.875	0.878	-0.004	0.875	-0.004	0.864	-0.014
# part-time (PT):	0.067	0.061	0.007	0.066	0.006	0.067	0.007
# unemployed (UE):	0.026	0.027	-0.001	0.025	-0.002	0.036	0.009*
# out-of-labor force (OLF) :	0.032	0.034	-0.002	0.034	0.000	0.033	-0.002
Mother employment status							
# full-time (FT):	0.220	0.239	-0.018	0.220	-0.019	0.303	0.064***
# part-time (PT):	0.521	0.489	0.032**	0.522	0.033**	0.457	-0.032**
# unemployed (UE):	0.061	0.065	-0.004	0.061	-0.003	0.068	0.004
# out-of-labor force (OLF):	0.198	0.208	-0.010	0.197	-0.011	0.172	-0.035***
Number of students	2,365	1854	-	2,175	-	2,999	-

Notes: Robust standard errors are shown in parentheses and significance levels are indicated by: *** 1%, ** 5%, * 10% + 15%. This table shows a *two-sample t-test* for comparing the main control variables in the pre-reform period between *Treatment Groups* **T/T1/T2** and *Control Group* ***C1*** (see Section 3.3 and Section 4.1). This is for *PISA-I* the respective pooled average of control variables for 2003 and 2006. This table illustrates the checks as described in Section 4.4.
Source: Author's calculations based on PISA 2003 and 2006.

Table A.6: Main Results for *Model Base*: 1^{st} step to Derive IEOp Measure for *Reading*

DEPENDENT VARIABLE:		**Control Group (C)**		**Treatment Group (T)**	
	Reading Test Scores in PISA (STDPVREAD3)	Before (2003-2006)	After (2009-2012)	Before (2003-2006)	After (2009-2012)
CONTROL: **Individual Characteristics (IC)**					
i)	Female	0.066	0.393***	0.288***	0.412***
		(0.105)	(0.086)	(0.040)	(0.038)
	Age in years	-0.059	-0.248**	-0.167**	-0.159***
		(0.183)	(0.119)	(0.065)	(0.040)
ii)	Migration Background	-0.234	-0.167*	-0.074	-0.105*
		(0.241)	(0.090)	(0.073)	(0.055)
	NO German spoken at home	-0.494	-0.153	-0.303***	-0.168**
		(0.530)	(0.201)	(0.113)	(0.071)
CONTROL: **Parental Characteristics (PC)**					
iii)	Parental Education: [Base: *ISCED-level (3-4)*]				
	# at most lower sec. educ. (ISCED-level (1-2))	-0.486**	0.107	-0.303***	-0.005
		(0.225)	(0.199)	(0.100)	(0.055)
	# tertiary educ. (ISCED-level (5-6))	-0.159	0.134	-0.009	-0.048
		(0.156)	(0.109)	(0.057)	(0.048)
CONTROL: **Socio-Economic Status (SES)**					
iv)	No. of books in household [Base: *101-500*]				
	# max 10 books	0.169	-0.441	-0.522**	-0.441***
		(0.395)	(0.272)	(0.225)	(0.125)
	# 11-100 books	-0.126	-0.079	-0.303***	-0.138***
		(0.214)	(0.120)	(0.053)	(0.043)
	# more than 500 books	0.204*	0.077	0.079	0.087*
		(0.117)	(0.069)	(0.060)	(0.051)
v)	Highest ISEI-level of Parental Jobs	0.007	0.002	0.001	0.003***
		(0.005)	(0.003)	(0.002)	(0.001)
CONTROL: **Family Characteristics (FC)**					
vi)	Family Structure [Base: *No*]				
	Single Parent Household	0.079	0.268**	0.066	0.127**
		(0.261)	(0.122)	(0.066)	(0.057)
vii)	Father: Employment [Base: *Full-time (FT)*]				
	# part-time (PT)	-0.300	-0.233	-0.089	-0.096
		(0.208)	(0.280)	(0.095)	(0.074)
	# unemployed (UE)	0.382	0.320	-0.023	0.106
		(0.441)	(0.373)	(0.187)	(0.154)
	# out-of-labor force (OLF)	0.014	-0.079	0.082	0.125
		(0.271)	(0.182)	(0.150)	(0.097)
	Mother: Employment [Base: *Full-time (FT)*]				
	# part-time (PT)	-0.013	-0.058	0.005	0.004
		(0.120)	(0.107)	(0.058)	(0.045)
	# unemployed (UE)	0.267	0.257	-0.062	0.136
		(0.240)	(0.490)	(0.127)	(0.097)
	# out-of-labor force (OLF)	-0.165	0.101	-0.053	-0.023
		(0.150)	(0.120)	(0.079)	(0.058)
	Constant	0.754	3.348*	2.611**	2.141***
		(2.969)	(1.874)	(1.054)	(0.638)
	School FE	yes	yes	yes	yes
	Observations	346	608	2356	3329
	R^2	0.242***	0.162***	0.180***	0.213***
		(0.057)	(0.034)	(0.031)	(0.020)
	$R^2 - adjusted$	0.172***	0.114***	0.154***	0.192***
		(0.062)	(0.036)	-0.032	(0.021)

Notes: Robust standard errors are shown in parentheses and significance levels are indicated by: *** 1%, ** 5%, * 10%, + 15%. This table shows the first stage OLS regressions to derive the R^2 as IEOp measure for conducting the DiD estimation approach, with the results shown in the first sub-panel in Table A.11. The dependent variable is *stdpvread*3, i.e. **standardized PISA reading test scores** for each test year with respect to students in *Gymnasium* that are part of the representative *grade-based* German PISA test cohort across the respective *Model Base* period (2003-2012) (Footnote 24). Columns (1) to (2) showing the results for Control Group (**C**) provide first-step regressions for the *Before-reform period (2003-2006)* in column 1 and *After-reform period (2009-2012)* in column 2. Columns (3) to (4) provide first-step regression results for Treatment Group (**T**) with *Before-reform period (2003-2006)* results in column 3 and *After-reform period (2009-2012)* results in column 4. Background variables used to derive R^2 are explained in Section 3.3 and listed in four groups with a total of seven subgroups (compare Appendix A.5.3). Observations are weighted according to the provided **population weights**. Standard errors are clustered at the federal state level, and inflated by the estimated measurement error in test scores (compare Appendix A.5.1 on their computation).
Source: Author's calculations based on PISA 2003, 2006, 2009 and 2012.

Table A.7: Main Results for *Model Base*: 1^{st} step to Derive IEOp Measure for *Mathematics*

DEPENDENT VARIABLE:		**Control Group (C)**		**Treatment Group (T)**	
	Mathematics Test Scores in PISA (STDPVMATH3)	Before (2003-2006)	After (2009-2012)	Before (2003-2006)	After (2009-2012)
CONTROL: **Individual Characteristics (IC)**					
i)	Female	-0.662***	-0.533***	-0.464***	-0.450***
		(0.116)	(0.066)	(0.044)	(0.046)
	Age in years	-0.110	-0.313***	-0.209***	-0.220***
		(0.112)	(0.072)	(0.045)	(0.037)
ii)	Migration Background	-0.084	-0.127	-0.128*	-0.162***
		(0.185)	(0.135)	(0.071)	(0.048)
	NO German spoken at home	-0.373	-0.170	-0.082	-0.198***
		(0.337)	(0.191)	(0.104)	(0.077)
CONTROL: **Parental Characteristics (PC)**					
iii)	Parental Education: [Base: *ISCED-level (3-4)*]				
	# at most lower sec. educ. (ISCED-level (1-2))	-0.454*	0.133	-0.169**	-0.092
		(0.275)	(0.152)	(0.080)	(0.069)
	# tertiary educ. (ISCED-level (5-6))	-0.230**	0.057	0.020	-0.001
		(0.101)	(0.148)	(0.045)	(0.037)
CONTROL: **Socio-Economic Status (SES)**					
iv)	No. of books in household [Base: *101-500*]				
	# max 10 books	0.342	-0.411	-0.398**	-0.316***
		(0.243)	(0.305)	(0.155)	(0.103)
	# 11-100 books	-0.120	-0.078	-0.253***	-0.134***
		(0.078)	(0.129)	(0.061)	(0.046)
	# more than 500 books	0.234	0.185	0.074	0.116***
		(0.149)	(0.120)	(0.056)	(0.037)
v)	Highest ISEI-level of Parental Jobs	0.007	0.002	0.002	0.004***
		(0.005)	(0.002)	(0.002)	(0.001)
CONTROL: **Family Characteristics (FC)**					
vi)	Family Structure [Base: *No*]				
	single parent household	0.058	0.195*	0.046	0.112**
		(0.176)	(0.103)	(0.054)	(0.052)
vii)	Father: Employment [Base: *Full-time (FT)*]				
	# part-time (PT)	-0.238	-0.413*	0.007	-0.117*
		(0.278)	(0.218)	(0.085)	(0.061)
	# unemployed (UE)	0.075	0.100	-0.210	0.109
		(0.353)	(0.428)	(0.138)	(0.139)
	# out-of-labor force (OLF)	-0.044	-0.174	-0.026	-0.028
		(0.308)	(0.146)	(0.143)	(0.106)
	Mother: Employment [Base: *Full-time (FT)*]				
	# part-time (PT)	0.096	0.025	-0.001	0.045
		(0.068)	(0.125)	(0.065)	(0.052)
	# unemployed (UE)	0.212	0.143	-0.003	0.232**
		(0.161)	(0.597)	(0.080)	(0.092)
	# out-of-labor force (OLF)	-0.072	0.203	-0.037	0.082
		(0.154)	(0.125)	(0.083)	(0.072)
	Constant	1.824	4.853***	3.581***	3.539***
		(1.755)	(1.121)	(0.711)	(0.583)
	School FE	yes	yes	yes	yes
	Observations	346	608	2356	3329
	R^2	0.353***	0.189***	0.267***	0.248***
		(0.060)	(0.041)	(0.033)	(0.026)
	$R^2 - adjusted$	0.294***	0.144***	0.244***	0.228***
		(0.065)	(0.043)	(0.034)	(0.027)

