Camarero Garcia, Sebastian

Working Paper

Inequality of educational opportunities and the role of learning intensity: Evidence from a quasi-experiment in Germany

ZEW Discussion Papers, No. 18-021

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Camarero Garcia, Sebastian (2018) : Inequality of educational opportunities and the role of learning intensity: Evidence from a quasi-experiment in Germany, ZEW Discussion Papers, No. 18-021, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, https://nbn-resolving.de/urn:nbn:de:bsz:180-madoc-456979

This Version is available at:
https://hdl.handle.net/10419/178498

Discussion Paper No. 18-021

Inequality of Educational Opportunities and the Role of Learning Intensity: Evidence from a Quasi-Experiment in Germany

Sebastian Camarero Garcia

ZEW

Zentrum für Europäische Wirtschaftsforschung GmbH

Centre for European Economic Research

Discussion Paper No. 18-021

Inequality of Educational Opportunities and the Role of Learning Intensity: Evidence from a Quasi-Experiment in Germany

Sebastian Camarero Garcia

Inequality of Educational Opportunities and the Role of Learning Intensity: Evidence from a Quasi-Experiment in Germany*

Sebastian Camarero Garcia†

ZEW Mannheim‡, University of Mannheim§ and CEP at LSE¶

May 2018

Abstract

Over the 2000s, many federal states in Germany shortened the duration of secondary school by one year while keeping the curriculum unchanged. Exploiting quasi-experimental variation due to the staggered introduction of this reform allows me to identify the causal effect of increased learning intensity on Inequality of Educational Opportunity (IEOp), the share in educational outcome variance explained by predetermined *circumstances* beyond a student's control. The reform-induced increase in learning intensity had no short-term effect on IEOp. In the medium term, however, IEOp increased as differences in parental resources gained importance through support opportunities like private tuition adapting to the intensified educational process. The effect is stronger for mathematics/science than for reading, implying the existence of subject-dependent curricular flexibilities. My findings point to the importance of accounting for distributional consequences when evaluating reforms aimed at increasing the efficiency of educational systems and to the role of learning intensity for explaining changes in educational opportunities influencing social mobility.

JEL-Classification: D04, D63, H75, I21, I24, I28, J18, J24, J62, O52

Keywords: *(In)Equality of Opportunity*, *Educational/Learning Intensity*, *shortening school duration*, *G-8 education reform*, *Education & Social Mobility*, *Germany*

*I would like to thank my supervisor Andreas Peichl. Moreover, I am thankful to Felix Chopra, 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 and Martin Ungerer 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 conference at CUNY, New York City and ECINEQ Winter School in Canazei 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 would also like to acknowledge financial support by the Cusanuswerk. My thanks also go to Elvira Babaeva and Marc Stadtherr for their excellent research assistance.

†E-Mail: sebastian.camarero.garcia@gess.uni-mannheim.de.

‡Centre for European Economic Research (ZEW), 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 income and wealth inequality 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 may 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) may be crucial for explaining 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, thus, it is of key policy interest to analyze educational systems in terms of equality, in particular to detect drivers of *inequality of opportunity* (as Chetty, Friedman, Saez, Turner, and Yagan (2017) for US colleges).

Yet in times of public spending constraints, accelerating growth of scientific knowledge and economic competition within OECD countries, educational policies have still shifted their attention onto how to make a country's educational system more efficient (Machin, 2014). In fact, recent reforms have started to focus on compressing educational processes, that is, increasing learning intensity.[2]

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 of compressing educational processes for social mobility. As 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 will adopt the concept as illustrated by Roemer and Trannoy (2015) stating that society has achieved *equality of opportunity* if what individuals achieve with respect to a desirable objective is determined by their *efforts* (e.g. how hard they study), instead of by *circumstances* that are beyond an individual's control (e.g. gender). IEOp[3] is hence 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. It is a relative measure of educational mobility.

[1] For instance, Chetty, Grusky, et al. (2017) provide evidence for falling absolute income mobility. OECD (Organization of Economic Cooperation and Development) data from 2012 confirm low absolute educational mobility, in particular Germany reaches only below average *upward* and *downward* mobility rates (Figure A.1 in Appendix A.5.).

[2] With respect to schooling, learning intensity can be defined as the ratio of curricular content that is covered in a given amount of instructional time. This can also be denoted as schooling intensity and would correspond to the intensive margin if the curricular content fixed for a given school track was regarded to be the extensive margin. But research on the role of this factor has been limited as it goes beyond issues of participation in education or duration.

[3] 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. Thus, if opportunities depend less on factors beyond an individual's control but more on their *efforts*, EOp will increase and IOp will decrease. In line with Brunori, Peragine, and Serlenga (2012), instead of IOp *in education* I use the expression IEOp and instead of EOp *in education*, Equality of Educational Opportunity (EEOp). In the following, 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. The 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 in the same year in which they received their university access diploma. However, 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. For that reason, the staggered introduction of the reform across federal states generates quasi-experimental variation that allows the application of a Difference-in-Differences estimation approach (DiD) to derive the causal effect of the increase in learning intensity on IEOp by comparing the respective treatment and control group over time.

For the purpose of measuring IEOp, I use Program for International Student Assessment (PISA) data which provides a representative sample of students in the ninth grade with standardized test scores in reading, mathematics and science comparable across time and federal states, as well as a rich set of family background variables that allow me to define relevant *circumstances*. Then, 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 reaching 20-30% of the variance in cognitive test scores that can be only attributed to *circumstances* corresponds to the levels of common estimates for inequality of opportunity in income. Second, the reform-induced increase in learning intensity did not affect IEOp in the short term. However, in the medium term, after an adjustment period of about three years, IEOp significantly increased by up to 10% in terms of explained test score variance for the fourth affected cohort onwards. 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 in favor of the existence of subject-dependent curricular flexibilities. General skills in mathematics and science tend to be more inflexible and thus more responsive to changes in curricular intensity than reading competency, which is in general less dependent on schooling as it is more often trained through its usage in everyday life. Finally, the results can be rationalized by differential compensation possibilities for higher learning intensity depending on parental resources in terms of both the capacity to pay for additional tuition and to invest time in supporting students with school work. This shows that there are important adverse distributional concerns with respect to providing equal opportunities that must be taken into account when designing reforms altering the intensity of educational processes.

This paper is among the first to provide an analysis of IEOp in a quasi-experimental setting going beyond its pure measurement. As Ramos and Van de gaer (2016) point out, the understanding of how institutions influence IEOp is still limited. My aim is to contribute to this issue by 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 may be realized if intense curricula proved to be an instrument to overcome the trade-off between educational spending and output.

This paper offers several contributions to the existing literature. First, I contribute to the still limited strand of research on measuring IOp with respect to educational outcomes by adding empirical evidence on how IEOp 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, Cappelen, Lind, Sørensen, and Tungodden (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 follow mostly a cross-country comparison approach using PISA data to achieve comparability of educational achievement measures over time and across countries (e.g. Raitano and Vona (2016); Oppedisano and Turati (2015); Ferreira and Gignoux (2013)). Instead, my study estimates IEOp for Germany exploiting quasi-experimental within-country variation. 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. Bratti, Checchi, and de Blasio (2008) and Brunori et al. (2012) on Italy). They find that both expanding higher education through opening more sites and reducing the length to get a first-level degree to have a positive effect on equality of educational opportunity. However, only a few studies investigate the impact of school reforms on IEOp (e.g. Edmark, Frölich, and Wondratschek (2014) for Sweden). Specifically for Germany, Riphahn and Trübswetter (2013) confirm the role of intergenerational persistence in educational achievement and that the school system did not improve in response to reunification regarding IEOp.[4] In this paper, I add evidence on how IEOp changed over time in Germany, but focus on causal estimates of educational intensity for the academic track in the secondary education 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)). 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.[5]

[4] Only 19% of 25-34 year-old Germans achieve higher degrees than their parents (OECD Education at Glance 2014).

[5] There have been a few studies exploiting quasi-random assignments of instructional time (e.g. due to timing of

In this context, reforms that shortened schooling while keeping curricular content unchanged, allow for the evaluation of the impact of increasing learning intensity. For instance, analyzing a similar school reform in parts of Canada, Krashinsky (2014) finds only low long-term effects on wages, which suggests that increased learning intensity might not affect earnings permanently.[6] 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 the 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 it, only a few studies have evaluated the G-8 reform and its effects on educational outcomes (e.g. Büttner and Thomsen (2015); Andrietti (2016); Huebener, Kuger, and Marcus (2017)). Those studies tend to find non-significant positive effects of the reform on a student's average cognitive skills and educational outcomes, such as final marks for the university access diploma.[7] Instead, my analysis shifts focus in the evaluation of the G-8 reform onto distributional concerns. This is relevant in the debate surrounding reforms to the secondary school system, because it provides policy suggestions for how to design curricula in terms of learning intensity taking into account both the effects on cognitive skill formation and on IEOp. For instance, implementing a whole-day school system may limit the role of parents for students to deal with compressed schooling.

Thirdly, my paper relates to the emerging literature aimed at explaining drivers of inequality in educational outcomes as one key determinant of recent trends in decreased social mobility (e.g. Philippis and Rossi (2017); Boneva and Rauh (2017a); Chetty, Friedman, et al. (2017); Rothstein (2018)). I contribute to this strand of research by providing evidence that the so far neglected factor of learning intensity may be a relevant policy channel for both the effectiveness of (non-)cognitive skill formation and the importance of *circumstances* for educational outcomes. Whereas my analysis mainly focuses on exploiting a school reform to derive estimates on how intensified instruction affects IEOp in Germany, the interpretation of these empirical 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 and which supports the integration of 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 the implications. Section 6 concludes.

school year, absence of teachers) that usually find similar positive effects of additional time on test scores. Marcotte (2007) is an example of a study exploiting quasi-experimental variation in time available to prepare for state-wide tests (due to differences in snow-related school closure) and finds positive effects of more school days on performance.

[6] Whether this is true due to schooling working primarily as a signal or whether increased intensity may compensate human capital accumulation in response to less schooling, is unclear. The fact that students could choose to complete high school in four or five years, however, raises doubts over whether the effect captures schooling intensity.

[7] For related literature which evaluates other outcomes of the G-8 reform, please refer to Appendix A.1.2.

2 Institutional Setting: the "G-8 reform"

Explaining the institutional background and implementation of the G-8 reform illustrates how it can be exploited as a quasi-experiment to analyze the effect 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*). That being said, most features are comparable across states. School starts usually at the age of six, when students enter primary school for a period of four years. Afterwards, students enter a tripartite secondary school system, where the choice of track is determined by their previous academic performance.[8]

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. The academic track, *Gymnasium*, which this paper focuses on, leads to a diploma (*Abitur*) granting access to university. Traditionally, the academic track used to last for 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 adjusted to 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).[9]

Then, in the early 2000s, the nine years were perceived as a competitive disadvantage for the economy, because they contributed to the relatively 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).[10]

[8] Primary schools issue recommendations for each student regarding which secondary school track the student should enter (Dustmann, Puhani, & Schönberg, 2017). Based on a student's performance in primary school, recommendations were binding in federal states for the time period 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 .

[9] In addition to the three different school tracks, federal states have recently started to provide a comprehensive school (*Integrierte Gesamtschule*). In comprehensive schools, 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), because during this time the vast majority of students achieving *Abitur* still attended *Gymnasium*. See Figure A.2 in Appendix A.5 for further details on the German education system.

[10] For further arguments discussed during the reform debate, please refer to Appendix A.1.1 in Appendix A.1.

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 2008/2009, implying that the resulting *double cohorts* graduated between 2006/2007 and 2015/2016.[11]

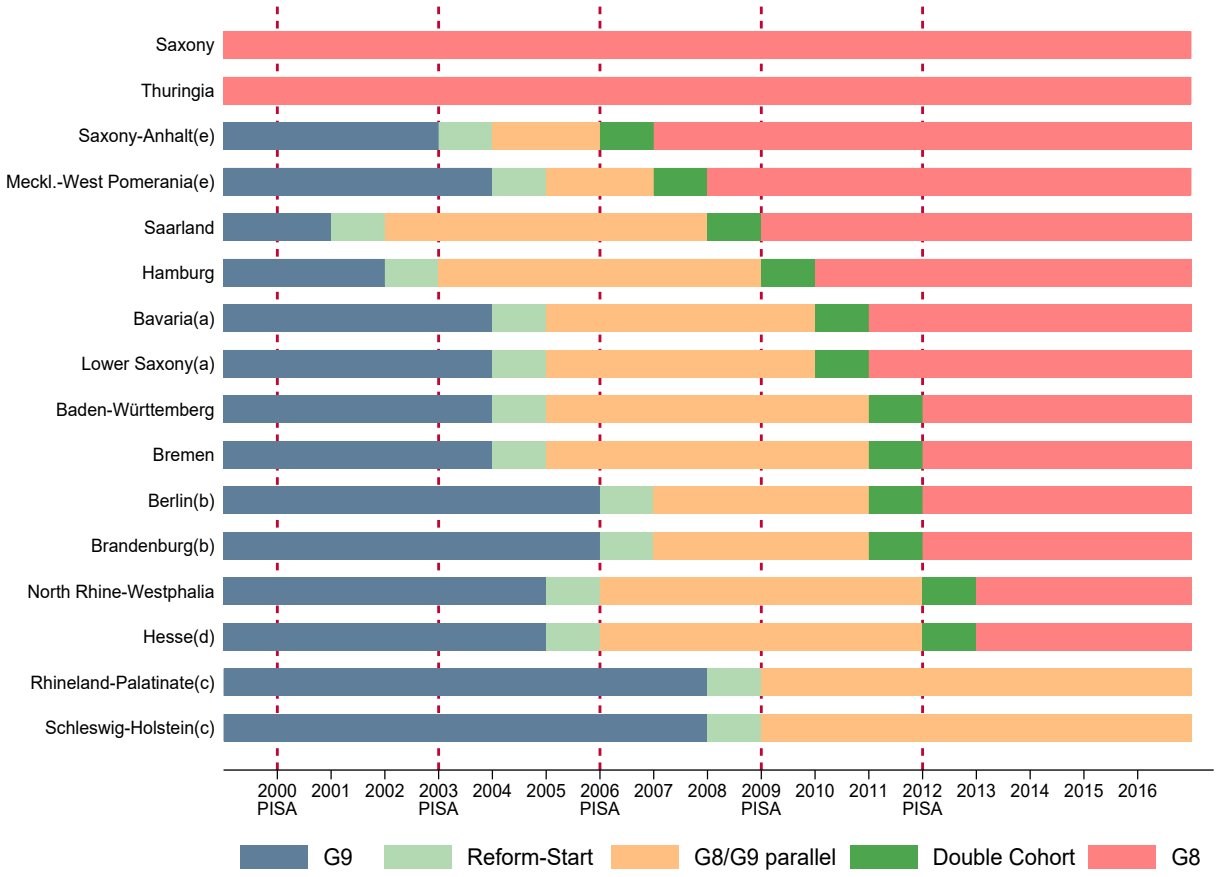
Second, although the academic track was reduced by one school year, the curricular content remained at the original level. In fact, education ministers decided that standards for the university access diploma (*Abitur*) were not to be lowered in response to the reform. Therefore, the minimum number of 265 instruction hours per week and school year across all grades was maintained and students still had to pass the same number of lessons in total before they could graduate from the *Gymnasium* (Education Ministers, 2016). This should ensure comparable nationwide standards 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, because the first G-8 and the last G-9 model cohort had to complete those grades together. Only marks during the final two years and marks in the *Abitur* exam count towards the university access GPA. Thus, school authorities chose to focus the compression on the first years instead, squeezing the material originally taught in the seven years between grades 5 and 11 into the six years between grades 5 and 10. In other words, 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. To keep the required total minimum weekly lessons unchanged for the new G-8 model, instructional time increased by about two hours a week per grade during grades 5-9 for G-8 model students compared to previous cohorts in the G-9 model.[12]

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 teach already 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, on which this study focuses on, in any of the federal states.

[11] For related literature which evaluates other outcomes of the G-8 reform, please refer to Appendix A.1.2.

[12] Andrietti (2016) provides a calculation based on the regulations set by the Ministry of Education on weekly hours: "*By the end of grade 9, G8 students have covered the curriculum corresponding to 6,460 (265/8 per week over 39 weeks for five grades) of the 10,335 instructional hours required for graduation. This means that they have accumulated on average 720 more instructional hours [...] than G9 students at the end of grade 9 (265/9 per week over 39 weeks for five grades, i.e. 5,740 hours) [...].*" 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 G-8 reform: on average hours per grade increased by about 2 hours, i.e. about 8-10% per grade and school year.

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 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 onwards.

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

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

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 9^{th} grade onwards.

Source: Based on facts as shown in Table 2 and the regulations explained in Table A.4. This figure corresponds to the geographical maps illustrating the implementation of the reform across time and space in Figure A.3 in Appendix A.5.

The G-8 reform exogenously led to a considerable increase in learning intensity over the first few years of the *Gymnasium*, that is, the amount of material covered per week increased for each grade.[13]

[13] As curricular content in the G-8 model began to change in the years after 2012 (cf. Table A.4 in Appendix A.4), this assumption would not necessarily hold for later G-8 cohorts. But by using data including ninth graders tested in 2012 or before, I focus on the very first cohorts affected by the reform, so these later changes do not affect the analysis.