Notes: Robust standard errors are shown in parentheses and significance levels are indicated by: *** 1%, ** 5%, * 10%, + 15%. This table shows the first stage OLS regressions to derive the R^2 as IEOp measure for conducting the DiD estimation approach, with the results shown in the second sub-panel in Table A.11. The dependent variable is *stdpvmath3*, i.e. **standardized PISA mathematics test scores** for each test year with respect to students in *Gymnasium* that are part of the representative *grade-based* German PISA test cohort across the respective *Model Base* period (2003-2012) (Footnote 24). Columns (1) to (2) showing the results for Control Group (**C**) provide first-step regressions for the *Before-reform period (2003-2006)* in column 1 and *After-reform period (2009-2012)* in column 2. Columns (3) to (4) provide first-step regression results for Treatment Group (**T**) with *Before-reform period (2003-2006)* results in column 3 and *After-reform period (2009-2012)* results in column 4. Background variables used to derive R^2 are explained in Section 3.3 and listed in four groups with a total of seven subgroups (compare Appendix A.5.3). Observations are weighted according to the provided **population weights**. Standard errors are clustered at the federal state level and inflated by the estimated measurement error in test scores (compare Appendix A.5.1 on their computation).
Source: Author's calculations based on PISA 2003, 2006, 2009 and 2012.

Table A.8: Main Results for *Model Base*: 1^{st} step to Derive IEOp Measure for *Science Scores*

DEPENDENT VARIABLE:		**Control Group (C)**		**Treatment Group (T)**	
	Science Test Scores in PISA (STDPVSCIE3)	Before (2003-2006)	After (2009-2012)	Before (2003-2006)	After (2009-2012)
CONTROL: **Individual Characteristics (IC)**					
i)	Female	-0.509***	-0.340***	-0.354***	-0.287***
		(0.113)	(0.064)	(0.037)	(0.040)
	Age in years	-0.093	-0.252***	-0.163***	-0.160***
		(0.118)	(0.082)	(0.062)	(0.053)
ii)	Migration Background	-0.297*	-0.287***	-0.087	-0.202***
		(0.163)	(0.096)	(0.082)	(0.054)
	NO German spoken at home	-0.347	-0.158	-0.222**	-0.195***
		(0.403)	(0.221)	(0.092)	(0.067)
CONTROL: **Parental Characteristics (PC)**					
iii)	Parental Education: [Base: *ISCED-level (3-4)*]				
	# at most lower sec. educ. (ISCED-level (1-2))	-0.529**	0.076	-0.259***	-0.047
		(0.243)	(0.184)	(0.092)	(0.061)
	# tertiary educ. (ISCED-level (5-6))	-0.153	0.116	0.030	0.005
		(0.123)	(0.126)	(0.055)	(0.040)
CONTROL: **Socio-Economic Status (SES)**					
iv)	No. of books in household [Base: *101-500*]				
	# max 10 books	0.101	-0.397	-0.306**	-0.525***
		(0.413)	(0.281)	(0.149)	(0.068)
	# 11-100 books	-0.122	-0.114	-0.282***	-0.200***
		(0.106)	(0.112)	(0.067)	(0.045)
	# more than 500 books	0.175	0.120	0.149**	0.163***
		(0.147)	(0.111)	(0.060)	(0.038)
v)	Highest ISEI-level of Parental Jobs	0.009**	0.003	0.003*	0.003**
		(0.004)	(0.003)	(0.002)	(0.002)
CONTROL: **Family Characteristics (FC)**					
vi)	Family Structure [Base: *No*]				
	Single Parent Household	0.051	0.207**	0.008	0.122*
		(0.241)	(0.084)	(0.076)	(0.063)
vii)	Father: Employment [Base: *Full-time (FT)*]				
	# part-time (PT)	-0.162	-0.223	-0.054	-0.141*
		(0.207)	(0.190)	(0.116)	(0.074)
	# unemployed (UE)	0.192	0.161	-0.022	0.091
		(0.426)	(0.378)	(0.125)	(0.166)
	# out-of-labor force (OLF)	0.036	-0.021	0.019	0.068
		(0.290)	(0.238)	(0.129)	(0.098)
	Mother: Employment [Base: *Full-time (FT)*]				
	# part-time (PT)	-0.083	-0.059	-0.016	0.011
		(0.097)	(0.098)	(0.064)	(0.046)
	# unemployed (UE)	0.184	0.101	-0.021	0.195*
		(0.156)	(0.307)	(0.101)	(0.102)
	# out-of-labor force (OLF)	-0.218	0.087	-0.089	0.006
		(0.139)	(0.113)	(0.067)	(0.059)
	Constant	1.593	3.933***	2.784***	2.614***
		(1.918)	(1.287)	(1.006)	(0.803)
	School FE	yes	yes	yes	yes
	Observations	346	608	2356	3329
	R^2	0.363***	0.173***	0.214***	0.209***
		(0.052)	(0.048)	(0.025)	(0.023)
	$R^2 - adjusted$	0.304***	0.125**	0.190***	0.188***
		(0.057)	(0.051)	(0.026)	(0.023)

Notes: Robust standard errors are shown in parentheses and significance levels are indicated by: *** 1%, ** 5%, * 10%, + 15%. This table shows the first stage OLS regressions to derive the R^2 as IEOp measure for conducting the DiD estimation approach, with the results shown in the third sub-panel in Table A.11. The dependent variable is $stdpvscie3$, i.e. **standardized PISA science test scores** for each test year with respect to students in *Gymnasium* that are part of the representative *grade-based* German PISA test cohort across the respective *Model Base* period (2003-2012) (Footnote 24). Columns (1) to (2) showing the results for Control Group (**C**) provide first-step regressions for the *Before-reform period (2003-2006)* in column 1 and *After-reform period (2009-2012)* in column 2. Columns (3) to (4) provide first-step regression results for Treatment Group (**T**) with *Before-reform period (2003-2006)* results in column 3 and *After-reform period (2009-2012)* results in column 4. Background variables used to derive R^2 are explained in Section 3.3 and listed in four groups with a total of seven subgroups (compare Appendix A.5.3). Observations are weighted according to the provided **population weights**. Standard errors are clustered at the federal state level and inflated by the estimated measurement error in test scores (compare Appendix A.5.1 on their computation).
Source: Author's calculations based on PISA 2003, 2006, 2009, and 2012.

Table A.9: *Robust Model* for T vs. C and C1

Subject	**Main Control Group C**			**Extended Control Group C1**		
Reading	**C**	**T**	**Δ (T-C)**	**C1**	**T**	**Δ (T-C1)**
Before	0.242	0.180	-0.062	0.163	0.180	0.016
	(0.057)	(0.031)	(0.065)	(0.032)	(0.031)	(0.044)
After	0.161	0.195	0.035	0.183	0.195	0.012
	(0.060)	(0.034)	(0.069)	(0.030)	(0.034)	(0.046)
Change in R^2	-0.081	0.016	**0.097**	0.020	0.016	**-0.004**
	(0.083)	(0.046)	(0.095)	(0.044)	(0.046)	(0.064)
Mathematics	**C**	**T**	**Δ (T-C)**	**C1**	**T**	**Δ (T-C1)**
Before	0.353	0.267	-0.086	0.216	0.267	0.052
	(0.060)	(0.033)	(0.068)	(0.029)	(0.033)	(0.044)
After	0.270	0.227	-0.043	0.233	0.227	-0.006
	(0.073)	(0.037)	(0.082)	(0.033)	(0.037)	(0.050)
Change in R^2	-0.084	-0.040	**0.043**	0.017	-0.040	**-0.057**
	(0.094)	(0.049)	(0.107)	(0.044)	(0.049)	(0.066)
Science	**C**	**T**	**Δ (T-C)**	**C1**	**T**	**Δ (T-C1)**
Before	0.363	0.215	-0.148	0.205	0.215	0.010
	(0.052)	(0.025)	(0.058)	(0.024)	(0.025)	(0.035)
After	0.257	0.201	-0.056	0.215	0.201	-0.014
	(0.067)	(0.034)	(0.075)	(0.039)	(0.034)	(0.051)
Change in R^2	-0.106	-0.014	**0.092**	0.010	-0.014	**-0.024**
	(0.085)	(0.042)	(0.095)	(0.046)	(0.042)	(0.062)

Notes: Table entries are R^2 measures of IEOp (Equation (7)). Robust standard errors are in parentheses and were calculated using replication weights following the method as explained in Appendix A.5.1, clustering at the federal state level. DiD results are estimated according to Equation (9) taking into account population weights. Positive changes in R^2 indicate increasing IEOp or decreasing EEOp and vice versa.
Background variables used to derive R^2:
(i) Individual Characteristics (IC) I: *age* and *gender*
(ii) Individual Characteristics (IC) II: *language spoken at home*; *migration background* (based on par. birth place)
(iii) Parental Characteristics (PC): *highest parents' qualification (ISCED-level 1-2/ISCED 3-4/ISCED 5-6)*
(iv) Socio-economic Status (SES) I: *no. of books in household (max. 11, 11-100, 101-500, more than 500)*
(v) Socio-economic Status (SES) II : *highest ISEI-level-index [0-90] of job in the family*
(vi) Family Characteristics (FC) I: *family structure - growing up in single parent household?*
(vii) Family Characteristics (FC) II: *mother/father: working part-time (PT) - unemployed (UE) - out of labor force (OLF)*
Compare: Due to space constraints first-step regressions for **T** vs. **C/C1** for the time period 2003-2006 vs. 2009 have been omitted, but they are available upon request from the author.
Source: Author's calculations based on PISA 2003, 2006, and 2009 (compare Section 3.1).

Table A.10: Robustness Checks: Placebo Tests (2003-2006) T vs. C

Subject	**With R^2 Measure**			**With $R^2_{adjusted}$ Measure**		
Reading	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before (2003)	0.288	0.139	-0.149	0.173	0.101	-0.071
	(0.115)	(0.047)	(0.125)	(0.134)	(0.049)	(0.143)
After (2006)	0.284	0.229	-0.055	0.178	0.199	0.022
	(0.072)	(0.039)	(0.082)	(0.083)	(0.041)	(0.092)
Change in R^2	-0.004	0.090	**0.094**	0.005	0.098	**0.093**
	(0.136)	(0.061)	(0.149)	(0.158)	(0.064)	(0.170)
Mathematics	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before (2003)	0.353	0.235	-0.118	0.249	0.202	-0.047
	(0.109)	(0.047)	(0.119)	(0.127)	(0.049)	(0.136)
After (2006)	0.362	0.293	-0.069	0.267	0.266	-0.001
	(0.054)	(0.048)	(0.072)	(0.062)	(0.050)	(0.079)
Change in R^2	0.009	0.058	**0.049**	0.018	0.064	**0.046**
	(0.122)	(0.067)	(0.139)	(0.141)	(0.070)	(0.157)
Science	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before (2003)	0.384	0.186	-0.198	0.285	0.151	-0.134
	(0.080)	(0.037)	(0.088)	(0.093)	(0.038)	(0.100)
After (2006)	0.383	0.251	-0.132	0.291	0.222	-0.069
	(0.074)	(0.037)	(0.083)	(0.085)	(0.039)	(0.093)
Change in R^2	-0.002	0.064	**0.066**	0.006	0.071	**0.065**
	(0.109)	(0.052)	(0.121)	(0.126)	(0.054)	(0.137)

Notes: Table entries are R^2 measures of IEOp (Equation (7)). Robust standard errors are in parentheses and were calculated using replication weights following the method as explained in Appendix A.5.1, clustering at the federal state level. DiD results are estimated according to Equation (9) taking into account population weights and the indicated school fixed effects. Positive changes in R^2 indicate increasing IEOp or decreasing EEOp and vice versa.
Background variables used to derive R^2:
(i) Individual Characteristics (IC) I: *age* and *gender*
(ii) Individual Characteristics (IC) II: *language spoken at home*; *migration background* (based on par. birth place)
(iii) Parental Characteristics (PC): *highest parents' qualification (ISCED-level 1-2/ISCED 3-4/ISCED 5-6)*
(iv) Socio-economic Status (SES) I: *no. of books in household (max. 11, 11-100, 101-500, more than 500)*
(v) Socio-economic Status (SES) II : *highest ISEI-level-index[0-90] of job in the family*
(vi) Family Characteristics (FC) I: *family structure - growing up in single parent household?*
(vii) Family Characteristics (FC) II: *mother/father working: part-time (PT) - unemployed (UE) - out of labor force (OLF)*
Compare: Due to space constraints first-step regressions for **T** vs. **C** for the time period 2003 vs. 2006 have been omitted, but they are available upon request from the author.
Source: Author's calculations based on PISA 2003 and 2006.

Table A.11: Robustness Check of Main Results: Testing Potential Sorting across Schools

Subject	*Model Base (2003-2012)* - **T** vs. **C** — (**Figure A.3**)					
	Main: School Fixed Effects			**Robustness: State Fixed Effects**		
Reading	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before	0.242	0.180	-0.062	0.180	0.121	-0.059
	(0.057)	(0.031)	(0.065)	(0.054)	(0.025)	(0.060)
After	0.162	0.213	0.051	0.131	0.140	0.009
	(0.034)	(0.020)	(0.039)	(0.034)	(0.019)	(0.039)
Change in R^2	-0.080	0.033	**0.113**	-0.049	0.019	**0.068**
	(0.066)	(0.037)	(0.076)	(0.064)	(0.032)	(0.071)
Mathematics	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before	0.353	0.267	-0.086	0.300	0.172	-0.128
	(0.060)	(0.033)	(0.068)	(0.059)	(0.026)	(0.064)
After	0.190	0.249	0.060	0.160	0.190	0.030
	(0.040)	(0.027)	(0.048)	(0.040)	(0.025)	(0.047)
Change in R^2	-0.164	-0.018	**0.146**	-0.140	0.018	**0.158**
	(0.072)	(0.042)	(0.083)	(0.071)	(0.036)	(0.080)
Science	**C**	**T**	**Δ (T-C)**	**C**	**T**	**Δ (T-C)**
Before	0.363	0.215	-0.148	0.295	0.148	-0.147
	(0.052)	(0.025)	(0.058)	(0.055)	(0.022)	(0.059)
After	0.173	0.210	0.037	0.128	0.142	0.013
	(0.048)	(0.023)	(0.053)	(0.038)	(0.019)	(0.042)
Change in R^2	-0.190	-0.005	**0.185**	-0.166	-0.006	**0.160**
	(0.071)	(0.034)	(0.079)	(0.066)	(0.029)	(0.073)

Notes: Table entries are R^2 measures of IEOp (Equation (7)). Robust standard errors are in parentheses and were calculated using replication weights following the method as explained in Appendix A.5.1, clustering at the federal state level. DiD results are estimated according to Equation (9) taking into account population weights and the indicated fixed effects. Positive changes in R^2 indicate increasing IEOp or decreasing EEOp and vice versa.
Background variables used to derive R^2:
(i) Individual Characteristics (IC) I: *age* and *gender*
(ii) Individual Characteristics (IC) II: *language spoken at home*; *migration background* (based on parental birth place)
(iii) Parental Characteristics (PC): *highest parents' qualification (ISCED-level 1-2/ISCED-level 3-4/ISCED-level 5-6)*
(iv) Socio-economic Status (SES) I: *no. of books in household (max. 11, 11-100, 101-500, more than 500)*
(v) Socio-economic Status (SES) II : *highest ISEI-level-index[0-90] of job in the family*
(vi) Family Characteristics (FC) I: *family structure - growing up in single parent household?*
(vii) Family Characteristics (FC) II: *mother/father working part-time (PT) - mother/father unemployed (UE) - mother/father out of labor force (OLF)*
Compare: The first-step regressions of the setting: treatment group **T** vs. control group **C** with school-fixed effects are provided in Table A.6, A.7 and A.8 in Appendix A.2.
Source: Author's calculations based on PISA 2003, 2006, 2009, and 2012.

Table A.12: Difference-in-Differences Results: Overview Control Group C

(1) Outcome	(2) Treatment	(3) Control	(4) Control set	(5) R^2 adj.	(6) R^2
reading	T	C	1	0.060	0.063
reading	T	C	2	0.073	0.078
reading	T	C	3	0.081	0.090
reading	T	C	4	0.086	0.095
reading	T	C	5	0.076	0.087
reading	T	C	6	0.096	0.113
reading	T1	C	1	0.058	0.062
reading	T1	C	2	0.072	0.078
reading	T1	C	3	0.080	0.089
reading	T1	C	4	0.085	0.095
reading	T1	C	5	0.075	0.086
reading	T1	C	6	0.095	0.112
reading	T2	C	1	0.036	0.041
reading	T2	C	2	0.044	0.051
reading	T2	C	3	0.051	0.062
reading	T2	C	4	0.056	0.067
reading	T2	C	5	0.046	0.059
reading	T2	C	6	0.067	0.087
mathematics	T	C	1	0.109	0.110
mathematics	T	C	2	0.121	0.124
mathematics	T	C	3	0.127	0.131
mathematics	T	C	4	0.134	0.139
mathematics	T	C	5	0.134	0.140
mathematics	T	C	6	0.136	0.146
mathematics	T1	C	1	0.101	0.102
mathematics	T1	C	2	0.114	0.117
mathematics	T1	C	3	0.120	0.125
mathematics	T1	C	4	0.128	0.133
mathematics	T1	C	5	0.127	0.133
mathematics	T1	C	6	0.129	0.139
mathematics	T2	C	1	0.097	0.099
mathematics	T2	C	2	0.106	0.110
mathematics	T2	C	3	0.109	0.115
mathematics	T2	C	4	0.117	0.123
mathematics	T2	C	5	0.116	0.123
mathematics	T2	C	6	0.120	0.132
science	T	C	1	0.153	0.153
science	T	C	2	0.156	0.158
science	T	C	3	0.160	0.164
science	T	C	4	0.169	0.172
science	T	C	5	0.162	0.166
science	T	C	6	0.177	0.185
science	T1	C	1	0.156	0.155
science	T1	C	2	0.159	0.160
science	T1	C	3	0.164	0.167
science	T1	C	4	0.173	0.176
science	T1	C	5	0.165	0.169
science	T1	C	6	0.182	0.189
science	T2	C	1	0.135	0.136
science	T2	C	2	0.135	0.137
science	T2	C	3	0.137	0.142
science	T2	C	4	0.144	0.149
science	T2	C	5	0.136	0.143
science	T2	C	6	0.155	0.164

Notes: This table shows *T/T1/T2 vs. C* for all 3 test score domains with *school* fixed effects and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$. Note that column (6) shows the DiD results using R^2, column (5) shows the same but using adjusted R^2 as IEOp measure.