3 Data and Measuring Inequality of Educational Opportunity (IEOp)

In this section, I first focus on which specific PISA data[14] are used for my analysis. Then, I will 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. Finally, I provide some basic descriptives 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 taking the same test on the same day 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, on the one hand, about 25 students of age 15 are randomly taken to be tested (*age-based* sample); on the other hand, within 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.[15] 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, a comparison among ninth graders is more appropriate for the strategy pursued in this paper because the G-8 reform affected students based on their grade in a certain school year.

Moreover, national PISA extensions were conducted for the years 2000, 2003 and 2006. Each of them consists of about 50,000 students. By oversampling less populated federal states, these extensions allow for more robust comparisons of educational performance between German federal states.[16] However, PISA-E was discontinued in 2009 and replaced by the *IQB federal state comparison test*.[17] Since then each extension 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.

Nevertheless, Andrietti (2016) 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, in the case of Andrietti (2016) only for the year 2009 and in the case of Huebener et al. (2017) additionally for 2012. Only *grade-based* PISA-I samples have all domains consistently available for each test year.

[14] Some background information on the OECD's PISA data, its advantages and disadvantages to measure educational outcomes as well as on the representativeness of these data across states and over time is provided in Appendix A.2.1.

[15] See Table A.1 in Appendix A.4 for an overview of available *grade-based* PISA-I datasets.

[16] For this purpose, one day after the tests taken for the PISA-I samples, in each federal state additional students were randomly selected to undergo the same testing procedures answering an additional national questionnaire. Combined with the original PISA-I samples (Table A.1), enlarged *grade-based* PISA-E samples emerged (Table A.2). However, the PISA-E 2006 data is based only on age and not on the grade (cf. Table A.2). Thus, one would need to focus on the subset of 15-year-olds who are in ninth grade. But this results in a smaller sample size and constitutes a departure from the *grade-based* sampling on which data for the other waves rely.

[17] From 2009 onwards, this *comparison test* aims to assess national educational standards determined by the Standing Conference of Ministers of Education (SC) of all federal states instead of by the OECD.

Therefore, to have more consistent comparability across the studies used, this paper avoids mixing PISA-E and PISA-I datasets. For these reasons, the main results will be based on *grade-based* PISA-I samples for all available waves (2003, 2006, 2009, 2012). The PISA-E data will only be used when the time period is extended as part of robustness checks, and their use will be restricted to the 2000 PISA-E wave (see Appendix A.2.2 for data sources used).[18]

As this paper focuses on the academic track (*Gymnasium*), only schools of this type are included in the sample. They make up about one third of the *grade-based* PISA-I sample which corresponds to the real share of secondary school students in *Gymnasium*. Finally, the analysis is restricted to variables derived from the questionnaire for students and their parents (the *student*-dataset).[19]

To sum up, this paper relies on the *grade-based* PISA-I samples to construct a representative repeated cross-section of German students in grade nine of the *Gymnasium* that allows me to analyze IEOp in response to the G-8 reform by using variables based on test scores and background characteristics.

Descriptive Statistics Regarding the main outcome variables, PISA test scores in the domains of reading, mathematics and science, by focusing on students in the academic track of secondary school, mean PISA test scores are above the German average. As expected, a typical ninth grader in a *Gymnasium* consistently achieves results that are about 60 points higher than for the average German ninth grader, which corresponds to about an entire *proficiency level*, that is, the value-added of two school years. Regarding the three testing areas, students perform worst in reading literacy. Moreover, they appear to stagnate or even slightly deteriorate in their reading skills between 2000 and 2012. This observation is in line with reports on German PISA test results for the years 2000-2009 illustrating that students perform better in mathematics compared to reading skills (cf. Klime et al., 2010; *Bilanz nach einem Jahrzehnt*), with average scores in mathematics (about 580) exceeding scores in reading (about 570). Students perform best in science, reaching up to 590 points.

Furthermore, in all three domains the median exceeds mean test scores. This indicates that there appears to be more variation at the lower end of the performance scale, with more students performing relatively badly, thus pushing the median down. The mean/median comparison and its development may be regarded as first sign for whether IEOp changes over time. The data show that median and mean deviate only slightly more after than before the reform. The same applies to the variance of student test scores which do not change significantly over time. Finally, the dataset in this paper 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.4 for an overview).

[18] In 2000, there was no specific *grade-sample based* PISA-I sample available from the Institut zur Qualitätsentwicklung im Bildungswesen (IQB). However, PISA-2000 being the PISA-2000-E dataset is *ninth grade-based* (Baumert, 2002; PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich). But instead of the usual 80 replication weights, one then has about 768 replication weights. One also needs to pay closer attention regarding weighting as in the larger samples (PISA-E or IQB-LV) student observations per test domain may vary and thus different final weights may be required for each test domain.

[19] The IQB does not provide access to all *teacher*-datasets (as for 2006). Thus, I refrain from using data based purely on teacher questionnaires (*teacher*-dataset in Table A.1). However, variables derived from questions reappearing for cross-checking purposes in *student*, *teacher* and *school principal* questionnaires (e.g. gender) are taken into account.

3.2 Outcome Measure: Inequality of Educational Opportunity (IEOp)

The idea that societies should distribute opportunities equally has a long tradition within political philosophy. Following Rawls (1971) seminal contribution and its discussion (e.g. Sen (1980)), the idea established that a prerequisite for measuring Inequality of Opportunity (IOp) is distinguishing whether a form of inequality is acceptable or not within a society.[20] 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 recent 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* 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, for which one cannot be held responsible, e.g. your family household's socio-economic status (SES), parental education, gender, ethnicity or innate ability/talents.

Consider a sample of N students indexed by $i \in \{1, ..., N\}$. 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*.[21]

Thus, we can represent the population by a $(n \times m)$ 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 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*.

Thus, the concept of Inequality of Educational Opportunity (IEOp) can be translated as follows: Assuming talents to be distributed normally across the whole population, students who work harder,

[20] There is strong experimental evidence that people distinguish acceptable (fair) and unacceptable (unfair) income inequality (Cappelen, Sørensen, & Tungodden, 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*.

[21] This model could be extended to the case of having continuous elements in the vectors of *circumstances*/*efforts*.

that is, putting in greater *efforts*, should be rewarded by achieving good educational results regardless of their specific *circumstances* characteristics. Hence unfair 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 educational achievement due to variations in individual *efforts* are acceptable (*reward principle*). Thus, IEOp resembles differences between students that can only be attributed to *circumstances* beyond their control.

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 literature, I conduct an *ex-ante*[22], *between-types inequality* measurement approach of IOp which is in line with the *indirect*[23] approach, because it is based solely on the observed marginal distribution of *advantages* (test scores) given by the vector $y = \{y_1, \ldots, y_N\}$ and on the joint distribution of *advantages* and *circumstances* over the sample population $\{y, C_n\}$. Consequently, I follow the measurement approach of Ferreira and Gignoux (2013), because it requires fewer assumptions (e.g. on how to form *tranches* without directly observing *effort*). Moreover, given the high requirements for data availability, applying a non-parametric approach to conduct a *within-tranche inequality* decomposition (Checchi & Peragine, 2010)[24] may bias the measure of IEOp, because the more precisely one tries to design the partition, the smaller *cells* become.

Consequently, this paper adopts a parametric, *ex-ante* estimation approach to derive IEOp measures. I will 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 $y = f(C, E(C))$. Thus, for instance, it should be noted that unobserved innate ability is taken into account in this framework and is considered to be 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.[25] In the PISA-data evaluating students in the ninth grade, the individuals involved are on average about 15 years old. Hufe, Peichl, Roemer, and Ungerer (2017) argue that choices made before an age of consent (16) are likely beyond an individual's control.

[22] 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-type inequality* satisfying *ex-ante* compensation/reward principle. 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. However, if *effort* is distributed independently from *circumstances*, *ex-post* will be equivalent to *ex-ante* EOp (Ramos & Van de gaer, 2016, proposition II).

[23] One distinguishes between *direct* and *indirect* measurement approaches. As the *direct* approaches aim to model the opportunity sets explicitly, their implementation has been difficult because opportunities are not directly observable. Instead, *indirect* approaches measure IOp based on the observed joint distributions of outcomes and *circumstances*.

[24] Their approach essentially involves four steps. First, one defines the *advantage* variable. Second, one has to choose *circumstances*-variables to form *type* and *tranche*, and thus the respective *cells*. Assuming that ideally the within-type distribution should be the same, thirdly, one removes the within-cell score inequality to obtain a *smoothed* distribution. Fourth, total inequality in the *smoothed* scores distribution is decomposed into fair and unfair.

[25] See Appendix A.3.5 in the Appendix for further discussions of ability in the the context of measuring IEOp.

Therefore, it is plausible to assume that these tested students are (if at all) only partially responsible for their choices. My model of measuring IEOp considers the role of *circumstances*, *efforts* and their interplay. Following Ferreira and Gignoux (2013) a linear functional form is used:

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

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

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*. The aim at this point being to estimate the mean score outcome of each *type* conditional on *circumstances*, one proceeds with:

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

This will create a new, simulated distribution of scores, $\widehat{y} = \{\widehat{y_1}, \ldots, \widehat{y_N}\}$. Thus, every student is assigned the value of her 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. Now, the remaining issue is what inequality index I(.) to use. The literature on IOp in income has used the Mean Log Deviation (MLD) index due to its desirable properties (e.g. path independence). However, as outlined in Ferreira and Gignoux (2013), the MLD is not appropriate for measuring inequality in PISA test score data. The reason for this is that it is not ordinally invariant to the standardization of PISA test scores. In this case, the authors instead show that the most appropriate measure for IEOp consists of the variance. Being an absolute measure of inequality itself, it is ordinally invariant in the test score standardization and it 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*.

Thus 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.[26] Second, as shown in Ferreira and Gignoux (2011), the R^2 results in a meaningful summary statistic, the lower bound of the true IEOp. As the subject of concern is the total joint effect of all *circumstances* on educational outcomes as measured by test scores, the object of interest is to understand what percentage of the variation in scores y is causally explained by the overall effect of *circumstances* (directly and indirectly via *efforts*). With *efforts* being treated as generally unobserved, omitted *circumstances* variables, if we observed them, 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).[27] Third, $\widehat{\theta}_{IEOp}$ is a relative measure of IEOp that is cardinally invariant to the standardization of test scores. Moreover, 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 used in the literature (e.g. Ferreira and Gignoux (2013)). One needs to include as control those variables that represent *circumstances*, factors which a student cannot influence, but which can determine the dependent variable of interest, cognitive skills as measured by test scores.[28] They can be divided into two main groups, 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 in the base model specification.

Concerning student-level controls, students are on average 15.42 years old. The share of female students is slightly greater than that of male ones. This reflects the fact that in recent years female participation in *Gymnasium* has been steadily higher compared to that of male students (Prenzel et al., 2013; PISA 2012 - Fortschritte und Herausforderungen in Deutschland).

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

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

[28] Obviously, it is never possible to observe all of these circumstances variables, which is the reason that the IEOp measure will be a lower bound estimate as discussed in Section 3.2 (Niehues & Peichl, 2014).

The variable *migration background* indicates that about 17% of students have at least one foreign-born parent. But the variable *language spoken at home* improves the extent to which one controls 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. Less than half of the number of students with foreign traits said they spoke a different language at home. 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, i.e. the highest educational qualification achieved by one or both of a student's parents. This seems to be 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 serves to identify whether at least one parent has achieved an academic degree, ISCED level 5 or 6, in which case they would constitute an *academic household*. Table 1 shows that about 60% of students live in such *academic households*. A medium category comprises students whose highest level of parental education is post-secondary, non-tertiary education (ISCED level 3/4). Finally, students whose parents do not have any qualification beyond lower-secondary school (ISCED level 1/2) form the low parental education category.

In order to take the socio-economic status (SES) of a student's family background into account, I exploit, first, the *number of books at home* as a variable indicating the SES environment in which a student grows up. This variable is generated in all PISA studies and has been shown to be a good alternative proxy for the family SES, as household income is highly correlated with the amount of books in a 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 aged 15 the *number of books* variable represents *circumstances* controlling for family SES. I take the range of 101-500 books as a base category for this variable, as about 50% of students in the sample live in such a household. Similarly, the International Socio-Economic Index of Occupational Status (ISEI) index can be taken into account as a further control variable for socio-economic background. Higher ISEI scores correspond to higher levels of occupational status.[29]

I also control for family structure characteristics. First, I take into account whether a student lives in a single parent household. About 14% of all students are raised under such *circumstances*. It is important to control for them because they may indicate whether a student has grown up in a more stressful environment. Second, 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 taken into account. In the sample, the vast majority of fathers work full-time (FT), whereas most mothers are part-time employed (PT) (about 44%).

[29]The International Standard Classification of Occupation (ISCO) can serve as an alternative for describing parental SES. This involved obtaining parents' occupational data by asking open-ended questions, the responses to which were coded into ISCO codes. But this is not available for all PISA datasets, in contrast to the mapping of ISCO into ISEI indexes. See Ganzeboom, De Graaf, and Treiman (1992) for further details on this methodology.

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

Base-MT (2003-2012)	**Mean**	**SD**	**Min-Max**	**Missings (SD)**
Individual Characteristics				
Female	0.5289	0.4989	[0-1]	-
Age in years	15.43	0.49	[13,75-17,25]	-
Language spoken at home *(Base cat. German)*	0.0552	0.2285	[0-1]	0.0060 (0.0774)
Migration background *(Base cat. German)*	0.1679	0.3738	[0-1]	
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-I-2003/2006/2009/2012) and is weighted by the sampling weights provided in the PISA dataset (compare Appendix A.2.1). 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 indicated by italics. Finally, the number of observations and the G-8 reform dummy share is provided.

This is consistent with the predominant family model in Germany during the 2000s consisting of the father as the main bread-winner and a part-time working mother mainly in charge of child care.[30]

[30]In fact, with a school system based mostly on half-day schooling and working parents faced with only a limited number of institutions to take care of children after school, this family structure with a FT working father and mostly PT working mother prevailing in the data has been predominant in West Germany for many decades. However, a slow extension of all-day schools has been underway since the late 2000s, which may change the situation of student

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. The second step exploits the quasi-experimental variation of the G-8 reform to identify the effect of increased learning intensity on IEOp, using a difference-in-differences strategy based on forming reasonable treatment and control groups.

4.1 Estimating IEOp

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

$$Y_{ist} = \beta_0 + \beta_1(Individual\ Characteristics)_{ist} + \beta_2(Parental\ Characteristics)_{ist} + \beta_3(Socio{-}Economic\ Status)_{ist} + \beta_4(Family\ Character)_{ist} + FE(state/school)_s + \epsilon_{ist} \quad (8)$$

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.[31] 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 (see Appendix A.2.1 for details on the test metric).

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, Duflo, & Mullainathan, 2004). This requires the data sample to be as representative as possible with respect to the population of German ninth graders attending a *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.[32] Second, given the sampling strategy there may be correlation among observations in the error term. 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. Following the OECD guidelines I explain in detail how to estimate standard errors for the PISA data used in this study, in Appendix A.3.1.

cohorts born during the 2000s - a group of students that is not part of this paper's analysis period (2003-2012).

[31] Note that until Section 5.2, I focus in notation on the main specification, the *Base-MT* models covering the time period (2003-2012) with the general reform time set to take effect between 2006 and 2009, as also defined in Section 4.2 (for an overview see Appendix A.3.3). Then, the regression model can also be estimated separately by treatment and control groups only twice for the pooled pre-reform ((2000)-2003-2006) and post-reform (2009-(2012)) samples.

[32] Baumert and Prenzel (2008) discuss the PISA sampling strategy and the generation of the population weights. They argue for the 2006 wave of the PISA-E data that certain student groups might have been over/underrepresented, and that weights should be used to correct for this. These arguments also apply to the PISA-I data used in my paper.

As noted in Section 3.3, the control variables measuring *circumstances* in Equation (8) fall into four categories: Individual Characteristics (IC), Parental Characteristics (PC), Socio-Economic Status (SES) and Family Characteristics (FC) (cf. overview in Appendix A.3.2). Individual Characteristics include the *circumstances* variables: *age*, *gender* and *migration background*. As students were sampled based on being in the ninth grade, by controlling for *age*, differences in school entrance age and grade repetition potentially due to ability 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 (Klime et al., 2010; *PISA 2009 - Bilanz nach einem Jahrzehnt*). 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 and in that regard *parental education* can be assumed to constitute *circumstances* capturing investments into a student's early childhood. Similarly, favorable SES as measured by higher ISEI index values and/or more books available in a household should be expected to have a positive impact on a student's test scores. Higher SES of the family in which a student grows up may be indicative of better and easier access to support for dealing with school-related work. Otherwise, growing up with a single parent or with unemployed parents might have a negative effect on test scores, because such family conditions are more likely to be associated with adverse factors for skills formation or limited access to out-of-school support opportunities.

Moreover, in addition to these controls at the level of student i, the model in Equation (8) includes federal fixed effects (FEs) at the state level, the school level, or both. State fixed effects take into account time-invariant differences in the outcome variables between federal states due to, for instance, distinct political preferences for school policies. Moreover, the federal state in which a student goes to secondary school represents a *circumstance* variable, because it is beyond a student's control where parents decide to reside. Finally, adding school fixed effects allows me to capture quality differences among schools that can also exist within a federal state and to control for other school-level *circumstances*.[33]

[33]Unlike for the federal state where the decision of residence is made by the parents, students may have some influence over which school they attend, though this influence is likely very limited at age 10. However, the estimation results do not change much using different combinations of FEs, so this is probably not a major concern. By applying school-FEs without state-FEs, one still controls for characteristics both on school and state level (states are in charge of school policy). As the PISA test is not conducted in the same schools over the years, the school-FEs are wave-specific.