Table A.13: Difference-in-Differences Results: Overview Control Group C-NT

(1) Outcome	(2) Treatment	(3) Control	(4) Control set	(5) R^2 adj.	(6) R^2
reading	T	C-NT	1	0.027	0.025
reading	T	C-NT	2	0.030	0.028
reading	T	C-NT	3	0.021	0.019
reading	T	C-NT	4	0.025	0.023
reading	T	C-NT	5	0.023	0.021
reading	T	C-NT	6	0.028	0.025
reading	T1	C-NT	1	0.025	0.023
reading	T1	C-NT	2	0.030	0.027
reading	T1	C-NT	3	0.021	0.018
reading	T1	C-NT	4	0.024	0.022
reading	T1	C-NT	5	0.022	0.019
reading	T1	C-NT	6	0.028	0.024
reading	T2	C-NT	1	0.003	0.002
reading	T2	C-NT	2	0.002	0.001
reading	T2	C-NT	3	-0.008	-0.010
reading	T2	C-NT	4	-0.006	-0.006
reading	T2	C-NT	5	-0.007	-0.007
reading	T2	C-NT	6	0.000	-0.002
mathematics	T	C-NT	1	-0.005	-0.006
mathematics	T	C-NT	2	-0.001	-0.003
mathematics	T	C-NT	3	-0.009	-0.010
mathematics	T	C-NT	4	0.000	-0.002
mathematics	T	C-NT	5	0.016	0.014
mathematics	T	C-NT	6	0.022	0.018
mathematics	T1	C-NT	1	-0.013	-0.014
mathematics	T1	C-NT	2	-0.008	-0.010
mathematics	T1	C-NT	3	-0.015	-0.017
mathematics	T1	C-NT	4	-0.006	-0.008
mathematics	T1	C-NT	5	0.009	0.007
mathematics	T1	C-NT	6	0.015	0.011
mathematics	T2	C-NT	1	-0.017	-0.017
mathematics	T2	C-NT	2	-0.016	-0.016
mathematics	T2	C-NT	3	-0.026	-0.027
mathematics	T2	C-NT	4	-0.018	-0.018
mathematics	T2	C-NT	5	-0.002	-0.003
mathematics	T2	C-NT	6	0.006	0.003
science	T	C-NT	1	0.040	0.038
science	T	C-NT	2	0.039	0.036
science	T	C-NT	3	0.026	0.023
science	T	C-NT	4	0.034	0.031
science	T	C-NT	5	0.038	0.034
science	T	C-NT	6	0.052	0.047
science	T1	C-NT	1	0.042	0.040
science	T1	C-NT	2	0.041	0.039
science	T1	C-NT	3	0.030	0.026
science	T1	C-NT	4	0.038	0.035
science	T1	C-NT	5	0.041	0.038
science	T1	C-NT	6	0.057	0.051
science	T2	C-NT	1	0.022	0.021
science	T2	C-NT	2	0.017	0.016
science	T2	C-NT	3	0.003	0.001
science	T2	C-NT	4	0.009	0.008
science	T2	C-NT	5	0.013	0.011
science	T2	C-NT	6	0.030	0.026

Notes: This table shows *T/T1/T2 vs. C-NT Never-Takers* for all 3 test score domains with *school* fixed effects and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$. Note that columns (6) shows the DiD results using R^2, columns (5) shows the same but using adjusted R^2 as IEOp measure.

A.3 Further Details on the G-8 Reform

A.3.1 Related Literature on the Reform

Despite the public controversy over the G-8 reform that has even induced some federal states to reverse it (last column in Table A.1 in Appendix A.2), few studies have evaluated the G-8 reform and its effects on outcomes such as educational achievement. To begin with, studies have analyzed the reform by comparing G-8 model and G-9 model cohorts within one federal state.

In most federal states, the statistical offices have conducted studies comparing students' results in central exit examinations (*Abitur*) in the *double cohort*, i.e. the year when the last G-9 model and the first G-8 model cohort graduated from *Gymnasium* (Figure 1). Generally, these statistical evaluations have found no systematic difference in central exit exam outcomes between students with eight or nine years of schooling. However, as grades in final exams are a useful performance indicator only within the same student cohort, GPA comparisons across years have limitations. In fact, school exams are usually graded based on a relative performance distribution of each cohort. Thus, using grades as outcome variable limits what can be learned about the reform's impact on cognitive skills, in contrast to standardized test scores (see Appendix A.4.1).

For the federal state of Saxony-Anhalt (ST), a small series of papers has analyzed different aspects of the G-8 reform (Thiel et al., 2014; Büttner & Thomsen, 2015; T. Meyer & Thomsen, 2016). In summary, they examine the reform's effects on academic achievement in central exit examinations 2007, when the *double cohort* graduated in ST (Table A.1). Findings show that - due to more intense schooling - exam results significantly deteriorated for mathematics but remained unaffected for German literature. This suggests that learning intensity ratios differ across subjects. Moreover, no significant negative effects on students' soft skills have been detected, opposing claims that increased learning intensity and accordingly reduced time for non-school related activities may have adversely affected non-cognitive skill formation. In line with this result, Quis and Reif (2017) show that the more intense schooling experience had only limited impact on students' health. However, due to reduced leisure time, G-8 model students were less able to relax and slightly more stressed compared to their peers in the G-9 model. Finally, T. Meyer and Thomsen (2016) find some heterogenous effects on post-secondary school education decisions. For instance, they find significant delays in the starting dates for a first university degree for female students who graduated from a G-8 model school. Instead, they were more likely to first complete a type of vocational education. Moreover, T. Meyer and Thomsen (2016) reveal that despite the G-8 reform, students continue to pursue their hobbies. However, they tend to work less outside of school. Recently, T. Meyer et al. (2018) extended the same analysis to cover all federal states of Germany. Their new findings remain similar to T. Meyer and Thomsen (2016). Conducting a comparable analysis for all German federal states, Marcus and Zambre (2019) show that the G-8 reform reduced enrollment rates at university and increased the likelihood of affected students to switch their major degree. Using the a similar setup as in Marcus and Zambre (2019), Huebener and Marcus (2017) evaluate the impact of the G-8 reform on GPA and graduation rate using aggregated adiministrative data on the full population of students considering all states in Germany. Their results indicate that the reform had adverse effects on educational outcomes as they find significant negative effects of the reform on average GPA but none on the overall graduation rate.

Recently, a few papers have started to use more representative data that are more independent from school system related characteristics or relative performance measurement issues arising with marks at school (e.g. PISA data). Moreover, identifying the G-8 reform effect by exploiting the variation in its implementation across states and over time, this approach allows overcoming the shortcomings of previous studies. For instance, the following two papers related to this project exploit the reform setting using standardized PISA test scores for academic-track students as educational outcome variable.[36]

Andrietti and Su (2019) uses this representative dataset in order to exploit the G-8 reform for conducting a Difference-in-Differences (DiD) estimation. They find that the average treatment effect of the reform is significant and positive in all three educational outcomes (mathematics, reading and science). Treated students in a G-8 model experience an improvement of about 0.067 to 0.089 standard deviations in PISA test scores. In a similar manner to Huebener et al. (2017), Andrietti and Su (2019) extend their analysis by conducting a quantile DiD to investigate heterogenous reform impacts. Their findings indicate that the reform could have worsened inequality as only high-performing students tend to have benefited from the reform.[37]

Huebener et al. (2017) use state regulations of timetables for secondary school to show that, due to the G-8 reform, weekly instruction hours for the average treated student increased by about 6.5 percent over a period of five years. They suggest that increased instruction time improved the average student performance in all three PISA test domains. However, the effect size is small, with about six percentage points of a standard deviation in scores. Moreover, the effects are insignificant for low-performing students, while their high-performing peers experience significant, but small, positive effects. This suggests that the performance gap among students in *Gymnasium* widened which is in line with the results of Andrietti and Su (2019). In that regard, Huebener et al. (2017) focus on the increased instruction time effect, whereas Andrietti and Su (2019) puts more emphasis on the increased learning intensity aspect of the reform.

In this paper, I use similar data as Huebener et al. (2017) with PISA test scores from 2000 to 2012. However, my focus is on analyzing the effects of increased learning intensity on educational outcomes in response to the G-8 reform (interpreting the reform similar to Andrietti and Su (2019)). While these studies estimate the direct reform effect on test scores, they do not tackle the question of whether increasing learning intensity may have changed Inequality of Educational Opportunity (IEOp). In this paper, I shift focus in the analysis of the G-8 reform onto distributional concerns, i.e. its consequences on IEOp. In other words, I answer the question of whether the G-8 reform is *selective*, i.e. a reform that at least maintains test score results, but at the same time increases IEOp; or whether the G-8 reform is *inclusive*, i.e. a reform that at least maintains test score results while decreasing IEOp (Checchi & van de Werfhorst, 2018). Thus, I am among the first to evaluate the G-8 reform based on Germany-specific PISA data in order to analyze its impact on IEOp.

[36] Back in 2012, Camarero Garcia (2012) appears to have been the first to combine the usage of PISA test scores as an outcome variable to analyze the effects of the G-8 reform on cognitive skills in a Difference-in-Differences (DiD) estimation framework, finding a positive effect of about 0.15 standard deviations in test scores similar to the later results by Andrietti and Su (2019). An extensive discussion on the impact of the G-8 reform on test scores is also conducted by Homuth (2012), whose findings are in line with the results of Camarero Garcia (2012) and Andrietti and Su (2019) because he shows that, on average, the G-8 reform had positive effects on reading skills.

[37] Andrietti and Su (2019a) argue that the "preparedness" of students influences the reform's effects on educational outcomes.

A.3.2 The Reform Debate

The first PISA-study in 2000 received broad public attention in Germany because it revealed that German students achieved test scores below the average of OECD countries (the so-called "PISA-shock"). Debates over how to improve the German school system ensued (e.g. Davoli and Entorf (2018)). Among the reform proposals, shortening the academic secondary school track (*Gymnasium*) from nine to eight years, the G-8 reform, remains controversial to this day. The last column in Appendix Table A.1 gives an overview on the status quo of the reform as of school year 2015/16.

Three main reasons were given for introducing the G-8 reform. First, it was intended to reduce the relatively high age of university graduates in Germany. This was said to increase their competitiveness in the labor market compared to the (on average) younger graduates in other OECD countries OECD (2005a). Furthermore, with students entering the job market one year earlier, working lifetime would be extended, augmenting social security contributions. Thus, the reform was said to contribute to stabilizing the social security system of a society facing demographic change. Second, as the most successful countries in the PISA test ranking, such as Finland, had a school system of twelve years, reduced schooling appeared to be both successful and efficient. Third, the G-8 reform was seen as a necessary adjustment of secondary school with regards to harmonizing tertiary education across Europe. As Büttner and Thomsen (2015) illustrate, the reform of shortening secondary school duration was also enacted in the context of the *Bologna Process*. This initiative aims to create a European Higher Education Area (EHEA) providing a more comparable, flexible European framework for tertiary education. Therefore, adjusting secondary school duration towards the average among other European nations was regarded to be sensible. Finally, the reform was said to serve as an incentive for then younger school graduates to strive for obtaining a university degree, bringing Germany's below average rate of university graduates per birth cohort in line with other OECD countries.

However, opponents of the reform claimed that the intensified educational experience might worsen the human capital skill formation for affected students. Parental complaints about increased stress for students (due to less free time) revealed further concerns. In fact, many parents said that compressed and intensified schooling might have negative impacts for their children, on both academic performance and the development of non-cognitive skills which are typically formed by recreational activities Thiel et al. (2014). However, the majority of East Germans support shortened duration of the academic track, whereas the opposite is true across West German federal states which only recently adopted the G-8 model Wössmann et al. (2015).

A.4 Further Details on the Data Used

A.4.1 Background Information on the PISA Data

Every three years since 2000, the OECD conducts the PISA study in order to measure the performance of 15 year-old students with respect to three basic competencies (*Life skills*), namely *reading*, *mathematical*, and *scientific* literacy. These skills are regarded to be of special importance for a person's future success and are tested when students approach the end of compulsory schooling age (cf. OECD (2010); OECD (2013a)). The idea of PISA is to evaluate the ability to apply knowledge, as acquired through the curriculum at school in the three tested domains, for solving real-world problems. This means to test the level of skills that students achieve until compulsory schooling ends and that are essential for participating in modern society (OECD, 2001).[38] Apart from cognitive test scores, PISA collects rich information on family and school characteristics. This is based on questionnaires that students, their parents, teachers, and school's principals fill out.

Concerning the PISA procedure, for each test cycle, the OECD chooses an international contractor who is responsible for the test's design and comparability across countries (e.g. that test questions are robust to cultural bias) and over time (making trend analysis possible (OECD, 2009b)). On the country level, a PISA National Project Manager is chosen to make sure that the test is conducted according to the strict OECD quality guidelines. The test procedure itself resembles a *two-stage stratified randomized survey test design*. First, as a primary sampling unit, schools with eligible students are randomly selected (with a minimum of 150 schools in each country) to create a representative sample of all school types across all regions within a country. Then, as second-step sampling units, eligible students (15-year-olds)[39] are randomly selected within the sampled schools to reach a minimum of 4500 observations. Each student within a school receives distinct combinations of approved test questions on all three PISA domains.[40] The level and scope of the test is identical for each student independent of the secondary school type attended. The paper-based test takes two hours, with additional 30 minutes dedicated for students to complete the questionnaire on their socio-economic background, school and on their attitude, motivation, or aspiration. After the test has been evaluated on the national level (supervised by the international contractor), the OECD publishes a cross-country comparison of official test scores.

To have comparable measures of latent ability in each PISA domain across and within countries, the raw answers to test questions, *items*, undergo some processing (cf. OECD (2005b), OECD (2009a), OECD (2012)).The so-called *Item Response theory (IRT)* is used to back out the distribution of the latent variable, cognitive skills (as measured by test scores), from individual *item* responses, taking into account the particular difficulty of an *item*. However, to address the issue of small-sample measurement error, for instance, as not all students answer all *items*, *Plausible Values* of test results are provided for each student.

[38] The underlying question of PISA is "*What is important for citizens to know and be able to do?*". More generally, in PISA the concept of "literacy" refers to "*students' capacity to apply knowledge and skills in key subjects, and to analyze, reason and communicate effectively as they identify, interpret and solve problems in a variety of situations*". For specific definitions of each tested domain, I refer to OECD (2004) and in particular to chapter 1 of OECD (2009b).

[39] This includes students who were aged between 15 years and 3 months and 16 years and 2 months at the beginning of the assessment period (plus/minus 1 month), who were enrolled in an educational institution (grade 7 or higher) (OECD, 2013b).

[40] For details on the international PISA test procedure, I refer to section 2 in Lavy (2015) and to publications on the PISA Assessment Framework or to one of the Technical Reports on the test, e.g. OECD (2013a) and OECD (2012).

First, the marginal distribution of the latent variable conditional on the *item* responses and a set of observables is estimated. Thus, for each student a probability distribution of test scores based on their answers is estimated. Second, M draws from this distribution are taken to become the *Plausible Values* of a student's test score. For PISA, in each test cycle, five *Plausible Values* are provided for each student in all three test domains ($M = 5$). Conducting estimations with PISA test scores, the OECD (2010) suggests estimating any statistic s by using each of M *Plausible Values* datasets separately (getting $\hat{s}_m$) and then averaging them over M to get a final estimate $\hat{s}$. After this IRT-adjustment, the plausible test scores are standardized, as follows:

$$y_{ij} = \hat{\mu} + \frac{\hat{\sigma}}{\sigma}(x_{ij} - \mu) \tag{A.1}$$

where, x_{ij} is the post-IRT, pre-standardized score for student i, in country j; μ (σ) are original mean (standard deviation) across all countries in the sample of the respective test year, and $\hat{\mu}$ ($\hat{\sigma}$) denote the estimated mean (standard deviation) for a country-specific sample based on the *Plausible Values*. This generates the normalized distribution of test scores with a mean value of 500 and a standard deviation of 100 score points.[41]

The PISA test scores have neither maximum nor minimum values and there are no thresholds for passing the test, as it is designed to provide a relative measure that allows us to compare skills in the three domains across students and over time. The interpretation of test scores is eased when one compares them to a standard, such as *proficiency levels*. For instance, in mathematics, a proficiency level is supposed to consist of about 70 points. This corresponds to about two years of schooling in the average OECD country (OECD, 2013b).[42] In contrast to GPA or final exam marks in school, which are only valid as relative measures of performance in the respective school, PISA test scores have the important advantage to be a representative measure of cognitive skills for tested student cohorts across schools. Thus, PISA test scores make it possible to compare student cohorts both over time and across or within countries (federal states).[43]

Nevertheless, three doubts on the validity of PISA test scores should be considered. First, if the student population from which the test participants are selected is not complete, as some students are excluded, this would threaten representativeness. However, the sampling standards of PISA require that participating countries cannot exclude more than 5% of students from the eligible population. Permissible reasons include only special cases, such as serious illnesses or lack of language skills due to recent immigration (e.g. asylum seekers). For Germany, with at least 97% of students in the eligible age (or in the ninth grade, see Section 3.1) being part of the initial student population, exclusion is not a concern for the validity of PISA data (OECD, 2010); (OECD, 2013a).