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, based on the implementation of the reform and the timing of this process across federal states, useful treatment and control groups can be formed.

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 2000	2003	2006	2009	2012	(if) Treatment cohort/grade affected 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**	**T**	-	**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**	**T**	-	**1st** cohort	**4th** cohort
Baden-Württemberg (BW)	2004/2005	2011/2012	**5**	**C**	**C**	**C**	**T**	**T**	-	**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
Hamburg (HB)	2002/2003	2009/2010	**5**	**C**	**C**	**C**	**T**1	**T**1	-	**3rd** cohort	**6th** cohort
Bremen (BR)	2004/2005	2011/2012	**5**	**C**	**C**	**C**	**T**1	**T**1	-	**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
Saxony-Anhalt (ST)	2003/2004	2006/2007	9								
		2007/2008	8						**1st** cohort		
		2008/2009	**7**	**C**	**C**	**T**	**T**	**T**	7^{th} *graders*		
		2009/2010	6							**2nd** cohort	**5th** cohort
		2010/2011	5	C	C	C	T	T		5^{th} *graders*	5^{th} *graders*
Mecklenburg-West Pomerania (MWP)	2004/2005	2007/2008	9								
		2008/2009	**8**	**C**	**C**	**T**	**T**	**T**	**1st** cohort		
		2009/2010	7						8^{th} *graders*		
		2010/2011	6							**1st** cohort	**4th** cohort
		2011/2012	5	C	C	C	T	T		5^{th} *graders*	5^{th} *graders*
Saxony (SN)	since 1949		**5**	**C**h	**C**h	**C**h	**C**h	**C**h	*hypothetical control group: always in treatment*		
Thuringia (TH)	since 1949		**5**	**C**h	**C**h	**C**h	**C**h	**C**h	*hypothetical control group: always in treatment*		
Saarland (SL)	2001/2002	2009/2010	**5**	**C**	**C**	**T**	**T**	**T**	**1st** cohort	**4th** cohort	**7th** cohort
Hesse (H)[a]	2004/05	2011/2012	5	**C**	**C**	**C**	***T***	**T**	-	*(less than 10%)*	$2^{nd}/3^{rd}/4^{th}$
	2005/06	2012/2013	***5***	**C**2	**C**2	**C**2	***C2***	**T**	-	**1st** cohort	**2nd/3rd/4th**
	2006/07	2013/2014	***5***	**C**	**C**	**C**	***C***	**T**	-	-	**2nd/3rd/4th**

[a] Hesse (H) only introduced the reform gradually across 3 school years (Table A.4 and Figure 1 as well as for regression specifications Figure A.4).
Notes: In this table, the Treatment/Control Groups are highlighted by rectangular boxes in the following way: For the **Base/Ext-ST/MT** Models: Treatment **T** ≡ red rectangle, **T1** ≡ red + magenta rectangle and **T2** ≡ red + magenta + violet rectangle. For the **Base/Ext-MT** Models: Control Group (**C**) ≡ blue rectangle; Control Group (**C1**) ≡ blue + green rectangle. Adding H to C1 would form *Control Group (C2)*. Finally, TH and S form a hypothetical *Control Group (Ch)* (always G-8 model). For an overview of treatment and control groups, please refer to Appendix A.3.3.

For seven out of fourteen federal states in which a reform took place, the introduction of the G-8 reform occurs between 2006 and 2009. PISA 2009 is thus the first post-treatment wave for the tested ninth graders in these states. The effect visible in 2009 will be denoted the "short-term" effect of the reform, while models including the 2012 wave capture the "medium-term" effect. Therefore, I define as *Baseline Model*: the *medium-term perspective (Base-**MT**)* model covering the time period (2003-2012) and the *short-term perspective (Base-**ST**)* model covering the time period (2003-2009).[34]

Medium-Term Model For the medium term, the reform takes effect in between 2006 and 2009. According to Table 2 seven federal states can be classified as the treatment group, because tested ninth graders were only in the G-8 model from 2009 onwards. *Treatment Group* **T2** 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, since many of the teachers working there were educated in the former German Democratic Republic (GDR). Focusing on West Germany only, the *Treatment Group* **T1** consists of BW, BV, LS, BR and HB. Finally, the most conservative set excludes the city states of HB and BR, because they may exhibit some inherently different characteristics from higher populated larger states. The *Treatment Group* **T**, that I focus on in the main specification, consists of just the three territorial West German states: BW, BV and LS.

The control group in the main specification, *Control Group* **C**, consists of two other 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 time period considered here, 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 their secondary school system after reunification. They form a *hypothetical Control Group* **Ch** to estimate the reform effect relative to 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.

Focusing on a treatment taking effect for students in grade nine from 2009 onwards, five federal states cannot be classified into either treatment or control group and are therefore excluded from the estimation. In Saarland (SL), the first West German state that implemented the reform, ninth graders were already being taught 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 states the reform affected students from ninth grade onwards, while in most other states students were affected starting in the fifth grade. In Hesse (H)[35] and North Rhine-Westphalia (NRW), ninth graders were only taught in a G-8 model from 2012 onwards, hence, after the 2006-2009 window.

[34]The *Extended Model* (Full-MT/ST) will additionally include the year 2000. For an overview, see Appendix A.3.3.

[35]Despite being 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, one could still classify Hesse as a control state in 2009 when only 10% of students tested had been already treated. Then the reform only occurred by 2012 and it is control state both in 2006 and 2009, whereas it can be classified as treatment state in 2012, when all ninth graders were in a G-8 model.

Short-Term Model For the short-term models, the reform time is set between 2006 and 2009. 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 *hypothetical Control Group* **Ch** made up of SN and TH. Now, North Rhine-Westphalia (NRW) as the federal state with the largest population in Germany can be added to the *Control Group* **C**, as in the *(Full)/Base-**ST*** model, ninth graders were taught in a G-9 model across the whole time period (2000)/2003 until 2009. This creates *Control Group* **C1** consisting of RP, SH, and NRW. One can add Hesse (H) to consider another territorial, West German state. For this purpose, one has to assume that Hesse is still part of control groups in 2009, as by then only ten percent of ninth graders had been treated (Table 2). Thus, *Control Group* **C2** is formed of RP, SH, NRW and H.

In summary, Table 2 already indicates that the most comparable setting for the **medium-term** models consists of the *Treatment Group* **T** and *Control Group* **C**, because it focuses on territorial, West German federal states that are very comparable in their relevant characteristics. Thereby, this setting still accounts for 37.6 out of 80.6 million people and thus about 50% of the German population and hence, will serve as the main specification for the *Base-**MT*** model.[36] The most comparable setting for the **short-term** models consists of the *Treatment Group* **T** and *Control Groups* **C** or **C1**. With the later group I account for 55.2 out of 80.6 million people and thus about 68% of the German population. Hence, there are two main control groups for the *Base-**ST*** model.[37]

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 the basic medium-term model, there are three states in the treatment group (Baden-Württemberg, Bavaria, and Lower Saxony) and two states in the control group (Rhineland-Palatinate, Schleswig-Holstein). Moreover, the pre-reform years are 2003 and 2006 ($before$), and 2009 and 2012 are the post-reform years ($after$). Then, the DiD strategy is implemented via the following regression model:

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

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 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 interaction term, being 1 if a student attends a *Gymnasium* in a treatment state after the

[36] However, in Section 5.3, I also conduct robustness checks using T1, T2 and Ch (Figure A.4 in Appendix A.5).

[37] However, robustness checks will be conducted using T1, T2 and C, Ch, C2 (Figure A.4 in Appendix A.5; for an overview see Appendix A.3.3).

implementation of the new G-8 model. Consequently, δ_1 is the coefficient of interest, that is, the causal reform effect of increased learning intensity on IEOp. X_{st} is a vector of potential state-level variables, including federal state and year fixed effects. In that way, one can take into account concerns whether there might have been differential "implementation" effects on the level of the federal states imposing the reform. For instance, certain school-system characteristics could have heterogeneously influenced the impact of the reform across federal states. Thus, I adjust the regression by including federal state fixed effects capturing any specific effects at the highest level of variation that is not captured by the DiD group specific means in Equation (9).

4.4 Selecting Appropriate Treatment/Control Group Settings

Internal Validity To begin with, German federal states share a similar legislative, cultural and economic framework and common qualification standards are coordinated by the Standing Conference of the Ministers of Education and Cultural Affairs (SC). Thus, exploiting variation in the reform implementation process across federal states can be considered more effective than relying on cross-national variation as in many existing studies (Wössmann, 2010).

Moreover, 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. For instance, half of all reform states started the transition into shortened school duration for 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.[38] Second, direct and indirect moving costs, including bureaucratic hurdles, appear to be reasons why few families with children move to another federal state in Germany. Third, strategic issues concerning competition for access to study programs also support the assumption that selection bias due to movement between states is unlikely. As a result of the reform several *double cohorts* graduating in between 2009 and 2016 temporarily increase the number of applicants for university studies in Germany. As this would inversely affect the probability of students immediately entering a study program of their choice, by completing the G-8 model students could at least insure themselves against the risk of having to take a gap year as their 14^{th} year of education.[39]

[38] The geographical maps in Figure A.3 in Appendix A.5 reveal the quick spread of the treatment across states.

[39] Instead of spending 13 years in school and having to wait 1 additional year before entering the study program of their choice, having 12 years of schooling before enrolling at a university even after 1 gap year would "save" 1 year.

By limiting the focus to a setting in which treatment states implemented the reform in school year 2004/2005, the quasi-experimental design is unlikely to suffer from estimation bias due to non-random reasons for introducing the G-8 reform slightly earlier or later among federal states.[40] Finally, for the internal validity of a DiD estimation to hold, the common time trend assumption requires that in the absence of the reform, both treatment and control group would have shown a parallel time trend. Therefore, I conduct Placebo Tests (Bertrand et al., 2004) and show pre-reform time trend graphs as robustness checks for the internal validity of my strategy in Section 5.3.

Treatment/Control Group Comparison Based on the quality of 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, significant observable pre-reform differences in the control variable sets might call the empirical strategy into question by suggesting the existence of unobserved compositional pre-reform differences. Following Imbens and Wooldridge (2009), Table 3 shows standardized means comparison tests for the control variable sets (Table 1) in the main treatment and control group settings to be considered when deriving the results. For the main time period models, *Model Base-MT/-ST*, the G-8 reform takes effect between 2006 and 2009. Hence, PISA waves 2003 and 2006 constitute the pre-treatment period. Table 3 shows that treatment and control groups have very similar characteristics in terms of the main *circumstances* variables used for the analysis. In particular, the main treatment group **T** and the main control group **C** appear to be very similar with no significant differences. There only is a small difference with respect to the ISEI index measure for SES - though the number of books measure of SES shows no significant difference between both groups (columns 1-3 in Table 3). This supports the internal validity of the strategy, because the main treatment and control groups consisting of West German states that are not city states turn out to be indeed comparable. Similarly, for *Model Base-ST*, when we extend the control group **C** by including additional large West German states, with North-Rhine Westphalia to form **C1** and in addition to Hesse to form **C2**, the control group not only gets larger in sample size, but remains comparable for nearly all *circumstances* variables. Finally, using enlarged treatment groups the results are still relatively robust in combination with standard control group (**C**; **C1/C2**).[41]

In summary, the pre-reform simple means comparison test for the main control variable set (Table 3) suggests that the DiD estimation approach outlined in Section 4.3 might be valid, at least for the following designs: for both *Model Base-MT/-ST* to compare *Treatment Group* **T** versus *Control Group* **C**; and for *Model Base-ST* to compare *Treatment Group* **T** versus *Control Group* **C1** or **C2**.

[40] Appendix A.3.4 shows that the treatment/control groups in *Base-ST/MT* are not different regarding the stability of the government that enforced the G-8 reform. Political preferences appear to stay similar across groups over the whole analysis period. Furthermore, systematic changes in the transitions flows between different tracks within the secondary school system appear to be neither of considerable concern (Huebener & Marcus, 2017).

[41] See Table A.5 for pre-reform tests of *circumstances* variables with enlarged treatment groups. Due to space constraints, pre-reform means comparisons between treatment and control groups for each point in time are not provided. But, they confirm that there have been no relevant differential changes across groups before the reform, thus supporting the validity of the estimation strategy. See also pre-trend DiD graph in Figure A.6 in Appendix A.5.

The comparability of pre-reform characteristics supports the assumption of internal validity that justifies exploiting the quasi-experimental design of the G-8 reform to estimate how the associated increase in learning intensity affected IEOp.

Table 3: Pre-Reform Treatment/Control Group Comparison of Control Variable Sets

	Base-MT (2003-2012)			**Base-ST (2003-2009)**				
	T	**C**	**T-C**	**T**	**C1**	**T-C1**	**C2**	**T-C2**
Individual characteristics								
Female	0.537	0.501	0.036	0.537	0.549	-0.011	0.543	-0.006
Age in years	15.488	15.468	0.020	15.488	15.464	0.025	15.474	0.015
Language at home not German	0.054	0.043	0.011	0.054	0.055	-0.002	0.054	-0.001
Migration background	0.183	0.144	0.039*	0.183	0.175	0.008	0.184	-0.000
Parental characteristics								
Parental Education (highest ISCED level):								
# ISCED-level (5-6):	0.662	0.637	0.025	0.662	0.648	0.014	0.642	0.019
# ISCED-level (3-4): [*Base*]	0.288	0.326	-0.037	0.288	0.288	0.000	0.291	-0.003
# ISCED-level (1-2):	0.044	0.026	0.018	0.044	0.036	0.008	0.037	0.007
Socio-Economic Status								
Number of books in household:								
# + 500:	0.226	0.233	-0.008	0.226	0.246	-0.020	0.243	-0.017
# 101-500: [*Base category*]	0.509	0.516	-0.007	0.509	0.481	0.028*	0.489	0.020
# 11-100:	0.246	0.228	0.019	0.246	0.228	0.018	0.222	0.025*
# max. 10:	0.010	0.014	-0.004	-0.001	0.010	0.015	0.016	-0.006*
Highest ISEI of parental job	59.103	57.072	2.031**	59.103	58.471	0.633	58.698	0.406
Family Characteristics								
Single Parent *(Base cat.: No)*	0.137	0.141	-0.004	0.137	0.150	-0.013	0.156	-0.019*
Father employment status								
# full-time (FT): [*Base cat.*]	0.854	0.841	0.013	0.854	0.843	0.012	0.845	0.009
# part-time (PT):	0.065	0.063	0.001	0.065	0.058	0.007	0.054	0.010
# unemployed (UE):	0.024	0.032	-0.007	0.024	0.026	-0.001	0.026	-0.001
# out-of-labor force (OLF) :	0.033	0.035	0.033	0.033	0.000	-0.005	0.034	-0.001
Mother employment status								
# full-time (FT): [*Base cat.*]	0.217	0.213	0.004	0.217	0.232	-0.015	0.232	-0.014
# part-time (PT):	0.515	0.501	0.014	0.515	0.476	0.040**	0.476	0.039***
# unemployed (UE):	0.061	0.075	-0.014	0.061	0.063	-0.002	0.061	-0.001
# out-of-labor force (OLF):	0.194	0.187	0.194	0.194	0.202	-0.008	0.204	-0.009
Number of students	2,175	347	-	2,175	1,861	-	2,334	-

Notes: This table shows a *two-sample t-test* for comparing the main control variables in the pre-reform period of the main specification to be considered in this paper between *Treatment* and *Control Group* (see Section 3.3 and 4.1). This is for both *T vs. C* in **Model Base-MT** and for *T vs. C/C1* in **Model Base-ST** the respective pooled average of control variables in *PISA-I-2003 and -2006*. Stars denote the significance of the simple mean difference in pre-reform characteristics in the form of *p-values* as follows: *** p<0.01; ** p<0.05; * p<0.1 ; *Source:* Author's calculations based on PISA-I-data 2003, 2006, 2009, 2012. Table A.5 contains the table with additional Treatment Groups (T1/T2) and Control Groups (Ch).

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* - taken from the representative *grade-based* PISA-I-dataset (Section 3.1).[42] Section 5.1 explains the first-step and Section 5.2 the second-step results for the main specifications (*Base-ST/MT*). Section 5.3 provides robustness checks with extended treatment and control group settings, while Section 5.4 rationalizes the results.

5.1 Main Results: First Step

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.[43] 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 take into account the stratified data structure and representativeness of each observation.

However, before measuring IEOp, it is also important to check how *circumstances* variables directly affect cognitive skills as measured by test scores. Detailed regression output per test domain is only provided for the main specification in the medium-term (*Base-**MT***) model: **T** versus **C** (Table A.6, A.7, A.8). As these output tables are sufficient to show the main patterns of how control variables affect test scores, *first-step* results are only provided for mathematics in the short-term (*Base-**ST***) model: **T** versus **C** (columns (1-4) in Table A.9) or **T** versus **C1/C2** (Table A.10).[44]

Medium-Term First-Step Results[45] Starting with the *Base-**MT** Model*, the following patterns can be observed concerning how *circumstances* affect cognitive skills as measured by 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

[42] That is, in the rest 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 rep. *grade-based* PISA test cohort in any of the test years that form the sample (2003, 2006, 2009, 2012 in *Base-MT*) (*stdpvsubject3*): This allows me to interpret the coefficients relative to the student performance across the sample period.

[43] In Section 5.3 for robustness check purposes, for all main specifications for 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.3.1 for details on computing standard errors.

[44] In the literature, it is common to focus on mathematics scores due to their comparability characteristics (Philippis & Rossi, 2017). However, *first-step* regressions for science/reading scores for *Model-ST* are available upon request.

[45] 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 (4) refer to *Control Group* **C**, the last four to *Treatment Group* **T**. Within both *Groups*, the first two columns refer to the "*Before*" reform period (2003-2006), the last two repeat regressions using only "*After*" reform (2009-2012) data. Each odd numbered column only includes federal state fixed effects, while each even one additionally controls for school fixed effects.

by 45-65% and in the science test by 30-50% in terms of an international standard deviation (SD), even though the effect size slightly declines in the post-reform period across both treatment and control group. However, female students increase their reading performance by 30-40% in terms of one international SD. This is consistent with the aforementioned evidence suggesting the existence of gender-specific achievement differences in educational outcomes (Niederle & Vesterlund, 2010).

All the other control variable estimates are fairly robust in their signs independent of the test domain. As expected, the *age effect* is negative. Those, for instance, that started school at an older age or that 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 lower in all three testing domains. Additionally controlling for whether a *foreign language is spoken at home*, the negative effect shrinks as expected. This suggests that the degree to which a migrant student's family is integrated may be key for test scores, in particular for the domain of reading. Looking at the socioeconomic-status (SES) of the household in which a student grows up, a higher *amount of books* than the base category (101-500) proves to be positively correlated with test score performance. Likewise, the higher the ISEI index of a job in the family, the higher is the positive effect on educational outcomes.[46] 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* should be 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). However, the effect for reading is insignificant. 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.[47] But Parental Characteristics have less effect on scores once individual *circumstances* are taken into account. Finally, *family structure* and *employment status* show no clear patterns.

Short-Term First-Step Results Repeating this exercise for the short term (*Base-**ST** Model*), with the preferred specification being **T/C1**, *first-step* regressions for mathematics test scores show that the same general patterns as described for the medium term can be observed. Female students perform better in reading, but significantly worse in mathematics and science tests. *Age* and *migration background* tend to be negatively correlated with educational achievement across all domains. A more favorable SES family background, such as growing up in an *academic household*, has a positive impact on cognitive skills as measured by test scores. (See Table A.10 in Appendix A.4).

In summary, *first-step* regressions demonstrate that for both the short and medium-term horizon most of the *circumstances* variables affect the PISA test scores into the expected directions.

[46] With the average family's highest job ISEI index being 45, an effect on test scores of 0.001 translates into 4.5% of an international standard deviation. See also Appendix A.2.1 in Appendix A.4 for further explanations.

[47] Furthermore, more 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.

The fact that these patterns are consistent over varying time horizons and across different control group settings confirms that the chosen *circumstances* variables were appropriately selected. Furthermore, the explanatory power of these *first-step* regressions remains in a range of about 15-35% across the different specifications. Consequently, the level of the IEOp measure found in this paper can be categorized in a lower bound within the range of few available IEOp estimates for European countries, such as Ferreira and Gignoux (2013), who find based on PISA-I-2006 data, that about 35% of test score variation in Germany can be attributed to *circumstances* or Carneiro (2008) finding about 40% IEOp for the case of Portugal. Moreover, using additionally school instead of federal states fixed effects (FE) increases the observed IEOp measures across all *first-step* regression specifications suggesting that school-level characteristics matter for educational outcome measures.

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

Now, we can switch our attention to the *second-step* of the estimation approach, the DiD framework. The IEOp measure that we just derived 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.

Medium-Term Results Starting with the main treatment/control group specification, the medium-term (*Base-**MT** Model*) results are shown in the second column of Table 4. The top panel shows 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 considerably increased due to the G-8 reform. That is, the share of inequality in test scores that can be attributed to *circumstances* has increased. 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 19% of educational outcomes can be additionally considered to constitute IEOp. These results are statistically significant, with standard errors computed as explained in Appendix A.3.1. Thus, given initial values of 20-30% in IEOp, DiD estimates would correspond to a relative increase in IEOp of about 50% in response to increased learning intensity as induced by the G-8 reform, that is, in the medium-term horizon this rise in IEOp is economically significant. Zooming in, one can further note that for Control Group **C** IEOp seems to have considerably decreased in the time period after the reform. In contrast, for the Treatment Group **T** the level of IEOp appears to have remained practically unchanged across all three domains. Thus, in this setting, due to the increase in learning intensity in treated states the role of *circumstances* remained constant, whereas in absence of shortening school duration, IEOp tends to have decreased. The *Base-**MT** Model* 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 (Table 2 & Figure 1).

Table 4: Main Results: *Short-Term (**ST**) vs. Medium-Term (**MT**) Reform Effect*

Subject	**Treatment Group (T) vs. Control Group (C)** with **School-FE**					
	Base-**ST** (2003-2009) as in **Figure A.4**			Base-**MT** (2003-2012) as in **Figure A.4**		
Reading	**C**	**T**	**Δ *(T - C)***	**C**	**T**	**Δ *(T - C)***
Before Reform	0.242	0.173	-0.068	0.242	0.173	-0.068
	(0.057)	(0.032)	(0.065)	(0.057)	(0.032)	(0.065)
After Reform	0.161	0.196	0.036	0.162	0.206	0.044
	(0.060)	(0.037)	(0.071)	(0.033)	(0.023)	(0.040)
Change in R2	-0.081	0.023	**0.104**	-0.079	0.033	**0.112**
	(0.083)	(0.048)	(0.096)	(0.066)	(0.039)	(0.077)
Mathematics	**C**	**T**	**Δ *(T - C)***	**C**	**T**	**Δ *(T - C)***
Before Reform	0.353	0.257	-0.097	0.353	0.257	-0.097
	(0.060)	(0.032)	(0.068)	(0.060)	(0.032)	(0.068)
After Reform	0.270	0.223	-0.047	0.190	0.232	0.042
	(0.073)	(0.041)	(0.084)	(0.039)	(0.028)	(0.048)
Change in R2	-0.084	-0.034	**0.050**	-0.163	-0.025	**0.139**
	(0.094)	(0.052)	(0.108)	(0.071)	(0.043)	(0.083)
Science	**C**	**T**	**Δ *(T - C)***	**C**	**T**	**Δ *(T - C)***
Before Reform	0.363	0.203	-0.160	0.363	0.203	-0.160
	(0.052)	(0.024)	(0.058)	(0.052)	(0.024)	(0.058)
After Reform	0.257	0.195	-0.062	0.173	0.202	0.028
	(0.067)	(0.033)	(0.075)	(0.047)	(0.024)	(0.053)
Change in R2	-0.106	-0.008	**0.098**	-0.189	-0.001	**0.188**
	(0.085)	(0.041)	(0.095)	(0.071)	(0.034)	(0.078)

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.3.1, 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/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* and *migration background* (based on (parental) birth place)
(iii) parental characteristics (PC): *highest parental education level (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) - mother/father unemployed (UE) - mother/father out of labor force (OLF)
Compare: First-step regressions are conducted according to Equation (8). The respective detailed results are only illustrated for mathematics test scores. The derivation of IEOp measures in the second sub-panel (mathematics) are provided in Table A.9. The first-step regression results for the other subjects are available upon request.
Source: Author's calculations based on PISA-I-data 2003, 2006, 2009, 2012.

Short-Term Results To learn about the short-term effects, it is interesting to see how results change for the main treatment/control group specification when conducting the same two-step estimation procedure for the *Base-**ST** Model* covering only years 2003 until 2009. Therefore, the left-hand panels in Table 4 show the short-term effects of increased learning intensity on IEOp focusing mainly on the first student cohorts treated by the G-8 reform. The DiD estimates remain

considerably positive across all test domains. Now, 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 considerably lower in the short term - if different from zero at all - compared to its significant size in the medium term (right-hand panel in Table 4). 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. In conclusion, the main treatment/control group specification suggests that increased learning intensity aggravates IEOp, not in the short term for the first treated student cohorts, but only after some time in the medium term, that is, for the fourth or later treated cohorts. The effects are stronger for mathematics/science than for reading.

To understand how the G-8 reform changed educational opportunities in the *Gymnasium*, it is useful to expand the short-term model analysis to consider treatment/control group specifications that bear even more external validity for the German school system. For instance, with **C1** about 68% of the German high school student population can be considered in the short-term reform analysis. DiD estimation results for this extended treatment/control group specification are presented in the left panel of Table 5. Interestingly, the short-term effect of the reform vanishes across all three test domains and for both fixed effects (FE) settings. Thus, there appears to be no effect on IEOp in response to the G-8 reform. However, the IEOp measures still range between 15 to 25%, their magnitude increasing from reading to science to mathematics. Students in both treatment and control group experience similar increases in IEOp, such that in total the DiD effect is canceled out. Furthermore, the DiD results reveal that changes in IEOp are very small. But as in the main specification with *Control Group* ***C***, these results are not statistically significant. Finally, extending the control group to include both North-Rhine-Westphalia (NRW) and Hesse (H), thus covering 75% of the German high school student population, results for this treatment/control group specification (**T**/**C2**) are shown in the right-hand column of Table 5. The DiD estimation findings on the effect of the G-8 reform are very similar across **T**/**C1** and **T**/**C2** 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 depends on the time horizon. Focusing on the short term (*Model Base-**ST*** (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 (Table 5). 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 Base-**ST*** setting, but results are barely statistically significant (Table 4). However, taking a medium-term perspective on the G-8 reform (*Model Base-**MT*** (2003-2012)) shows that increased learning intensity induced by the reform causally increases the IEOp measures.

Table 5: Main Results: *Treatment Group (**T**) vs. Control Group (**C1**) or Control Group (**C2**)*

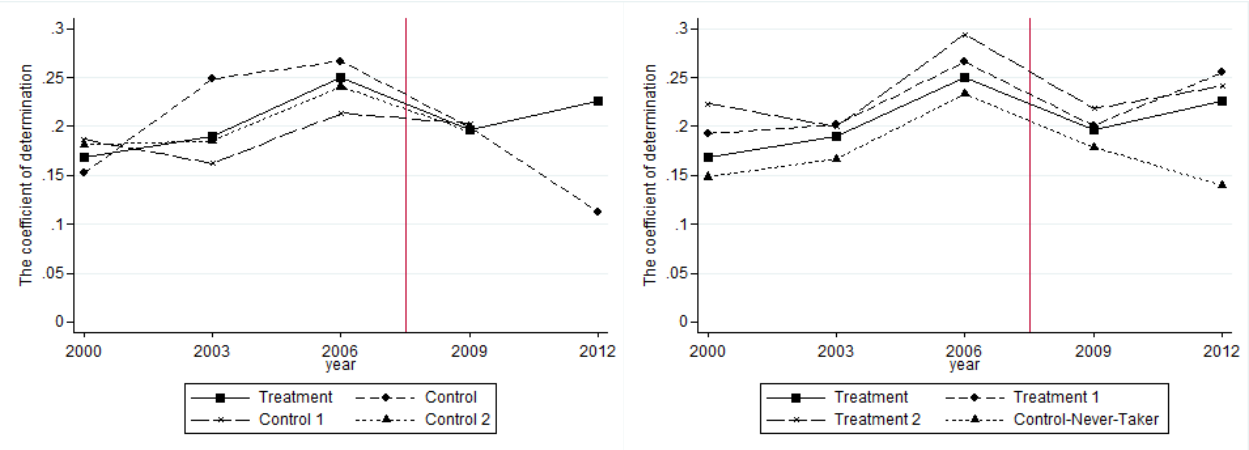
Subject	*Short-Term (**ST**) (2003-2009) Model* - **T** vs. **C1/C2** — (**Figure A.4**)					
	T vs. C1: with School-FE			T vs. C2: with School-FE		
Reading	**C1**	**T**	**Δ *(T - C1)***	**C2**	**T**	**Δ *(T - C2)***
Before (2003-2006)	0.163	0.173	0.010	0.175	0.173	-0.002
	(0.032)	(0.032)	(0.045)	(0.029)	(0.032)	(0.043)
After (2009)	0.183	0.196	0.013	0.179	0.196	0.018
	(0.030)	(0.037)	(0.047)	(0.021)	(0.037)	(0.042)
Change in R2	0.020	0.023	**0.003**	0.004	0.023	**0.019**
	(0.044)	(0.048)	(0.065)	(0.035)	(0.048)	(0.060)
Mathematics	**C1**	**T**	**Δ *(T - C1)***	**C2**	**T**	**Δ *(T - C2)***
Before (2003-2006)	0.216	0.257	0.041	0.243	0.257	0.014
	(0.029)	(0.032)	(0.043)	(0.031)	(0.032)	(0.045)
After (2009)	0.233	0.223	-0.010	0.222	0.223	0.001
	(0.033)	(0.041)	(0.053)	(0.027)	(0.041)	(0.049)
Change in R2	0.017	-0.034	**-0.051**	-0.020	-0.034	**-0.013**
	(0.044)	(0.052)	(0.068)	(0.041)	(0.052)	(0.066)
Science	**C1**	**T**	**Δ *(T - C1)***	**C2**	**T**	**Δ *(T - C2)***
Before (2003-2006)	0.205	0.203	-0.002	0.214	0.203	-0.011
	(0.024)	(0.024)	(0.034)	(0.021)	(0.024)	(0.032)
After (2009)	0.215	0.195	-0.020	0.195	0.195	0.000
	(0.039)	(0.033)	(0.051)	(0.030)	(0.033)	(0.045)
Change in R2	0.010	-0.008	**-0.018**	-0.020	-0.008	**0.011**
	(0.046)	(0.041)	(0.061)	(0.037)	(0.041)	(0.055)

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.3.1, clustering at the federal state level. DiD results are estimated according to Equation (9) taking into account population weights and school fixed effects. Positive changes in R^2 indicate increasing IEOp/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 parental education level (ISCED-level 1-2/ 3-4/ 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) - mother/father unemployed (UE) - mother/father out of labor force (OLF)
Compare: First-step regressions are conducted according to Equation (8). Due to space constraints, the respective detailed results are only illustrated for mathematics test scores. The derivation of IEOp measures in the second sub-panel (mathematics) are provided in Table A.10. The first-step regression results for the other subjects are available upon request from the author. *Source:* Author's calculations based on PISA-I-data 2003, 2006, 2009.

The observed rise in inequality of opportunity is statistically significant and covers about half of the general IEOp measure estimated for students attending German secondary schools. Results reveal that for students at the *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.[48]

[48]For a graphical illustration of main results, please refer to Figure A.8 in Appendix A.4. To investigate whether

Figure 2: DiD Pre-Trend Graphs of IEOp measure for main Treatment/Control Groups



Notes: This figure shows the DiD graphs for IEOp measures based on mathematics test scores confirming that the parallel trend assumption holds. Due to the data as discussed in Section 3, in the main regression settings I only use the time frame 2003 to 2009 for the short-term or 2003 to 2012 for the medium-term analysis. Compare also Figure A.5 and Figure A.6 for IEOp measures based on the other test domains and Figure A.7 in Appendix A.5.

5.3 Robustness Checks

Placebo Test To evaluate the plausibility of the quasi-experimental identification strategy that allows a causal interpretation of the the effects of the increased learning intensity induced by the G-8 reform 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**/**C**/(C1/C2) in Table A.12 in Appendix A.4). 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. Thus, Placebo Tests confirm the plausibility for interpreting the main estimation results as causal effects of the reform on IEOp.

To investigate the robustness of the previous section's 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 the treatment group. Third, I show how results change based on enlarged control groups consisting of states that never changed their academic track.[49]

this increase in IEOp is long-lasting, one would ideally need to consider longer time periods, that are not yet available. However, once shifting attention to student cohorts that are far away from the first treated ones, potential curricular additional reforms undertaken in response to the initial G-8 reform (Table A.4) 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).

[49]The main output tables for robustness checks are shown in Appendix A.4: Table A.13 to A.19. All of these 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. First, each table shows results only with respect to one time period (e.g. *Model*

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 in particular *adjusted* R^2 measures of IEOp behave.The adjusted R^2 can help to detect which *Control set* 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 (Table A.13, A.14, A.15, A.16). This supports the empirical strategy taken for the main results based on using always all six variable sets in the *first-step* regression, as it is not rendering estimates that differ much from 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.

Finally, multi-level regressions confirm that school level *circumstances* are indeed already considered by school fixed effects. Moreover, using school fixed effects or only federal state effects to measure IEOp does not change DiD results. This indicates that sorting based on schools appears to be not a concern, which also supports the internal validity of the empirical strategy taken (see Figure A.8).