[41] This means that across all OECD countries, the typical student scored 500 points in mathematics and about two-thirds of students in OECD countries between 400 and 600 points. Thus, 100 points constitute a huge difference in skills. To deal with difficulties in constructing meaningful measures of IEOp based on these standardized test scores, the variance is a useful index as explained by Ferreira and Gignoux (2013).

[42] For instance, in *PISA-I-2012* the average interquartile range in mathematics tests of students within OECD countries is 128 score points. However, most differences related to socio-demographic characteristics are smaller than an entire *proficiency level*. For example, across all OECD members in *PISA-I-2012*, on average, boys outscore girls in mathematics by 11 points and native students score about 34 points higher than their peers with a migration background. Socio-economically advantaged students (in top quarter of SES) score an average of 90 points higher than their disadvantaged peers (bottom quarter) (see Table II.2.4a in OECD (2013b)).

[43] For a discussion on how the meaning of grades changes due to reforms that affect curricular intensity, see also Hübner et al. (2020).

Second, one may be concerned that the actual participation rate of randomly selected students may be low, such that systematic selection may affect representativeness. However, for most developed countries the rate of compliers is above 80% for selected students and 85% for selected schools, surpassing OECD quality thresholds for the sampling process. In Germany, the participation rate of selected students is well above 80% (on average 92%), for schools, it has usually been even 100%. Moreover, there is no evidence for selection on observables for those selected who do actually not take the test (Klieme et al., 2011).

Third, another concern is that schools or, more specifically, teachers may bias comparability of scores, if they systematically train or motivate students for the test. However, based on student information about their motivation for the test and based on the information about how teachers prepared students for the test, as provided in the questionnaires of PISA test studies 2000-2012, such concerns are unwarranted (Klieme et al., 2011). Most teachers report that they tried to make students familiar with general testing strategies but did not train them specifically for the test. In fact, affected students and teachers are only informed about their participation in the PISA test around two months before the test takes place. Moreover, given the general low probability of being selected for the test and as there are no incentives for neither teachers nor students to prepare for it, potential preparation could have only very limited effects on results.[44] Moreover, Klieme et al. (2011) show that the correlation between test motivation and scores is zero (on average 0.05) and did not change as more tested students were taught in the G-8 model. Thus, test results in Germany are not systematically influenced by any preparation behavior or test motivation (Wössmann, 2010).

In conclusion, the advantages of using PISA data as measure of cognitive skills dominate any potential caveats, which is the reason I decided to use them - in line with the studies mentioned in Appendix A.3.1. For the purpose of analyzing the effect of increased learning intensity (due to the G-8 reform) on IEOp, I use the Germany-specific versions of the PISA as explained in Section 3.1.

A.4.2 Data Sources

For more information on the Germany-specific PISA data of each test cycle and the availability of these datasets, the reader is recommended to refer to the IQB.

- PISA-2000:
  Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J., & Weiß, M. (2009). *Program for International Student Assessment 2000 (PISA 2000)*. Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz [Dataset]. `http://doi.org/10.5159/IQB_PISA_2000_v1`
- PISA-2003:
  Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2007): *Program for International Student Assessment 2003 (PISA 2003)*. Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz [Dataset]. `http://doi.org/10.5159/IQB_PISA_2003_v1`

[44] Only half of the teachers indicated that they had talked with their students about PISA and those who did started not earlier than one month before the test. Vice versa, only 25% of participating students indicate to have prepared for the reading part, only 13% for mathematics, and only 8% for the science section in the test.

- PISA-2006:
  Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (2010): *Program for International Student Assessment 2006 (PISA 2006)*. Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz [Dataset]. `http://doi.org/10.5159/IQB_PISA_2006_v1`
- PISA-2009:
  Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W., & Stanat, P. (2013): *Program for International Student Assessment 2009 (PISA 2009)*. Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz [Dataset]. `http://doi.org/10.5159/IQB_PISA_2009_v1`
- PISA-2012:
  Sälzer, C., Klieme, E., Köller, O., Mang, J., Heine, J.-H., Schiepe-Tiska, A., & Müller, K. (2015): *Program for International Student Assessment 2012 (PISA 2012)*. Version: 2. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz [Dataset]. `http://doi.org/10.5159/IQB_PISA_2012_v2`

A.5 Empirical Strategy and Robustness

A.5.1 On the Computation of Standard Errors Including Replication Weights

Throughout the paper, for both steps of the DiD regressions (Section 4), standard errors are computed in a way to take into account that student performance is reported in Plausible Values (PVs) of PISA test scores. Although, taking the average of five PVs as a measure of individual performance guarantees that estimates of group level means and regression coefficients remain unbiased, measures of dispersion should consider the within-student variability in PVs.

As explained by the (OECD, 2009b), standard errors are computed by regressing five times on the dependent variable, individual test scores, thereby using all Plausible Values (PVs) in turn. For each regression, the sampling variance (SV) estimate is clustered at the federal state level. The final SV is given by the average of sampling variances obtained with the five PVs. In addition, standard errors are inflated by the imputation variance (IV) because test scores measure latent cognitive skills with error. The IV is estimated as the average squared deviation between the estimates obtained with each Plausible Value and the final estimate (using the average of PVs), with the appropriate degree of freedom adjustment ($IV = \frac{1}{4}\sum\left(\hat{\theta}_i - \hat{\theta}\right)^2$ where $\hat{\theta}_i$ is the estimate for each of the five PVs and $\hat{\theta}$ is the final estimate). Then, as shown by (OECD, 2009b), the final error variance TV can be obtained by combining the sampling and imputation variance as follows:

$$TV = SV + (1 + \frac{1}{K}) * IV = SV + 1.2 * IV \tag{A.2}$$

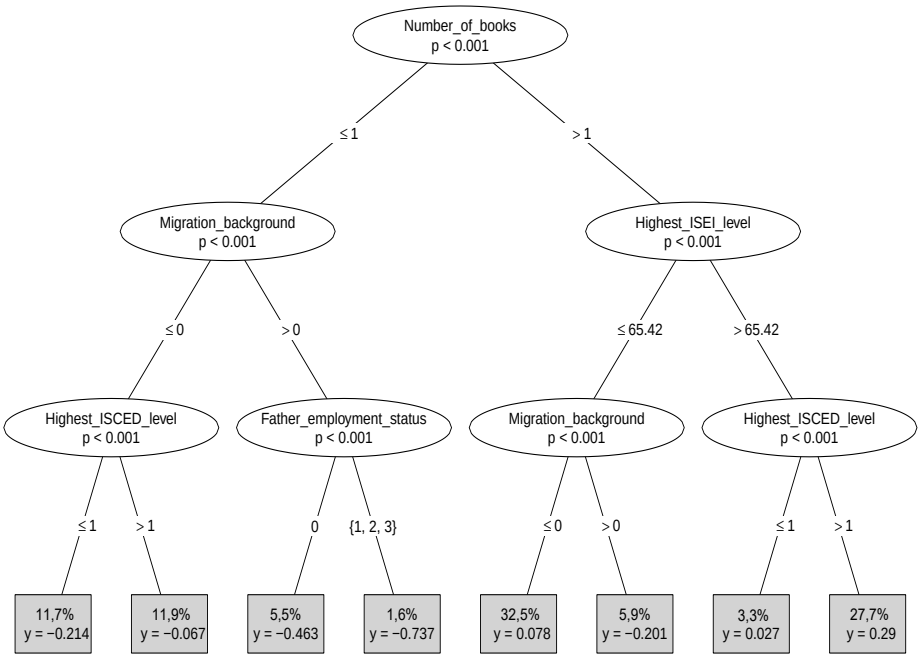
where $K = 5$ is the number of Plausible Values for each student. The final standard errors are given by the squared roots of the final error variances. To estimate SV, one can apply Fay's variant of the *Balanced Repeated Replication (BRR) method*, which directly considers the two-stage stratified sampling design of the PISA test. Therefore, each regression is iterated over the 80 sets of *replication weights* provided in the PISA dataset. Then, the SV estimate is given by the average squared deviation between the replicated estimates and the estimate obtained with final weights, with a degree of freedom correction depending on the Fay coefficient (a parameter governing the variability between different sets of replication weights, set at 0.5 in PISA).

Standard errors in all *first-step* and *second-step* regressions are based on this method. For computational convenience and similar to Philippis and Rossi (2019), I use the "unbiased shortcut" procedure described in OECD (2009b). It relies on only one set of Plausible Values (PVs) for estimating the sampling variance (whereas the imputation is estimated using all five sets). Andrietti and Su (2019) rely on clustering standard errors on the state level and argues that a wild t-bootstrap procedure produces similar results.[45] Huebener et al. (2017) also focus on clustering methods. However, given the sampling strategy used to generate PISA scores, estimating standard errors considering both replication weights and PVs is more reliable.