Extending Treatment Groups Next, it is useful to repeat the estimations with extended treatment groups to investigate the potential external validity of this paper's main results. Therefore, all main regressions discussed in Section 5.2 are rerun with *Treatment Group* **T1** including the two West German city states Hamburg and Bremen, and for *Treatment Group* **T2**, which is **T1** plus Berlin and Brandenburg. Increasing the treatment group, on average DiD reform effects become smaller, for instance, in the regression settings with *Control Group* **C** (Table A.13, A.14) both in *Model Base-MT* and *-ST*, the increasing effect on IEOp declines as we move from **T** to **T1** to **T2** consistently within each test domain and across all *Control sets*. However, the general direction of the reform effect as found in Section 5.2 remains (right panel in Figure 2). Moreover, *Model Base-ST* regressions with respect to **C1** (Table A.15) or **C2** (Table A.16) confirm the zero reform effect on IEOp across the enlarged treatment groups. 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 no longer carry implications that are likely to be valid for the entire German secondary school system.

Base-MT (2003-2012)) and one control group. Second, in each overview table, results are provided for all three testing domains. Third, for each combination of treatment/control group and test domain, six rows of results are provided as indicated by column (5), the *Control set*. *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.3.2). Then, subsequently additional control variables are added, until in set 6 all available *circumstances* are applied together in the *first-step* regression. Finally, four DiD estimates are presented in each row: Column (6) provides the standard R^2 measure based DiD estimate that only takes into account federal states fixed effects (FE); Column (8) presents the same estimate but taking additionally school FE into account.Column (7/9) provide the IEOp estimates relying on *adjusted* R^2 measures.

Extending Control Groups As mentioned in Section 4.2, one could also compare treatment groups with federal states that always maintained the same length for the *Gymnasium*. When using *hypothetical Control Group* **Ch**, the DiD results change and in all specifications a decrease in IEOp can be found. This effect is strongest in the short term (*Model Base-ST*), but rather vanishes in the medium term (*Model Base-MT*). In that regard, the pattern is consistent with the normal control groups because when taking the *hypothetical Control Group* into account, in relative terms, the reform effect tends to shift towards more IEOp in the medium term compared to the short term.[50]

Extending Time Period Studied Taking also PISA-E-2000 data into account, I rerun the main estimation framework. It is reassuring that the results are similar to the DiD estimation for *Model* ***Base****-MT*. Focusing on the main specification with *Control Group* **C**, a slightly increasing or likely zero impact of increased learning intensity induced by the G-8 reform on IEOp can be found as we extend *Model* ***Base****-MT* by including an additional pre-reform period, that is, for the full time period *Model* ***Full****-MT* (**2000**-2012). In conclusion, results based on *Model* ***Base****-MT* and the most convincing treatment/control group specifications carry on their validity also for a slightly broader time period setting. This in turn supports the external validity of the main estimation findings.[51]

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 ($y_{i,t-1}$):

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

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) in Section 3.2) as long as the *circumstances* vector contains mostly family background variables. In this regard, $\widehat{\theta}_{IEOp}$ can be connected to measures of intergenerational educational immobility, which can be used to measure social (im)mobility (such as β Equation (10)).

In analogy, this is also related to the findings that childhood wealth can serve as a proxy for *circumstances* explaining future wealth inequality (Boserup, Kopczuk, & Kreiner, 2016). Moreover, intergenerational income elasticity and the Gini coefficient of incomes 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.

[50] For completeness, the DiD estimation results are provided in Table A.17 for *Model Base-MT* and Table A.18 for *Base-ST*. For the *hypothetical Control Group* the DiD estimate is negative in the short term and gets less negative in the medium term, whereas for control group C the effect is positive in the short term and rises in the medium term. However, pre-reform tests for Ch (Table A.5) indicate that any results based on it should be interpreted cautiously.

[51] For *Control Group* **C**, please refer to Table A.19 and for completeness for *Ch* tables can be provided upon request.

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 equality of educational opportunities would be welfare enhancing in a society with meritocratic preferences. However, 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 the importance of *circumstances* for the determination of a student's educational outcome decreases.

Returning to the G-8 reform, we 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.[52] As the implementation process of the reform suggests, the reform-induced increase in learning intensity surprised affected students and their parents 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 can not 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 may have been more relevant for the first affected student cohorts.

Second, in the medium term, after favorable *circumstances* had time to adapt and provide support to the associated students, *upward* mobility would be lessened in conjunction with *downward* mobility. 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 to 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 abov (cf. Figure A.9). There has been a rise in extra tuition following the reform, with this effect being stronger in the treatment compared to the control group federal states. Moreover, the increase in extra tuition has been more pronounced for students from more privileged family environments (*circumstances*), such as those living in academic households. A further potential mechanism involves the time investment made by mothers depending on their educational background into their children.

[52]The high correlation of parental education and a student's probability of entering the *Gymnasium* has been shown (e.g. Klime et al. (2010; *PISA 2009*) to be persistent in the German school system at least over the last two decades.

More highly educated women were more likely to work part-time in order to support their children in treatment states after the reform compared to the situation in control group states (Figure A.10).

Moreover, looking across the medium-term effect evidence (Table 4) DiD estimates of the effect of increased learning intensity on IEOp reveal some subject-related patterns. The level of IEOp is consistently higher for both mathematics and science compared to 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 indirectly in other school courses as well as in everyday life - reading being often a necessary prerequisite to simply 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 which are mainly accumulated through taught courses at school and less likely to be learnt indirectly through other courses at school or in everyday life. Thus, for the complementary skills set required by mathematics/science, it seems to be plausible that positive *circumstances* such as growing up in an academic household are relatively more important than for reading. Beneficial resources improving the accumulation of skills relevant for mathematics/science tend to be more exclusive than those important for reading. 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. It might raise the question of whether in order to improve reading skills, current curricula and teaching methods need to be adjusted. But, 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, 2016; 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 only gender-specific differences may be reinforced. The fact, that there appear to be no SES-specific differences supports my findings: the observed overall increase in IEOp seems to be mainly driven by heterogeneous parental support opportunities to deal 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 German secondary school system, in particular the academic track, the *Gymnasium*, to 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.

The first step of the analysis involves measuring IEOp as share in the variance of standardized PISA test scores that can be only attributed to *circumstances* beyond an individual's control. In that regard, the paper contributes to the still limited literature on measuring IOp with respect to educational outcomes by adding evidence on how IEOp has changed over time in Germany. 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.

For the second step a DiD estimation strategy with treatment and control groups chosen according to the implementation of the G-8 reform can be conducted to derive causal estimates. 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, Spieβ, and Staneva (2016). 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 until today. While there have been a few studies that tried to detect the direct effect of the G-8 reform on PISA test scores, they do not focus on the question of how the reform-induced increase in learning intensity may have changed the equality of educational opportunities. Thus, the analysis in this paper 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). This is of policy relevance in the debate about more intense schooling. To lower IEOp despite higher learning intensity, whole-day schooling and methods reducing the dependence of educational support on *circumstances* may be a solution (Deckers, Falk, Kosse, Pinger, & Schildberg-Hörisch, 2017).

Alternatively, to maintain equality of opportunity, when reducing the length of 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 mechanism of how IEOp and social mobility interact is likely to be very important for understanding phenomena such as the high persistence in the observed intergenerational transmission of educational achievement. For that reason, it is relevant to investigate further what aspects of increased learning intensity may have differential impact on mediocre but privileged students in terms of their background variables (potentially pushing *downward* mobility) compared to students with less advantageous *circumstances*. Moreover, it would be interesting to evaluate if in the long term increased learning intensity tends to inhibit *upward* mobility by increasing dependence on favorable *circumstances*. Generally, the fact that social mobility consists of both an *upward* and *downward* component seems to be still neglected, in the sense that focus appears to have shifted onto how to improve *upward* mobility, ignoring that this cannot be discussed independently from removing rigidities that potentially limit *downward* mobility. Thus, it is important to understand the effects of compressing education on IEOp, and its implications for social mobility.[53]

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 may involve unintended and underestimated welfare costs. Learning intensity is a key factor for the design of an educational process or system and has implications for both the effectiveness and efficiency of (non-)cognitive skill formation. A better understanding of the relationship between schooling duration, in particular intensity, and IEOp would also be important in the context of evaluating the conditions under which *Signaling* or respectively *Human Capital Theory* may be more important in evaluating the welfare benefits and cost 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, 2017; Boneva & Rauh, 2017b), it is economically desirable to achieve more equality of educational opportunities. Therefore, this paper shows that one so far neglected policy margin involves implementing an appropriate level of educational intensity, taking into account not only efficiency considerations, but also 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, 2018). Furthermore, additional work is needed to establish how IEOp translates into social mobility. This in turn may then allow us to assess the welfare effects of IEOp with respect to its impact on future income and wealth inequality - allowing the evaluation of new policy recommendations aimed at improving equality of opportunity in order to tackle challenges surrounding high levels of inequality.

[53]Figure A.1 in Appendix A.5 shows numbers on absolute educational mobility across OECD countries. Ideally, the theory of how learning (duration and intensity) and IEOp as well as how IEOp and social mobility are linked together would allow quantifying precisely the role of learning intensity for absolute educational mobility, thus social mobility.

References

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

Anderson, G., Fruehauf, T., Pittau, M. G., & Zelli, R. (2015). Evaluating Progress Toward an Equal Opportunity Goal: Assessing the German Educational Reforms of the First Decade of the 21st Century. *Working Paper University of Toronto*(552), 1–27.

Andrietti, V. (2016). The causal effects of an intensified curriculum on cognitive skills: Evidence from a natural experiment. *SSRN Electronic Journal*, 1–47. doi: 10.2139/ssrn.2774520

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

Boca, D. D., Piazzalunga, D., & Pronzato, C. (2017). Early Childcare, Child Cognitive Outcomes and Inequalities in the UK. *HCEO Working Paper Series*, 1–16.

Boneva, T., & Rauh, C. (2017a). Parental Beliefs about Returns to Educational Investments - The Later the Better? *Journal of the European Economic Association*, 1–52. doi: 10.2139/ssrn.2764288

Boneva, T., & Rauh, C. (2017b). *Socio-Economic Gaps in University Enrollment: The Role of Perceived Pecuniary and Non-Pecuniary Returns* (Doctoral dissertation, University College London).

Boserup, S. H., Kopczuk, W., & Kreiner, C. T. (2016). Born with a Silver Spoon? Danish Evidence on Wealth Inequality in Childhood. *NBER Working Paper*(22549), 1–40. doi: 10.3386/w22549

Bratti, M., Checchi, D., & de Blasio, G. (2008). Does the Expansion of Higher Education Increase the Equality of Educational Opportunities? Evidence from Italy. *Labour*, *22*(s1), 53–88. doi: 10.1111/j.1467-9914.2008.00411

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

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).

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., Saez, E., Turner, N., & Yagan, D. (2017). Mobility Report Cards: The Role of Colleges in Intergenerational Mobility. *NBER Working Paper*(23618), 1–94. doi: 10.3386/w23618

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. (2017). Socio-Economic Status and Inequalities in Children's IQ and Economic Preferences. *Discussion Paper Series IZA*(11158), 1–60.

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

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

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., Spieβ, 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.

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

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

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

Oppedisano, V., & Turati, G. (2015). What are the causes of educational inequality and of its evolution over time in Europe? Evidence from PISA. *Education Economics*, *23*(1), 3–24. doi: 10.1080/09645292.2012.736475

Philippis, M. D., & Rossi, F. (2017). *Parents, Schools and Human Capital Differences across Countries* (Doctoral dissertation, London School of Economics).

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

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), 1–30.

Raitano, M., & Vona, F. (2016). Assessing students' equality of opportunity in OECD countries: the role of national- and school-level policies. *Applied Economics*, *48*(33), 3148–3163. doi: 10.1080/00036846.2015.1136396

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.

Riphahn, R. T., & Trübswetter, P. (2013). The intergenerational transmission of education and equality of educational opportunity in East and West Germany. *Applied Economics*, *45*(22), 3183–3196. doi: 10.1080/00036846.2012.703314

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

Roemer, J. E., & Trannoy, A. (2015). Equality of Opportunity. In *Handbook of income distribution* (Vol. 2, pp. 217–300). doi: 10.1016/B978-0-444-59428-0.00005-9

Rothstein, J. (2018). Inequality of Educational Opportunity? Schools as Mediators of the Intergenerational Transmission of Income. *NBER Working Paper*, *24537*, 1–35.

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), 234–270.

List of Abbreviations

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.
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 increased learning intensity in such a way that by the end of grade 9 post-reform, students have received about the same amount of instruction, and covered the same curriculum as students that had completed two-thirds of grade 10 pre-reform. Learning intensity, thus, corresponds to the *intensive* margin as it reflects the amount of content (curriculum) to be studied in a constant amount of instruction time, whereas school duration (e.g. number of years/days) refers to the *extensive* margin.

Plausible Value Following OECD (2009; PISA Data Analysis Manual) 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 On the G-8 reform

A.1.1 The Reform Debate

The first PISA-study in 2000 received broad public attention in Germany, because it revealed that German students achieved within the OECD only below average test scores (the so-called "PISA-shock"). Debates over how to improve the German school system ensued (cf. Anderson, Fruehauf, Pittau, and Zelli (2015)). 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. I refer to the last column in Table A.4 for an overview on the status quo of the reform as of school year 2015/16.

Mainly three reasons were given to justify its introduction. First, it was intended to reduce the relatively high age of university graduates in Germany. For one, this was said to increase their competitiveness on the labor market compared to the (on average) younger graduates in other OECD countries (OECD, 2005; *Education at a Glance 2005*). Furthermore, with students entering the job market one year earlier, working lifetime would be extended, such that the reform was said to contribute to stabilizing the social security system of a society facing demographic changes. For instance, younger university graduates would start paying social security contributions earlier and over a longer timespan. 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. For this purpose, adjusting secondary school duration towards the average among other European nations appeared to be sensible. Moreover, it was thought that the reform would serve as an incentive for then younger school graduates to strive for obtaining a university degree, thereby increasing Germany's below average rate of university graduates per birth cohort compared to other OECD countries.

Opponents, however, have argued that the reform may worsen the human capital skills formation among students due to the intensified educational experience (cf. Section 2.2). Furthermore, parental complaints about increased stress for students due to less free time emphasize the fear of negative impacts on both academic performance and the development of non-cognitive skills typically formed by non-academic recreational activities (Thiel, Thomsen, & Büttner, 2014). However, the majority of East Germans support shortened duration of the academic track, whereas the opposite is true across the West German federal states that only recently adopted the new system (Wössmann et al., 2015).

A.1.2 Related Literature on the Reform

Despite the public controversy over this reform that has even induced some federal states to reverse it (last column in Table A.4 in Appendix A.4), 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 respective statistical offices have conducted studies comparing students' results in central exit examinations (*Abitur*) in the *double cohort*, that is the year when both the last G-9 model and the first G-8 model cohorts graduated from the *Gymnasium* (Figure 1). Generally, these statistical evaluations have found no systematic performance differences in central exit exams between students with eight or nine years of schooling.[54]

Furthermore, for the federal state of Saxony-Anhalt (ST), a small series of papers have analyzed different aspects of the G-8 reform (Thiel et al., 2014; Büttner & Thomsen, 2015; T. Meyer & Thomsen, 2016). In summary, they analyze the reform's effects on academic achievement in central exit examinations 2007, when the *double cohort* graduated in ST. Findings show that - due to more intense schooling - exam results deteriorated significantly for mathematics, but remained unaffected for German literature suggesting that learning intensity ratios differ across subjects. Moreover, no significant negative effects on student's soft skills are detected, opposing claims that increased learning intensity and associated reduced time for non-school related activities may have adversely affected non-cognitive skills formation. In line with this result, Quis and Reif (2017) show that the more intense schooling experience had only limited impacts 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 no negative effects of the G-8 reform on the ability, motivation and likelihood of students' entering university education.[55]

Recently, a few papers have started to use more representative data that might be 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, e.g. by applying methods taking into account variation across states (e.g. DiD). For instance, two studies related to this paper exploit the reform setting using as variables for educational outcome, the standardized PISA-test scores for academic-track students.[56]

[54] For instance, there are federal states with no observed difference (Saarland (SL), North Rhine-Westphalia (NRW)), states where G-9 model students remained slightly better (Baden-Württemberg (BW)), states with the opposite pattern (Hesse (HE), Berlin (BE)) and states where results differ depending on the subject (Bavaria (BV)).

[55] But the reform influenced post-secondary school 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, because they are now 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.

[56] 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 DiD-estimation framework, finding a positive effect of about 0.15 standard deviations in test scores, with stronger effects for students with a migration background similar to the later results by Andrietti (2016).

Andrietti (2016) uses this representative data set in order to exploit the G-8 reform for conducting a DiD-estimation of increased learning intensity on test scores. He finds that the average treatment effect of the reform is significant and positive in all three educational outcomes.[57]

In contrast to Huebener and Marcus (2017), Andrietti (2016) finds no evidence for a significant increase in general grade retention rates. Instead, his results suggest that grade repetition may only have increased for boys and students with a migration background. This may indicate that the G-8 reform caused distributional changes in educational outcomes and thus may have affected IEOp, however, Andrietti (2016) does not really address this issue. Huebener et al. (2017) show based on state regulations of timetables for secondary school, 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 almost five years. They suggest that this increased instruction time improved student performance on average in all three PISA test domains. However, the effect size is small, with about 6 percent of a standard deviation in scores. Moreover, for low-performing students positive effects are insignificant, whereas their high-performing peers experience significant, positive effects indicative of a widening of the performance gap among students in the *Gymnasium*. In that regard, Huebener et al. (2017) focus on the increased instruction time effect, whereas Andrietti (2016) 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 regarding the reform effect follows Andrietti (2016) with emphasis on the effects of increased learning intensity. While these studies estimate the direct reform effect on test scores, they do not tackle the question if and how increasing learning intensity may have changed IEOp. In this paper, I shift focus in the analysis of the G-8 reform onto distributional concerns, that is its consequences on IEOp. In other words, I answer the question of whether the G-8 reform may be considered to be a *selective*, i.e. at least maintaining test score results and at the same time increasing IEOp or an *inclusive* reform, i.e. that at least maintains test score results while decreasing IEOp (Checchi & van de Werfhorst, 2018). In that regard, this paper is among the first evaluating the G-8 reform based on German specific PISA data in order to analyze its impact on IEOp.

[57] Treated students in a G-8 model experience an improvement of about 0.095-0.145 standard deviations in PISA-test scores. Furthermore, the author tries to estimate the effects of the approximate pure instruction time increase on scores and finds similar results: a twenty-hour increase distributed over grades 5-9 or a ten-hour increase distributed over grades 8-9, correspond on average to an improvement of 0.08-0.15 standard deviations, depending on the subject.

A.2 Data

A.2.1 Background Information on the PISA data

Since 2000, the OECD conducts every three years the PISA study in order to measure the performance of 15 year-old students with respect to three basic competencies (*Life skills*) regarded to be of special importance for a person's future success when approaching the end of compulsory schooling age, namely *reading*, *mathematical* and *scientific* literacy (cf. Klime et al. (2010; *PISA 2009 - Bilanz nach einem Jahrzehnt*); OECD (2010, *PISA 2009 Assessment Framework*); OECD (2013a, *PISA 2012 Assessment Framework*)). Instead of testing if students master particular curricular contents, the idea of PISA is to evaluate more general skills, such as the ability to apply knowledge in the three tested domains for solving real-world problems, i.e. skills that students should learn before leaving school as they are essential for participating in modern society (OECD, 2001; *Knowledge and Skills for Life*).[58] Apart from general cognitive skills, PISA also collects rich information on family and school characteristics, based on the fact that students, their parents, teachers and school's principals are supposed to fill out questionnaires.

Concerning the PISA procedure, for each test cycle, the OECD chooses an international contractor 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, 2009; *PISA Data Analysis Manual*). On the national level, a PISA National Project Manager is chosen to be in charge of the test implementation. The test procedure itself resembles a ***two-stage stratified randomized survey test design***. First, as a primary sample unit, schools with eligible students are randomly selected (with a minimum of 150 schools in each country) to get a random sample representative of all school types across all regions within a country. Then, as second-step sampling units, eligible students (15-year-olds)[59] are randomly selected within sampled schools to reach a minimum of 4500 students in the sample. Each student within a school receives distinct combinations of approved test questions on all three PISA domains.[60]

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 providing information concerning their family, school and

[58] 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; *The PISA 2003 Assessment Framework*) and in particular to chapter 1 of OECD (2009; *PISA Data Analysis Manual - SPSS, Second Edition*).

[59] 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), regardless of the type of institution and of whether they were in full-time or part-time education (OECD, 2013b; *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed*).

[60] 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 test, e.g. OECD (2013a, *PISA 2012 Assessment and Analytical Framework*) and OECD (2012; *PISA-2009 Technical Report*).

socio-economic background as well as on their attitudes, motivations or aspirations. After the test has been evaluated on the national level (supervised by the international contractor), results are published by the OECD in 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 (*PISA-2003 Technical Report*, *PISA-2006 Technical Report*, *PISA-2009 Technical Report*)). As illustrated by Ferreira and Gignoux (2013), the so-called *Item Response theory (IRT)* is used to back out the distribution of the latent variable, cognitive skills (measured as test scores per domain), from individual *item* responses, taking into account the particular difficulty of an *item*. However, to address the issue of small-sample measurement error, because, for instance, not all students answer all *items*, for each student so called *Plausible Values* of test results are provided. First, the marginal distribution of the latent variable conditional on the *item* responses and a set of observables is estimated, that is, 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 ($M = 5$).[61]

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 standardized distribution of test scores with values of 500 (100).[62]
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 said to consist of about 70 points. This corresponds to about two years of schooling in the average OECD country (OECD, 2013b).[63]

[61] Conducting estimations with PISA-test scores, the OECD (2010, *PISA 2009 Assessment Framework - Key Competencies in Reading, Mathematics and Science*) suggests estimating any statistic s by using each of M (*Plausible Value*) datasets separately (getting $\hat{s}_m$) and then averaging them over M to get a final estimate $\hat{s}$.

[62] 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. 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 allow us to compare skills in the three domains across students and over time. To deal with difficulties in constructing meaningful measures of IEOp based on these standardized test scores, the variance appears to be a useful index as explained by Ferreira and Gignoux (2013).

[63] For instance, in *PISA-I-2012* the average difference in mathematics test scores between top and bottom quarters of students in OECD countries is 128 score points. However, most differences related to socio-demographic characteristics are smaller than an entire *proficiency level*. For instance, 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 the 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; *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed*)), and students in city schools score about 31 points higher than those in rural ones, on average (see Table II.3.3a in OECD (2013b; *PISA 2012 Results*)).

The advantage of using PISA test scores as measures for cognitive skills, in contrast to alternatives like GPA, is that its representativeness makes it possible to compare student cohorts both over time and across or within countries (federal states). However, 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 (Gamboa & Waltenberg, 2012). However, the sampling standards of PISA require that participating countries not exclude more than 5% of students from the population eligible to be tested. 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 serious concern for the representativeness of PISA test data (OECD, 2010; OECD, 2013a).

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% (Klime et al., 2010; *PISA 2009 - Bilanz nach einem Jahrzehnt*). Moreover, there is no evidence for selection on observables for those selected who do actually not take the test.

Third, another concern is that schools or more specifically teachers may bias comparability of scores across time and regions, if they systematically train or motivate students for the test. However, based on student information about their motivations 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 2000-2009, such concerns seem not to be relevant (Klime et al., 2010). The majority of teachers report that they tried to make students familiar with general testing strategies, but did not train them specifically for the test. Only half of teachers indicated that they had trained their students for PISA at all and those who did started no earlier than one month before the test. Vice versa, only 25% of participating students indicate having prepared for the reading part, only 13% for mathematics, and only 8% for the science section in the test.[64] As Klime et al. (2010) show, it is plausible to assume that test results in Germany are not systematically influenced by any preparation behaviour. Although questionnaires provide evidence on test motivation having slightly increased between 2000 and 2009, the correlation between test motivation and scores remains very low (on average 0.05). Thus, it is unlikely that test motivation biases results (Wössmann, 2010).

In conclusion, the advantages of using PISA data as measure of cognitive skills seem to dominate potential caveats, which is the reason, I decided to use them - in line with the studies mentioned in Appendix A.1.2. 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.

[64] Note that as affected students and teachers are only informed about two months before the PISA test takes place, and given the general limited probability of being selected for the test and as there are no particular incentives for neither teachers nor students to prepare for the test, the effect of potential preparation on scores appears to be limited.

A.2.2 Data Sources

For more information on the German specific PISA-data of each test cycle and availability of these datasets, the reader is recommended to refer to the Institut zur Qualitätsentwicklung im Bildungswesen (Institute for Educational Quality Improvement) (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. 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. http://doi.org/10.5159/IQB_PISA_2003_v1
- 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. 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. 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. http://doi.org/10.5159/IQB_PISA_2012_v2

A.3 Empirical Strategy and Robustness - Appendix

A.3.1 On the computation of standard errors including replication weights

Throughout the paper, standard errors for both steps of the DiD regressions (Section 4) using PISA data are constructed taking into account that student performance is reported through Plausible Values. Even though the average of five plausible values as a measure of individual performance guarantees that estimates of group level means and regression coefficients remain unbiased, measures of dispersion must take into account the within-student variability in Plausible Values (PVs).

As explained in the manuals provided by the OECD (2009; *PISA Data Analysis Manual - SPSS, Second Edition*), one should compute standard errors by running regressions with individual test scores as a dependent variable five times, thereby using all plausible values in turn. For each regression I employ an estimator for the sampling variance clustered at the federal state level. The final sampling variance, SV, is given by the average of sampling variances obtained with the five plausible values. In addition, standard errors are inflated by the imputation variance (IV) because test scores measure the latent cognitive skills of students with error. The imputation variance, IV , is estimated as the average squared deviation between the estimates obtained with each Plausible Value and the final estimate (obtained using the average of Plausible Values), with the appropriate degree of freedom adjustment ($IV = \frac{1}{4}\left(\hat{\theta}_i - \hat{\theta}\right)$ where $\hat{\theta}_i$ is the estimate for each of five Plausible Values and is the final estimate).

Finally, as shown by OECD (2009; *PISA Data Analysis Manual - SPSS, Second Edition*), 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 takes into account the two-stage stratified sampling design of the PISA test. For this method, each regression is iterated over the 80 sets of *replication weights* provided in the PISA dataset. The SV estimate is then 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 that governs the variability between different sets of replication weights, set at 0.5 in the PISA study).

Standard errors in all *first-step* and in all *second-step* regressions are based on this method. For computational convenience and similar to Philippis and Rossi (2017), I use the "unbiased shortcut" procedure described in OECD (2009; *PISA Data Analysis Manual - SPSS, Second Edition*), which relies on only one set of plausible values for estimating the sampling variance (whereas the imputation is estimated using all five sets). One should note that for estimating effects of the reform on test scores, Andrietti (2016) relies only on clustering standard errors on the state level and argues that applying a wild t-bootstrap procedure produces rather similar results. Huebener et al. (2017) also focus on clustering methods. However, given the arguments provided on the sampling strategy used to generate the PISA scores, estimating standard errors taking into account both replication weights and Plausible Values is more reliable.

A.3.2 List of *Circumstances* Variables

1. Individual Characteristics (IC):
   - (I) *gender* [Base: male] and *age* (in years)
   - (II) *migration background* [Base: none] 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*
4. Family Characteristics (FC)
   - (VI) *single parent household* [Base: none]
   - (VII) *mother/father employment status* [Base: FT]

A.3.3 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-**MT**): covers time period (**2003-2012**)
   - *Extended Model: medium-term perspective* (Full-**MT**): covers time period (**2000-2012**)
   - ***Baseline Model: short-term*** *perspective* (Base-**ST**): covers time period (**2003-2009**)
   - *Extended Model: short-term perspective* (Full-**ST**): covers time period (**2000-2009**)
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**)*
   - *Treatment Group (**T1**)*: ***BW***, ***BV***, ***LS***, *Bremen (**BR**)*, *Hamburg (**HB**)*
   - *Treatment Group (**T2**)*: ***BW***, ***BV***, ***LS***, ***BR***, ***HB***, *Berlin (**BE**)*, *Brandenburg (**BB**)*
   - *Control Group (**C**)*: *Rhineland-Palatinate (**RP**)*, *Schleswig-Holstein (**SH**)*
   - only for the **short-term** models ((Full)/Base-**ST**):
     - *Control Group (**C1**)*: ***RP***, ***SH***, *North Rhine-Westphalia (**NRW**)*
     - *Control Group (**C2**)*: ***RP***, ***SH***, ***NRW***, *Hesse (**H**)*
       One can add **H** to the *Control Group C1* to consider another West German state. One has to make the assumption that **H** can still be classified into the *Control Group* in 2009, as by then only 10% of ninth graders may have been treated (Table 2).

- *hypothetical Control Group_(**Ch**)*: *Saxony (**SN**)*, *Thuringia (**TH**)*
- *Never-Takers Control Group_(**C-NT**)*: ***RP***, ***SH***, ***SN***, ***TH***

3. *Neither Treatment nor Control Group:*
   - for the **medium-term** models: **SL**, **ST**, **MWP**, **H**, **NRW**
   - for the **short-term** models: **SL**, **ST**, **MWP**

A.3.4 Further Aspects on the Internal Validity of Empirical Strategy

It should be noted that 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 **medium-/short-term** models, i.e. **T/C** in Base-**MT** (2003-2012) or **T/C1** in Base-**ST** (2003-2009). In fact, though it is true that federal states governed by conservative parties (CDU) tended to be the first to introduce the G-8 reform, Table 2 shows that all states implemented the reform within a short time frame and maintained the same government for the time period of my analysis. Therefore, for the first affected cohorts considered in this paper, the whole reform period was usually dominated by the same government. Moreover, by controlling for federal state fixed effects and conducting a DiD, one takes into account general differences due to political parties governing the respective state and in charge of implementing the reform. The fact that there have not been systematic changes in federal state governments across treatment and control groups shortly before the respective reform implementation is supportive evidence that for the period considered it is plausible to assume a comparability in the stability of respective federal state 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 Greens/SPD: thus in the analysis period, the reform was implemented by the same government party and it is plausible to assume that, due to the time lag for government policy to take effect, that school policy up until school year 2012/2013 was conducted by the same party.
  - **BV**: Conservative (CSU)-led government for decades until the present day and thus plausible to assume that school policy was mainly conducted by the same political party.
  - **LS**: Conservatives (CDU) led the government over the whole analysis period (2003-2013); afterwards/beforehand the government was led by SPD. Thus, plausible to assume that for the whole analysis period, school policy was influenced by the same political party.
  - **BR**: Social-Democratic (SPD) government for decades until present day and thus plausible to assume that school policy was mainly conducted by the same political party.

  - **HB**: Social-Democratic (SPD) government for decades until 2001 and again since 2011; in between Conservatives led the government 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/C2)
  - **RP**: Social-Democratic (SPD) government for decades until present day and thus plausible to assume that school policy was mainly conducted by the same political party.
  - **SH**: Social-Democratic (SPD) government for decades (1988 - 2005) and in present day (since 2012). From 2005 until 2012, the government was led by Conservatives, but from 2010-2012 in a grand coalition with the Social-Democrats. Due to the narrow majorities, school policy for *Gymnasium* remained similar during the analysis period.
  - **NRW**: Social-Democratic (SPD) government for decades until 2005 and again from 2010 to 2017. In between the government was led by Conservatives (CDU). However, the reform had already been enacted under the Social-Democrat government, and despite the intermediate change, school policy remained similar for the analysis period. In particular, as I only take NRW into account for the Control Group in the short-term period model.
  - **H**: Social Democrats led the government until 1999. From 1999 onwards, the Conservatives led the government and, after some turmoil in 2009, they continued to govern from 2010 until the present day. They were thus in charge for the implementation of the reform.

Thus, as mentioned in the main text, focusing on the analysis period that covers only the first affected cohorts, the main DiD assumption appears to be plausible. However, given the reversal decisions in some federal states in recent years after the analysis period considered in this paper, a similar evaluation may be less plausible over time due to recent policy changes. The reform has become a topic on the political agenda in most federal states starting in 2010 until the present day (cf. last column in Table A.4). But for the very first affected cohorts, there is no systematic change in governments comparing treatment and control group states in the time period (2003-2012).

A.3.5 On ability in the context of measuring IEOp and within the DiD framework

Though, one may have concerns about differences in ability, one should take into account the following. First, the measurement framework takes into account any unchangeable features of cognitive skills as unobserved *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, in particular during early childhood, through epigenetic processes. This may explain, in the spirit of the Human Capital Theory literature (Heckman), that, for instance, Boca, Piazzalunga, and Pronzato (2017) find that attending childcare institutions can significantly improve children's cognitive skills, in particular those from disadvantaged SES. The measurement framework fully takes into account the role of ability, both as unobserved *circumstance* and *effort*. Thus, it is a lower bound measure.

Concerning the DiD, the only assumption that I need to make is the innocuous one that generally, the distribution in cognitive abilities of students between 2003 and 2012 did not systematically change between German federal states. Given the fact that moving behavior across federal states is unlikely to have occurred, this means that we simply assume that cognitive skills did not suddenly change across states in these years for any other reason than the reform. There is no way to provide evidence on whether there are systematic differences in ability across federal states. However, even if they existed, the DiD framework would take that into account. Therefore, there are not many plausible reasons given the short time period and the controls enacted via the DiD to think that there may have been some significant changes in cognitive skills differing among federal states that may somehow bias any results. In any case, these thoughts should be of much less concern in this quasi-experimental setting than in other settings of published journal articles that measure IEOp across countries. Moreover, as the reform only affects students from age 10 onwards, and treatment merely involves more intense instruction, but not different content, I claim that these concerns which can neither be addressed by empirical methods nor available data, are negligible and comparable to those in accepted published articles estimating returns to schooling.

A.4 Supplementary Tables

Table A.1: Available grade-sample based PISA-I datasets

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

[a] Number of observations for students as included in the PISA datasets (2000, 2003, 2006, 2009, 2012) as available from the IQB based on the *grade-based sample* (see also Appendix A.2.2). Note, that here the *student*-dataset includes both the original student questionnaire answers and their parental ones.

[b] Note that for the year 2000, there was no specific *grade-based PISA-I-sample* available from IQB. However, *PISA-2000* being the *PISA-2000-E* dataset is *ninth grade-based* (Baumert, 2002; *PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich*). Therefore, it has a lower number of variables, but a higher number of observations than the other datasets.