A.5.2 Detecting Important Circumstances Variables with Machine Learning

Machine Learning (ML) can be helpful to identify a model specification based on its advantages of being a data-driven, transparent, theory-agnostic, non-parametric approach. I apply the ML method of conditional inference regression trees to test the importance of my chosen *circumstances* in Section 3.3. This exercise confirms that the selected *circumstances* are indeed relevant for explaining differences in cognitive skills as measured by PISA tests. My ML algorithm follows the approach of Brunori et al. (2018), and I refer to their paper for more details on the technicalities. In summary, the tree algorithm

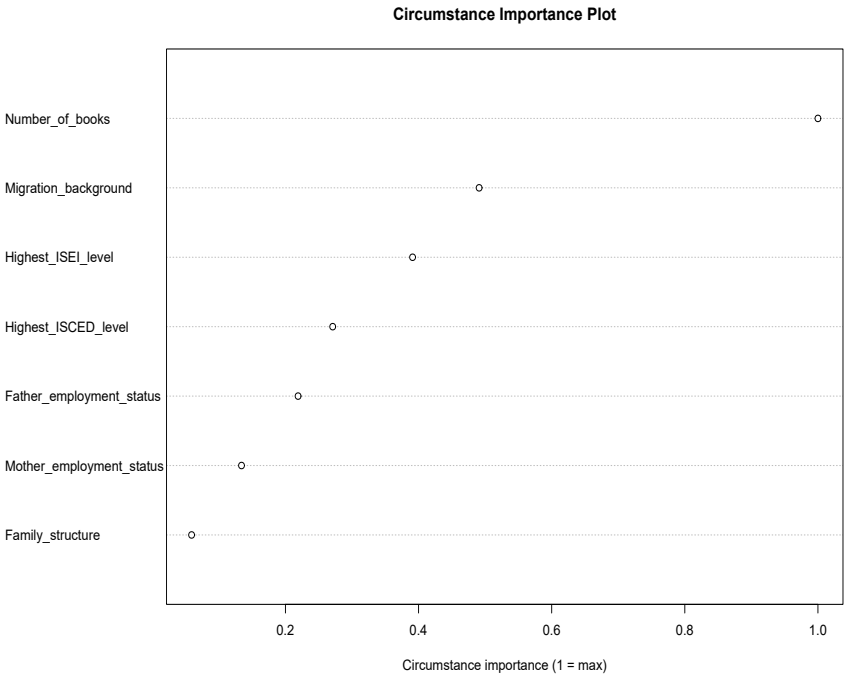
[45] Please refer to footnote 12 in (Andrietti & Su, 2019).



Note: This is an **Opportunity Tree** for students in *Gymnasium* considering PISA-I waves 2003, 2006, 2009, and 2012: Variables inside the white circles depict the *circumstances* on which the algorithm has chosen to split. The splitting criterion value is shown on the tree-branches and is based on the p-value (at the one percent level) of the difference in test scores between the *circumstances* groups. Terminal nodes are depicted by grey boxes: The first number shows the respective group's percentage share of the total weighted sample size; the second number shows the group's predicted standardized mathematics score. The tree algorithm splits the dataset into groups if the null hypothesis of EEOp is rejected. For illustrative reasons, the tree depicted considers a maximum depth of 3. To more clearly identify all drivers of IEOp, gender, as main driver of differences in test scores, was left out.

splits the dataset into groups if the null hypothesis of Equality of Educational Opportunity (EEOp) is rejected. This is best illustrated in the figure at the bottom of the previous page: It depicts an opportunity tree which is calculated for the standardized PISA mathematics score as outcome variable and is based on the PISA-I dataset. The tree shows that, for instance, students living in households that own less than 100 books achieve significantly worse results in mathematics compared to those from households with more books. Generally, the tree reveals that there are groups of certain *circumstances* along the lines of socioeconomic status, parental education, and migration background.

To check if the results obtained by the tree are stable, I further conduct a conditional inference regression forest machine learning procedure. The method is similar to the regression tree; however, forests calculate many trees and then average the obtained effect over the identified subgroups. Therefore, only a variable importance plot can be depicted (see the figure at the bottom of the page). The importance is calculated by the permutation principle of the mean decrease in accuracy whereby the variable importance is adjusted to depict relative terms. Hereby, the *circumstance* variable with greatest importance equals 1. As in the tree algorithm, the number of books seems to be an important factor influencing PISA test scores of students in Germany. Moreover, migration background, the highest ISEI and ISCED level of the household in which a student grows up, turn out to be a relevant *circumstance*. Thus, the machine learning algorithm confirms to include the depicted variables as controls which is in line with the arguments provided in Section 3.3.



Note: This figure depicts the **Circumstance Importance Plot** for students considering PISA-I waves 2003, 2006, 2009 and 2012: The plot indicates the importance of the listed *circumstances* with respect to the standardized PISA mathematics scores of students in *Gymnasium*. The importance is calculated by the permutation principle of the mean decrease in accuracy whereby the variable importance is adjusted to depict relative terms. Note that qualitatively similar results can be obtained using PISA test scores in reading or science as outcome variable instead of mathematics scores only. Due to space constraints, the respective graphs for reading and science are only available upon request from the author.

Furthermore, one can use the importance of different *circumstances* (as revealed by the plot graph) to derive refined interaction terms in order to detect heterogeneity in the causal reform effect on IEOp. In that regard, for instance, the regression forest exercise indicates that both the number of books in a household and the highest parental jobs' ISEI level should be used as important *circumstances* variables to control for social status. Results in Section 5.4 show that these *circumstances* explain heterogeneity in the effect of higher learning intensity on test scores which is in line with the estimated increase in IEOp.

A.5.3 List of *Circumstances* Variables

1. Individual Characteristics (IC):
   - (I) *gender* [Base: male] and *age* (in years)
   - (II) *migration background* [Base: German] and *language spoken at home* [Base: German]
2. Parental Characteristics (PC)
   - (III) *education*: highest *ISCED-index* level in 3 categories [Base: ISCED-level (3-4)]
3. Socio-Economic Status (SES)
   - (IV) *number of books in household* [Base: 101-500]
   - (V) *highest ISEI-index level* [scale: 0-90]
4. Family Characteristics (FC)
   - (VI) *single parent household* [Base: none]
   - (VII) *mother/father employment status* [Base: FT]

A.5.4 Overview of Definitions and T/C-Groups

1. Concerning the time periods possible, one can define the following models:
   - *Baseline Model: medium-term perspective* (Base): covers time period (**2003-2012**)
   - *Robustness Model: short-term perspective* (Robust): covers time period (**2003-2009**)
2. Concerning *Treatment* and *Control Groups*, the following groups can be formed (Table 2)
   - *Treatment Group (**T**)*: *Baden-Württemberg* (**BW**), *Bavaria (**BV**)*, *Lower Saxony (**LS**)*, *Bremen (**BR**)*, *Hamburg (**HB**)*
   - *Treatment Group (**T1**)*: *Baden-Württemberg* (**BW**), *Bavaria (**BV**)*, *Lower Saxony (**LS**)*
   - *Treatment Group (**T2**)*: ***BW***, ***BV***, ***LS***, ***BR***, ***HB***, *Berlin (**BE**)*, *Brandenburg (**BB**)*
   - *Control Group (**C**)*: *Rhineland-Palatinate (**RP**)*, *Schleswig-Holstein (**SH**)*
   - for short-term *Model Robust Control Group (**C1**)*: ***RP***, ***SH***, *North Rhine-Westphalia (**NRW**)*
   - *hypothetical Control Group (**Ch**)*: *Saxony (**SN**)*, *Thuringia (**TH**)*
   - *Never-Takers Control Group (**C-NT**)*: ***RP***, ***SH***, ***SN***, ***TH***
3. *Neither Treatment nor Control Group:*
   - *Saarland* (**SL**), *Saxony-Anhalt* (**ST**), *Mecklenburg-West Pomerania* (**MWP**), *Hesse* (**H**)

A.5.5 Further Aspects on the Internal Validity of Empirical Strategy

There were no specific changes in the political parties forming the government of federal states that form my main treatment and control group settings in both the *Model Base* (2003-2012) or *Model Robust* (2003-2009). Moreover, by conducting a Difference-in-Differences estimation (DiD) and controlling for federal states, general differences in the political parties in charge of implementing the reform are taken into account. The fact that there have not been systematic changes in governments across treatment and control groups around the respective reform time is supportive evidence that, for the period considered, it is plausible to assume a comparability in the stability of each federal state's educational policies.

- Treatment Groups (T/T1)
  - **BW**: Conservatives (CDU) led the government for decades until 2011, followed by (2011-2016) a coalition government of the Green Party/Social-Democrats (SPD): The reform was implemented by the CDU, and it is plausible to assume that, due to the time lag for new government policy to take effect, educational policy up until year 2012/2013 was made by the same party.
  - **BV**: Conservatives (CSU) led the government over the whole analysis period (2003-2012), thus, it is plausible to assume that school policy was conducted by the same party.
  - **LS**: Conservatives (CDU) led the government over the whole analysis period (2003-2012); afterwards/beforehand the government was led by the SPD. It is plausible to assume that for the whole analysis period, school policy was made by the same party.
  - **BR**: Social-Democrats (SPD) led the government over the analysis period (2003-2012), and thus, it is plausible to assume that school policy was made by the same party.
  - **HB**: Social-Democrats (SPD) led the government for decades (until 2001, since 2011). In between Conservatives governed and thus it is plausible to assume that for the analysis period (2003-2012), school policy was mainly conducted by the same party.
- Control Groups (C/C1)
  - **RP**: Social-Democrats (SPD) led the government over the analysis period (2003-2012), thus, it is plausible to assume that school policy was conducted by the same party.
  - **SH**: Social-Democrats (SPD) led the government for decades (1988-2005, 2012-2017). In between (2005-2012), the government was led by Conservatives, from 2010-2012 in a grand coalition with the SPD. School policy remained similar during the analysis period.
  - **NRW**: Social-Democrats (SPD) led the government for decades (until 2005, 2010-2017). They had already enacted the reform, when for five years the government changed to the Conservatives (CDU) who continued the implementation of the reform. School policy remained similar, in particular, when taking NRW as control for the period 2003-2009.

Thus, by focusing on the analysis period (2003-2012) that covers only the first affected cohorts, the main DiD assumptions appear to hold. However, as some federal states decided to reverse the reform in recent years, a similar evaluation may be less plausible for the time period after 2012. The reform has become a debated topic in most federal states since the early 2010s (cf. last column in Table A.1). But for the very first affected cohorts, there are no systematic changes in governments when comparing treatment and control group states over the time period (2003-2012).

A.5.6 On Ability in the Context of Measuring IEOp and Within the DiD Framework

Even though one may have concerns about differences in ability (or talents) when it comes to measure IEOp, one should consider the following. First, the IEOp measurement framework takes any time-invariant features of cognitive skills into account as part of the unobserved component of *circumstances*. Second, recent literature in the field of neuroscience suggests that in the spirit of the Human Capital Theory, cognitive skills appear to be malleable through epigenetic processes, in particular during early childhood. This may explain why, for instance, Boca et al. (2017) find that attending childcare institutions can significantly improve children's cognitive skills, in particular for those from disadvantaged SES. Thus, the IEOp measurement framework fully takes the role of ability into account, both as unobserved *circumstance* and *effort*. Consequently, it is a lower bound measure. Moreover, skills are defined as mixture of *circumstances* and *efforts* .

Concerning the Differences-in-Differences estimation approach (DiD), the only assumption that I make is that, in general, the distribution in cognitive abilities of students between 2003 and 2012 did not systematically change across German federal states. Given the fact that moving behavior between federal states is unlikely to have occurred (Section 4), this means we assume that cognitive skills did not suddenly change across states during the analyzed time period for any other reason than the reform treatment. Moreover, even if general systematic differences in ability across federal states existed, the DiD framework would control for any general level differences in ability.

Therefore, given the short time period and the controls enacted via the DiD framework, it is hard to find plausible reasons why there should have been any significant changes in cognitive abilities that differ among federal states and could bias results. In any case, these thoughts should be of less concern in this quasi-experimental setting than in the settings of other research papers that measure IEOp across countries. Moreover, as the reform only affects students from age 10 onward, and treatment merely involves more intense instruction but not different contents, I claim that these concerns - which can neither be addressed by empirical methods nor available data (e.g. there are no representative data on IQs in Germany) - are of second order importance and comparable to those in other studies estimating returns to schooling.

Online References

Andrietti, V., & Su, X. (2019a, jan). Education Curriculum and Student Achievement: Theory and Evidence. *Education Economics*, *27*(1), 4–19. doi: 10.1080/09645292.2018.1527894

Andrietti, V., & Su, X. (2019b, sep). The Impact of Schooling Intensity on Student Learning: Evidence from a Quasi-Experiment. *Education Finance and Policy*, *14*(4), 679–701. doi: 10.1162/edfp_a_00263

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., ... Weiß, M. (2002). *PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

Baumert, J., & Prenzel, M. (2008). *Vertiefende Analysen zu PISA 2006* (Vol. 10). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/978-3-531-91815-0

Black, S. E., & Rothstein, J. (2019). Policy Brief 3: An Expanded View of Government's Role in Providing Social Insurance and Investing in Children. *econfip*(January).

Boca, D. D., Piazzalunga, D., & Pronzato, C. (2017). Early Childcare, Child Cognitive Outcomes and Inequalities in the UK. *HCEO Working Paper Series*(2017-005).

Brunori, P., Hufe, P., & Mahler, D. G. (2018). *The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees*. The World Bank. doi: 10.1596/1813-9450-8349

Büttner, B., & Thomsen, S. L. (2015). Are We Spending Too Many Years in School? Causal Evidence of the Impact of Shortening Secondary School Duration. *German Economic Review*, *16*(1), 65–86. doi: 10.1111/geer.12038

Camarero Garcia, S. (2012). *Does Shortening Secondary School Duration Affect Student Achievement and Educational Equality? - Evidence from a Natural Experiment in Germany: The G-8 Reform* [Bachelor Thesis, University of St. Gallen].

Checchi, D., & van de Werfhorst, H. G. (2018). Policies, Skills and Earnings: How Educational Inequality Affects Earnings Inequality. *Socio-Economic Review*, *16*(1), 137–160. doi: 10.1093/ser/mwx008

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2020, aug). Income Segregation and Intergenerational Mobility Across Colleges in the United States*. *The Quarterly Journal of Economics*, *135*(3), 1567–1633. doi: 10.1093/qje/qjaa005

Davoli, M., & Entorf, H. (2018). The PISA Shock, Socioeconomic Inequality, and School Reforms in Germany. *IZA Policy Paper*(140).

Ferreira, F. H. G., & Gignoux, J. (2013). The Measurement of Educational Inequality: Achievement and Opportunity. *The World Bank Economic Review*, *28*(2), 210–246. doi: 10.1093/wber/lht004

Homuth, C. (2012). Der Einfluss des achtjährigen Gymnasiums auf den Kompetenzerwerb. , 1–22. doi: 10.1007/BF01778681

Huebener, M., Kuger, S., & Marcus, J. (2017). Increased Instruction Hours and the Widening Gap in Student Performance. *Labour Economics*, *47*(1561), 15–34. doi: 10.1016/j.labeco.2017.04.007

Huebener, M., & Marcus, J. (2017). Compressing Instruction Time into Fewer Years of Schooling and the Impact on Student Performance. *Economics of Education Review*, *58*, 1–14. doi: 10.1016/j.econedurev.2017.03.003

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., ... Stanat, P. H. (2011). *PISA 2009 - Bilanz nach einem Jahrzehnt*. Münster: Waxmann Verlag. doi: 10.1787/9789264095359-de

Lavy, V. (2015). Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries. *The Economic Journal*, *125*(588), F397–F424. doi: 10.1111/ecoj.12233

Marcus, J., & Zambre, V. (2019). The Effect of Increasing Education Efficiency on University Enrollment. *Journal of Human Resources*, *54*(2), 468–502. doi: 10.3368/jhr.54.2.1016.8324R

Meyer, T., & Thomsen, S. L. (2016). How Important is Secondary School Duration for Postsecondary Education Decisions? Evidence from a Natural Experiment. *Journal of Human Capital*, *10*(1), 67–108. doi: 10.1086/684017

Meyer, T., Thomsen, S. L., & Schneider, H. (2018, mar). New Evidence on the Effects of the Shortened School Duration in Germany: An evaluation of Post-School Education Decisions. *German Economic Review*, *9507*(9507). doi: 10.1111/geer.12162

OECD. (2001). *Knowledge and Skills for Life - First Results from PISA 2000*. Paris: OECD Publishing. doi: 10.1787/9789264195905-en

OECD. (2004). *The PISA 2003 Assessment Framework - Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Paris: OECD Publishing. doi: 10.1787/9789264101739-en

OECD. (2005a). *Education at a Glance 2005 - Home*. Retrieved from: `http://www.oecd.org/2005.htm`. (Last access: September 16, 2020)

OECD. (2005b). *PISA 2003 Technical Report*. Paris: OECD Publishing. doi: 10.1787/9789264010543-en

OECD. (2009a). *PISA 2006 Technical Report*. Paris: OECD Publishing. doi: 10.1787/9789264048096-en

OECD. (2009b). *PISA Data Analysis Manual: SPSS, Second Edition*. OECD Publishing. doi: 10.1787/9789264056275-en

OECD. (2010). *PISA 2009 Assessment Framework - Key Competencies in Reading, Mathematics and Science*. Paris: OECD Publishing. doi: 10.1787/9789264062658-en

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing. doi: 10.1787/9789264167872-en

OECD. (2013a). *PISA 2012 Assessment and Analytical Framework - Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing. doi: 10.1787/9789264190511-en

OECD. (2013b). *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed* (Vol. II; PISA, Ed.). Paris: OECD Publishing. doi: 10.1787/9789264201132-en

Philippis, M. D., & Rossi, F. (2019). Parents, Schools and Human Capital Differences across Countries. *Journal of the European Economic Association (forthcoming)*, 1–43.

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann Verlag.

Quis, J. S., & Reif, S. (2017). Health Effects of Instruction Intensity - Evidence from a Natural Experiment in German High-Schools. *FAU Discussion Papers in Economics*(12-2017), 1–30.

Thiel, H., Thomsen, S. L., & Büttner, B. (2014). Variation of Learning Intensity in Late Adolescence and the Effect on Personality Traits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(4), 861–892. doi: 10.1111/rssa.12079

Wössmann, L. (2010). Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries. *Jahrbücher fur Nationalökonomie und Statistik*, *230*(2).

Wössmann, L., Lergetporer, P., Kugler, F., Oestreich, L., & Werner, K. (2015). Deutsche sind zu Grundlegenden Bildungsreformen bereit – Ergebnisse des ifo Bildungsbarometers 2015. *ifo Schnelldienst*, *68*(17), 03–24.

Download ZEW Discussion Papers from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de