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

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

Table A.2: Available grade-sample PISA-E datasets

	Before reform			After reform	
Dataset	PISA-2000-E [b]	PISA-2003-E	PISA-2006-E [c]	IQB-LV-2008/2009 [d]	IQB-LV-2012
Student-dataset:					
# of variables	914	698	883	494	911
# of students[a]	34,754	46,185	39,573	39,663	44,584
test scores[d]	reading	mathematics	science	reading	mathematics
School-dataset:					
# of variables	470	633	387	-	176
# of schools	1,342	1,411	1,496	-	1,048
Teacher-dataset:					
# of variables	-	-	194	503	422
# of teachers	-	-	14,572	3,376	4,213

[a] Number of observations for students as included in the *PISA-E-Datasets (2000, 2003, 2006)* and *IQB-Ländervergleichsstudie(LV)-Sprachen-2009*, *IQB-Ländervergleichsstudie(LV)-2012* as available from the IQB based on the *grade-based sample* (see also Appendix A.2.2). Note, that here the *student*-dataset includes only the original student questionnaire answers as the parental ones are only provided for *PISA-2006-E*.

[b] For years 2000 and 2003, the *teacher*-dataset was not part of the Germany-specific PISA dataset provided by the IQB, as in other years. Similarly, the *school*-dataset was not provided for the *IQB-LV-Sprachen-2008/2009*.

[c] Note that for *PISA-2006-E*, the IQB only provides an *age-based sample*.

[d] For years 2009, only readings scores were assessed, whereas in 2012 only the mathematics and science test scores can be compared with previous *PISA-E* results.

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

Table A.3: Descriptive Statistics: Outcome Variables and Sample Size

	Before reform			**After reform**	
Student Test Scores 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	572.6754	584.7017	571.1871	580.472	576.1879
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
Number of federal states	16	16	16	16	16
Number of schools	409	62	67	68	78
Number 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.2.1 and OECD (2012; *PISA-2009 Technical Report*)). 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 (Table A.1), 2003, 2006, 2009, 2012).

Table A.4: Overview of the "G-8 reform" across federal states, sorted by year of *double cohort*

Federal state	Type of Federal State			Reform Timeline		Gymnasium		Reform reversal[c]
	West/East	City/Terr.	Population[a]	Begins	Ends	Type	Grade	yes/no
Saxony (SN)	East	territorial	4,0 mio	-	-	5-12	-	no[d]
Thuringia (TH)	East	territorial	2,2 mio	-	-	5-12	-	no[d]
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[e]
Hamburg (HB)	West	city state	1,7 mio	2002/2003	2009/2010	5-12	5^{th}	no[f]
Bavaria (BV)[b]	West	territorial	12,5 mio	2004/2005	2010/2011	5-12	5^{th},6^{th}	yes[g]
Lower Saxony (LS)[b]	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}	[o]
Schleswig-Holstein (SH)	West	territorial	2,8 mio	2008/2009	2015/2016	5-13	5^{th}	[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] In Bavaria (BV) and Lower Saxony (LS), the 6^{th} and 5^{th} grade were allocated into the G-8 model in the same school year, suggesting that educational intensity might be slightly stronger for the then 6^{th} graders that had to compensate the shortened school duration over 7 instead of 8 years, as the then 5^{th} grade students. However, the 9^{th} graders in 2009 in BV and LS were affected by the reform right from the 5^{th} grade.

[c] 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

[d] Since 1949, these states have implemented a G-8 model in the GDR and never had a G-9 model.

[e] Gymnasium remains in G-8 model, but in a comprehensive school G-13 model is possible.

[f] Gymnasium remains in G-8 model, whereas the so-called Stadtschule as a comprehensive school offers a G-13 model.

[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 intro. in # % of all Normal Gymnasium (5-12) 2004/2005: 10%; 2005/2006: 60%; 2006/2007: 30% double cohorts: 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 onwards

[o] Always maintained schools with G-9 model (G-13 model), but since 2008/2009 G-8 model offered at 19 Gymnasien

[p] *Since 2011/2012* schools are allowed by state law to offer a G-9 model (11/ 99 schools), G-8 model or both (4/99 schools).

Table A.5: Pre-Reform Treatment/Control Group Comparison of Variables for additional Groups

	Base-MT (2003-2012)			Model Base-ST (2003-2009)				
	T	**Ch**	**T-Ch**	**T1**	**C1**	**T1-C1**	**T2**	**T2-C1**
Individual characteristics								
Female	0.537	0.560	-0.023	0.533	0.549	-0.016	0.535	-0.014
Age in years	15.488	15.514	-0.026	15.492	15.464	0.028*	15.474	0.010
Lang. at home not German	0.054	0.018	0.036***	0.056	0.055	0.000	0.056	0.001
Migration background	0.183	0.059	0.124***	0.188	0.175	0.013	0.184	0.009
Parental characteristics								
Parental Education (highest ISCED level):								
# ISCED-level (5-6):	0.662	0.641	0.021	0.654	0.648	0.006	0.658	0.011
# ISCED-level (3-4): [*Base*]	0.288	0.310	-0.021	0.285	0.288	-0.003	0.280	-0.008
# ISCED-level (1-2):	0.044	0.012	0.033***	0.046	0.036	0.010	0.045	0.009
Socio-Economic Status								
Number of books in household:								
# + 500:	0.226	0.153	0.073***	0.229	0.246	-0.017	0.220	-0.026**
# 101-500: [*Base cat.*]	0.509	0.448	0.061***	0.501	0.481	0.019	0.496	0.015
# 11-100:	0.246	0.341	-0.095***	0.244	0.228	0.015	0.257	0.029**
# max. 10:	0.010	0.023	-0.013**	0.010	0.015	-0.005	0.011	-0.004
highest ISEI-level of job in the family								
# Highest ISEI-level:	59.103	55.590	3.514***	58.975	58.471	0.504	58.656	0.185
Family Characteristics								
Single Parent *(Base: No)*:	0.137	0.176	-0.039**	0.140	0.150	-0.010	0.168	0.018
Father: *employment status*								
# full-time (FT): [*Base cat.*]	0.854	0.841	0.013	0.847	0.843	0.004	0.832	-0.011
# part-time (PT):	0.065	0.036	0.029***	0.065	0.058	0.007	0.065	0.007
# unemployed (UE):	0.024	0.058	-0.033***	0.025	0.026	-0.001	0.034	0.009*
# out-of-labor force (OLF):	0.033	0.026	0.007	0.031	0.033	-0.001	0.031	-0.001
Mother: *employment status*								
# full-time (FT): [*Base cat.*]	0.217	0.614	-0.397***	0.216	0.232	-0.016	0.297	0.065***
# part-time (PT):	0.515	0.198	0.318***	0.511	0.476	0.036**	0.448	-0.027*
# unemployed (UE):	0.061	0.096	-0.035***	0.060	0.063	-0.003	0.067	0.004
# out-of-labor force (OLF):	0.194	0.063	0.132***	0.195	0.202	-0.008	0.169	-0.033***
Number of students	2,175	607	-	2,365	1,861	-	2,999	-

Notes: This table shows a *two-sample t-test* for comparing the main control variables of the additional specification between the *Treatment* and *Control Group* in the pre-reform period apart from Table 3.
This is for both *T vs. Ch* in **Model Base-MT** and for *T1/T2 vs. C1* in **Model Base-ST** the respective pooled average of control variables in *PISA-I-2003 and -2006*. Stars denote significance of the simple mean difference in pre-reform characteristics in the form of *p-values* as follows: *** p<0.01; ** p<0.05; * p<0.1; *Source:* Author's calculation based on PISA-I-data 2003, 2006, 2009, 2012.

Table A.6: Main Results for *Model Base-**MT***: 1^{st} step to derive IEOp measure for *reading scores*

DEPENDENT VARIABLE:		**Control Group (C)**				**Treatment Group (T)**			
	READING Test Scores in PISA (STDPVREAD3)	Before (2003-2006) (1)	(2)	After (2009-2012) (3)	(4)	Before (2003-2006) (5)	(6)	After (2009-2012) (7)	(8)
CONTROL: **Individual Characteristics (IC)**									
i)	female	-0.035	0.066	0.384***	0.395***	0.268***	0.294***	0.385***	0.399***
		(0.115)	(0.105)	(0.081)	(0.084)	(0.040)	(0.041)	(0.038)	(0.040)
	age in years	-0.047	-0.059	-0.251**	-0.250**	-0.199***	-0.163**	-0.169***	-0.179***
		(0.159)	(0.183)	(0.099)	(0.114)	(0.060)	(0.068)	(0.037)	(0.039)
ii)	migration background	-0.252	-0.234	-0.158*	-0.170*	-0.102	-0.061	-0.102*	-0.078
		(0.266)	(0.241)	(0.095)	(0.090)	(0.080)	(0.077)	(0.059)	(0.058)
	NO German spoken at home	-0.511	-0.494	-0.065	-0.168	-0.308***	-0.339***	-0.139	-0.155**
		(0.581)	(0.530)	(0.199)	(0.195)	(0.118)	(0.113)	(0.086)	(0.073)
CONTROL: **Parental Characteristics (PC)**									
iii)	Parental Education: [Base: *ISCED-level (3-4)*]								
	# at most lower sec. educ. (ISCED-level (1-2))	-0.512**	-0.486**	0.087	0.105	-0.274**	-0.285**	-0.028	0.003
		(0.216)	(0.225)	(0.196)	(0.192)	(0.129)	(0.112)	(0.060)	(0.058)
	# tertiary educ. (ISCED-level (5-6))	-0.144	-0.159	0.159*	0.147	-0.002	-0.011	-0.022	-0.036
		(0.177)	(0.156)	(0.093)	(0.106)	(0.058)	(0.061)	(0.048)	(0.049)
CONTROL: **Socio-Economic Status (SES)**									
iv)	No. of books in household [Base: *101-500*]								
	# max 10 books	0.132	0.169	-0.556**	-0.459*	-0.458*	-0.489**	-0.509***	-0.441***
		(0.410)	(0.395)	(0.267)	(0.273)	(0.266)	(0.228)	(0.142)	(0.127)
	# 11-100 books	-0.162	-0.126	-0.132	-0.084	-0.305***	-0.307***	-0.172***	-0.141***
		(0.195)	(0.214)	(0.113)	(0.122)	(0.056)	(0.054)	(0.048)	(0.043)
	# more than 500 books	0.195	0.204*	0.093	0.081	0.099	0.082	0.079	0.084
		(0.121)	(0.117)	(0.068)	(0.067)	(0.066)	(0.064)	(0.048)	(0.053)
v)	highest ISEI-level of parental jobs	0.007	0.007	0.002	0.001	0.001	0.001	0.004**	0.003***
		(0.006)	(0.005)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)
CONTROL: **Family Characteristics (FC)**									
vi)	family structure [Base: *No*]								
	single parent household	-0.032	0.079	0.319***	0.272**	0.092	0.09	0.109*	0.122*
		(0.294)	(0.261)	(0.117)	(0.120)	(0.061)	(0.061)	(0.061)	(0.065)
vii)	father: employment [Base: *full-time (FT)*]								
	# part-time (PT)	-0.352*	-0.3	-0.228	-0.241	-0.095	-0.091	-0.130*	-0.104
		(0.194)	(0.208)	(0.266)	(0.280)	(0.103)	(0.096)	(0.074)	(0.078)
	# unemployed (UE)	0.397	0.382	0.252	0.315	0.018	-0.031	0.02	0.068
		(0.449)	(0.441)	(0.372)	(0.366)	(0.205)	(0.188)	(0.190)	(0.191)
	#out-of-labor force (OLF)	0.046	0.014	-0.045	-0.085	0.106	0.093	0.119	0.116
		(0.259)	(0.271)	(0.174)	(0.184)	(0.139)	(0.152)	(0.117)	(0.099)
	mother: employment [Base: *full-time (FT)*]								
	# part-time (PT)	-0.019	-0.013	-0.073	-0.062	0.053	0.006	-0.001	0.013
		(0.137)	(0.120)	(0.099)	(0.104)	(0.059)	(0.060)	(0.041)	(0.039)
	# unemployed (UE)	0.167	0.267	0.139	0.252	-0.052	-0.065	0.174*	0.154
		(0.292)	(0.240)	(0.482)	(0.487)	(0.132)	(0.136)	(0.105)	(0.103)
	# out-of-labor force (OLF)	-0.158	-0.165	0.026	0.097	0.024	-0.032	-0.033	-0.021
		(0.122)	(0.150)	(0.122)	(0.116)	(0.072)	(0.080)	(0.071)	(0.060)
	Constant	0.655	0.754	3.434**	3.435*	3.114***	2.575**	2.320***	2.472***
		(2.604)	(2.969)	(1.538)	(1.785)	(0.996)	(1.097)	(0.619)	(0.626)
	Year FE	yes	yes	yes	yes	yes	yes	yes	yes
	Federal States FE	yes	yes	yes	yes	yes	yes	yes	yes
	School FE	no	yes	no	yes	no	yes	no	yes
	Observations	346	346	608	608	2168	2168	3093	3093
	R2	0.180***	0.242***	0.131***	0.162***	0.113***	0.174***	0.129***	0.206***
		(0.054)	(0.057)	(0.033)	(0.033)	(0.025)	(0.032)	(0.022)	(0.023)
	R2-adjusted	0.121**	0.172***	0.097***	0.114***	0.103***	0.147***	0.122***	0.184***
		(0.058)	(0.062)	(0.034)	(0.035)	(0.025)	(0.032)	(0.022)	(0.023)

Notes: 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 Appendix A.4. The dependent variable is *stdpvread3*, 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-***MT** period (2003-2012) (Footnote 42).
Columns (1) to (4) showing the results for Control Group (**C**) provide first-step regressions for the *Before-reform period (2003-2006)* in columns (1-2) and *After-reform period (2009-2012)* in columns (3-4). Columns (5) to (8) provide first-step regression results for Treatment Group (**T**) with *Before-reform period (2003-2006)* results in columns (5-6) and *After-reform period (2009-2012)* results in columns (7-8). Even columns in addition to their odd predecessors additionally take into account school fixed effects (SFE). 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.3.2). 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.3.1 on their computation) and provided in parentheses with $***$ denoting significance at the 1%, $**$ at the 5%, $*$ at the 10% and $+$ at the 15% level. *Source:* Author's calculations base on PISA-I-data 2003, 2006, 2009, 2012.

Table A.7: Main Results for *Model Base-**MT***: 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)	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CONTROL: **Individual Characteristics (IC)**									
i)	female	-0.744***	-0.662***	-0.553***	-0.536***	-0.509***	-0.461***	-0.498***	-0.458***
		(0.132)	(0.116)	(0.066)	(0.065)	(0.046)	(0.049)	(0.051)	(0.050)
	age in years	-0.064	-0.11	-0.289***	-0.318***	-0.207***	-0.210***	-0.214***	-0.237***
		(0.093)	(0.112)	(0.067)	(0.071)	(0.061)	(0.045)	(0.033)	(0.037)
ii)	migration background	-0.111	-0.084	-0.107	-0.135	-0.186**	-0.114	-0.152**	-0.141**
		(0.206)	(0.185)	(0.132)	(0.135)	(0.084)	(0.076)	(0.062)	(0.058)
	NO German spoken at home	-0.46	-0.373	-0.166	-0.189	-0.088	-0.091	-0.183**	-0.199**
		(0.397)	(0.337)	(0.167)	(0.171)	(0.116)	(0.110)	(0.088)	(0.082)
CONTROL: **Parental Characteristics (PC)**									
iii)	Parental Education: [Base: *ISCED-level (3-4)*]								
	# at most lower sec. educ. (ISCED-level (1-2))	-0.486*	-0.454*	0.018	0.126	-0.241***	-0.170**	-0.173**	-0.107
		(0.270)	(0.275)	(0.159)	(0.147)	(0.090)	(0.078)	(0.072)	(0.068)
	# tertiary educ. (ISCED-level (5-6))	-0.186	-0.023	0.061	0.067	0.045	0.014	0.021	0.017
		(0.121)	(0.101)	(0.133)	(0.139)	(0.052)	(0.047)	(0.038)	(0.040)
CONTROL: **Socio-Economic Status (SES)**									
iv)	No. of books in household [Base: *101-500*]								
	# max 10 books	0.318	0.342	-0.511	-0.435	-0.433**	-0.387**	-0.335***	-0.295***
		(0.271)	(0.243)	(0.320)	(0.303)	(0.192)	(0.154)	(0.124)	(0.112)
	# 11-100 books	-0.156*	-0.12	-0.123	-0.086	-0.271***	-0.256***	-0.163***	-0.140***
		(0.085)	(0.078)	(0.131)	(0.134)	(0.065)	(0.063)	(0.045)	(0.047)
	# more than 500 books	0.264*	0.234	0.184	0.191	0.041	0.06	0.116***	0.117***
		(0.135)	(0.149)	(0.122)	(0.122)	(0.051)	(0.055)	(0.038)	(0.039)
v)	highest ISEI-level of parental jobs	0.007*	0.007	0.001	0.001	0.001	0.001	0.004***	0.004***
		(0.004)	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
CONTROL: **Family Characteristics (FC)**									
vi)	family structure [Base: *No*]								
	single parent household	-0.024	0.058	0.242**	0.201*	0.031	0.043	0.092	0.091
		(0.188)	(0.176)	(0.115)	(0.110)	(0.057)	(0.053)	(0.056)	(0.058)
vii)	father: employment [Base: *full-time (FT)*]								
	# part-time (PT)	-0.278	-0.238	-0.382*	-0.421*	-0.053	-0.024	-0.162***	-0.127**
		(0.273)	(0.278)	(0.206)	(0.215)	(0.099)	(0.090)	(0.058)	(0.060)
	# unemployed (UE)	0.06	0.075	0.133	0.097	-0.283	-0.23	0.037	0.061
		(0.361)	(0.353)	(0.411)	(0.424)	(0.178)	(0.142)	(0.148)	(0.147)
	#out-of-labor force (OLF)	0.011	-0.044	-0.134	-0.184	-0.067	-0.036	-0.028	-0.026
		(0.332)	(0.308)	(0.155)	(0.149)	(0.157)	(0.141)	(0.121)	(0.119)
	mother: employment [Base: *full-time (FT)*]								
	# part-time (PT)	0.11	0.096	0.036	0.018	0.044	0.002	0.061	0.063
		(0.096)	(0.068)	(0.124)	(0.124)	(0.068)	(0.068)	(0.049)	(0.049)
	# unemployed (UE)	0.126	0.212	0.167	0.132	-0.047	-0.019	0.267***	0.271***
		(0.197)	(0.161)	(0.577)	(0.594)	(0.080)	(0.079)	(0.101)	(0.100)
	# out-of-labor force (OLF)	-0.04	-0.072	0.189*	0.197	0.06	-0.021	0.09	0.088
		(0.167)	(0.154)	(0.109)	(0.119)	(0.083)	(0.087)	(0.078)	(0.075)
	Constant	1.156	1.824	4.585***	5.001***	3.658***	3.650***	3.527***	3.859***
		(1.436)	(1.755)	(1.058)	(1.108)	(0.964)	(0.692)	(0.540)	(0.583)
	Year FE	yes	yes	yes	yes	yes	yes	yes	yes
	Federal States FE	yes	yes	yes	yes	yes	yes	yes	yes
	School FE	no	yes	no	yes	no	yes	no	yes
	Observations	346	346	608	608	2168	2168	3093	3093
	R2	0.300***	0.353***	0.161***	0.190***	0.158***	0.257***	0.168***	0.232***
		(0.059)	(0.060)	(0.039)	(0.039)	(0.022)	(0.032)	(0.025)	(0.028)
	R2-adjusted	0.250***	0.294***	0.128***	0.143***	0.148***	0.233***	0.162***	0.211***
		(0.063)	(0.065)	(0.041)	(0.041)	(0.022)	(0.033)	(0.025)	(0.029)

Notes: 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 Appendix A.4. The dependent variable is *stdpvmath*3, 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-***MT** period (2003-2012) (Footnote 42).
Columns (1) to (4) showing the results for Control Group (**C**) provide first-step regressions for the *Before-reform period (2003-2006)* in columns (1-2) and *After-reform period (2009-2012)* in columns (3-4). Columns (5) to (8) provide first-step regression results for Treatment Group (**T**) with *Before-reform period (2003-2006)* results in columns (5-6) and *After-reform period (2009-2012)* results in columns (7-8). Even columns in addition to their odd predecessors additionally take into account school fixed effects (SFE). 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.3.2). 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.3.1 on their computation) and provided in parentheses with $***$ denoting significance at the 1%, $**$ at the 5%, $*$ at the 10% and $+$ at the 15% level.
Source: Author's calculations base on PISA-I-data 2003, 2006, 2009, 2012.

Table A.8: Main Results for *Model Base-**MT***: 1^{st} step to derive IEOp measure for *science scores*

DEPENDENT VARIABLE:		**Control Group (C)**				**Treatment Group (T)**			
	SCIENCES Test Scores in PISA (STDPVSCIE3)	Before (2003-2006) (1)	(2)	After (2009-2012) (3)	(4)	Before (2003-2006) (5)	(6)	After (2009-2012) (7)	(8)
CONTROL: **Individual Characteristics (IC)**									
i)	female	-0.592***	-0.509***	-0.373***	-0.337***	-0.402***	-0.354***	-0.318***	-0.297***
		(0.123)	(0.113)	(0.067)	(0.064)	(0.039)	(0.039)	(0.042)	(0.043)
	age in years	-0.046	-0.093	-0.213***	-0.252***	-0.178***	-0.154**	-0.158***	-0.182***
		(0.105)	(0.118)	(0.067)	(0.078)	(0.066)	(0.066)	(0.047)	(0.052)
ii)	migration background	-0.323*	-0.297*	-0.293***	-0.295***	-0.132	-0.075	-0.194***	-0.170***
		(0.171)	(0.163)	(0.109)	(0.097)	(0.091)	(0.086)	(0.061)	(0.059)
	NO German spoken at home	-0.464	-0.347	-0.152	-0.182	-0.217**	-0.240**	-0.160**	-0.196***
		(0.447)	(0.403)	(0.208)	(0.214)	(0.101)	(0.097)	(0.070)	(0.068)
CONTROL: **Parental Characteristics (PC)**									
iii)	Parental Education: [Base: *ISCED-level (3-4)*]								
	# at most lower sec. educ. (ISCED-level (1-2))	-0.594**	-0.529**	-0.014	0.069	-0.261***	-0.250***	-0.122	-0.036
		(0.251)	(0.243)	(0.165)	(0.175)	(0.091)	(0.093)	(0.089)	(0.063)
	# tertiary educ. (ISCED-level (5-6))	-0.108	-0.153	0.077	0.119	0.041	0.032	0.021	0.025
		(0.149)	(0.123)	(0.112)	(0.123)	(0.062)	(0.062)	(0.044)	(0.043)
CONTROL: **Socio-Economic Status (SES)**									
iv)	No. of books in household [Base: *101-500*]								
	# max 10 books	0.099	0.101	-0.509	-0.425	-0.236	-0.272*	-0.577***	-0.506***
		(0.389)	(0.413)	(0.318)	(0.292)	(0.172)	(0.153)	(0.087)	(0.073)
	# 11-100 books	-0.166	-0.122	-0.147	-0.12	-0.295***	-0.282***	-0.241***	-0.209***
		(0.106)	(0.106)	(0.119)	(0.115)	(0.067)	(0.068)	(0.040)	(0.040)
	# more than 500 books	0.189	0.175	0.116	0.125	0.132**	0.142**	0.160***	0.159***
		(0.132)	(0.147)	(0.117)	(0.112)	(0.059)	(0.057)	(0.043)	(0.041)
v)	highest ISEI-level of parental jobs	0.009**	0.009**	0.002	0.002	0.002	0.003	0.004**	0.003*
		(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)
CONTROL: **Family Characteristics (FC)**									
vi)	family structure [Base: *No*]								
	single parent household	-0.042	0.051	0.271***	0.214***	0.026	0.022	0.118*	0.1
		(0.264)	(0.241)	(0.081)	(0.082)	(0.079)	(0.080)	(0.071)	(0.069)
vii)	father: employment [Base: *full-time (FT)*]								
	# part-time (PT)	-0.211	-0.162	-0.203	-0.23	-0.099	-0.065	-0.172**	-0.151*
		(0.198)	(0.207)	(0.170)	(0.191)	(0.130)	(0.120)	(0.074)	(0.081)
	# unemployed (UE)	0.183	0.192	0.215	0.16	-0.025	-0.032	0.058	0.064
		(0.441)	(0.426)	(0.357)	(0.370)	(0.166)	(0.149)	(0.198)	(0.192)
	#out-of-labor force (OLF)	0.093	0.036	0.068	-0.031	0.012	0.013	0.064	0.072
		(0.287)	(0.290)	(0.230)	(0.244)	(0.125)	(0.127)	(0.119)	(0.110)
	mother: employment [Base: *full-time (FT)*]								
	# part-time (PT)	-0.08	-0.083	-0.029	-0.067	0.04	-0.002	0.025	0.031
		(0.112)	(0.097)	(0.095)	(0.097)	(0.060)	(0.064)	(0.040)	(0.040)
	# unemployed (UE)	0.091	0.184	0.107	0.087	-0.042	-0.014	0.244**	0.217**
		(0.152)	(0.156)	(0.294)	(0.302)	(0.106)	(0.108)	(0.101)	(0.110)
	# out-of-labor force (OLF)	-0.189*	-0.218	0.09	0.081	0.013	-0.052	0.019	0.016
		(0.110)	(0.139)	(0.096)	(0.108)	(0.060)	(0.067)	(0.067)	(0.064)
	Constant	0.919	1.593	3.440***	4.020***	3.080***	2.685**	2.607***	2.986***
		(1.770)	(1.918)	(1.053)	(1.216)	(1.078)	(1.060)	(0.701)	(0.783)
	Year FE	yes	yes	yes	yes	yes	yes	yes	yes
	Federal States FE	yes	yes	yes	yes	yes	yes	yes	yes
	School FE	no	yes	no	yes	no	yes	no	yes
	Observations	346	346	608	608	2168	2168	3093	3093
	R2	0.295***	0.363***	0.129***	0.173***	0.133***	0.203***	0.130***	0.202***
		(0.055)	(0.052)	(0.037)	(0.047)	(0.020)	(0.024)	(0.018)	(0.024)
	R2-adjusted	0.244***	0.304***	0.095**	0.126**	0.123***	0.177***	0.123***	0.180***
		(0.059)	(0.057)	(0.039)	(0.050)	(0.020)	(0.025)	(0.018)	(0.025)

Notes: 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 Appendix A.4. 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-***MT** period (2003-2012) (Footnote 42).
Columns (1) to (4) showing the results for Control Group (**C**) provide first-step regressions for the *Before-reform period (2003-2006)* in columns (1-2) and *After-reform period (2009-2012)* in columns (3-4). Columns (5) to (8) provide first-step regression results for Treatment Group (**T**) with *Before-reform period (2003-2006)* results in columns (5-6) and *After-reform period (2009-2012)* results in columns (7-8). Even columns in addition to their odd predecessors additionally take into account school fixed effects (SFE). 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.3.2). 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.3.1 on their computation) and provided in parentheses with $***$ denoting significance at the 1%, $**$ at the 5%, $*$ at the 10% and + at the 15% level.
Source: Author's calculations base on PISA-I-data 2003, 2006, 2009, 2012.

Table A.9: Results for *Model Base-**ST*** vs. ***-MT***: 1^{st} step to derive IEOp for *mathematics scores*

		Model Base-**ST** (**2003-2006** vs. **2009**)				Model Base-**MT** (**2003-2006** vs. **2009-2012**)			
DEPENDENT VARIABLE:		Control Group (**C**)		Treatment Group (**T**)		Control Group (**C**)		Treatment Group (**T**)	
	MATHEMATICS Test Scores in PISA (STDPVMATH3)	Before (1)	After (2)	Before (3)	After (4)	Before (5)	After (6)	Before (7)	After (8)
CONTROL: **Individual Characteristics (IC)**									
i)	female	-0.662***	-0.676***	-0.461***	-0.478***	-0.662***	-0.536***	-0.461***	-0.458***
		(0.116)	(0.118)	(0.049)	(0.059)	(0.116)	(0.065)	(0.049)	(0.050)
	age in years	-0.11	-0.358***	-0.210***	-0.264***	-0.11	-0.318***	-0.210***	-0.237***
		(0.112)	(0.113)	(0.045)	(0.070)	(0.112)	(0.071)	(0.045)	(0.037)
ii)	migration background	-0.084	0.097	-0.114	-0.177**	-0.084	-0.135	-0.114	-0.141**
		(0.185)	(0.204)	(0.076)	(0.089)	(0.185)	(0.135)	(0.076)	(0.058)
	NO German spoken at home	-0.373	-0.443	-0.091	-0.287**	-0.373	-0.189	-0.091	-0.199**
		(0.337)	(0.321)	(0.110)	(0.140)	(0.337)	(0.171)	(0.110)	(0.082)
CONTROL: **Parental Characteristics (PC)**									
iii)	Parental Education: [Base: *ISCED-level (3-4)*]								
	# max. lower sec. educ. (ISCED-level (1-2))	-0.454*	0.466	-0.170**	-0.320**	-0.454*	0.126	-0.170**	-0.107
		(0.275)	(0.291)	(0.078)	(0.133)	(0.275)	(0.147)	(0.078)	(0.068)
	# tertiary educ. (ISCED-level (5-6))	-0.023	0.049	0.014	0.008	-0.023	0.067	0.014	0.017
		(0.101)	(0.211)	(0.047)	(0.055)	(0.101)	(0.139)	(0.047)	(0.040)
CONTROL: **Socio-Economic Status (SES)**									
iv)	No. of books in household [Base: *101-500*]								
	# max 10 books	0.342	0.014	-0.387**	-0.203	0.342	-0.435	-0.387**	-0.295***
		(0.243)	(0.258)	(0.154)	(0.176)	(0.243)	(0.303)	(0.154)	(0.112)
	# 11-100 books	-0.12	0.055	-0.256***	-0.184***	-0.12	-0.086	-0.256***	-0.140***
		(0.078)	(0.179)	(0.063)	(0.059)	(0.078)	(0.134)	(0.063)	(0.047)
	# more than 500 books	0.234	0.278	0.06	0.102	0.234	0.191	0.06	0.117***
		(0.149)	(0.216)	(0.055)	(0.075)	(0.149)	(0.122)	(0.055)	(0.039)
v)	highest ISEI-level of parental jobs	0.007	0.005	0.001	0.004	0.007	0.001	0.001	0.004***
		(0.005)	(0.004)	(0.002)	(0.002)	(0.005)	(0.002)	(0.002)	(0.001)
CONTROL: **Family Characteristics (FC)**									
vi)	family structure [Base: *No*]								
	single parent household	0.058	0.058	0.043	0.211***	0.058	0.201*	0.043	0.091
		(0.176)	(0.156)	(0.053)	(0.078)	(0.176)	(0.110)	(0.053)	(0.058)
vii)	father: employment [Base: *full-time (FT)*]								
	# part-time (PT)	-0.238	-0.519*	-0.024	-0.037	-0.238	-0.421*	-0.024	-0.127**
		(0.278)	(0.285)	(0.090)	(0.089)	(0.278)	(0.215)	(0.090)	(0.060)
	# unemployed (UE)	0.075	-0.076	-0.23	0.023	0.075	0.097	-0.23	0.061
		(0.353)	(0.461)	(0.142)	(0.239)	(0.353)	(0.424)	(0.142)	(0.147)
	#out-of-labor force (OLF)	-0.044	-0.175	-0.036	-0.103	-0.044	-0.184	-0.036	-0.026
		(0.308)	(0.214)	(0.141)	(0.158)	(0.308)	(0.149)	(0.141)	(0.119)
	mother: employment [Base: *full-time (FT)*]								
	# part-time (PT)	0.096	0.219	0.002	0.001	0.096	0.018	0.002	0.063
		(0.068)	(0.160)	(0.068)	(0.079)	(0.068)	(0.124)	(0.068)	(0.049)
	# unemployed (UE)	0.212	-0.076	-0.019	0.308*	0.212	0.132	-0.019	0.271***
		(0.161)	(0.598)	(0.079)	(0.160)	(0.161)	(0.594)	(0.079)	(0.100)
	# out-of-labor force (OLF)	-0.072	0.247*	-0.021	0.057	-0.072	0.197	-0.021	0.088
		(0.154)	(0.137)	(0.087)	(0.086)	(0.154)	(0.119)	(0.087)	(0.075)
	Constant	1.824	5.357***	3.650***	4.380***	1.824	5.001***	3.650***	3.859***
		(1.755)	(1.732)	(0.692)	(1.071)	(1.755)	(1.108)	(0.692)	(0.583)
	Year FE	yes	yes	yes	yes	yes	yes	yes	yes
	Federal States FE	yes	yes	yes	yes	yes	yes	yes	yes
	School FE	yes	yes	yes	yes	yes	yes	yes	yes
	Observations	346	308	2168	1467	346	608	2168	3093
	R2	0.353***	0.270***	0.257***	0.223***	0.353***	0.190***	0.257***	0.232***
		(0.060)	(0.073)	(0.032)	(0.041)	(0.060)	(0.039)	(0.032)	(0.028)
	R2-adjusted	0.294***	0.199**	0.233***	0.196***	0.294***	0.143***	0.233***	0.211***
		(0.065)	(0.080)	(0.033)	(0.042)	(0.065)	(0.041)	(0.033)	(0.029)

Notes: This table shows the first stage OLS regressions to derive the R^2 as IEOp measure for conducting the DiD estimation approach as shown in the second sub-panel of Table 4. 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*-**ST** (2003-2009) and -**MT** time period (2003-2012) (Footnote 42).
Columns (1-2) show the first-step regression results for Control Group (**C**), columns (3-4) for Treatment Group (**T**), both with respect to (**ST**) model (2003-2006 vs. 2009). Columns (5-6) provide first-step regression results for **C**, columns (7-8) for **T**, both with respect to the (**MT**) model (2003-2006 vs. 2009-2012). All regressions take into account *year*, *federal state* and *school* fixed effects. Even columns show results for the *After-reform*, odd ones for the *Before-reform* time period. 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.3.1 on their computation) and provided in parentheses with $***$ denoting significance at the 1%, $**$ at the 5%, $*$ at the 10% and $+$ at the 15% level.
Source: Author's calculations base on PISA-I-data 2003, 2006, 2009.

Table A.10: Results for *Model Base-**ST*** - **T** vs. **C1/C2**: 1^{st} step to derive IEOp for *maths*

		Model Base-**ST** (**2003-2006** vs. **2009**)				Model Base-**ST** (**2003-2006** vs. **2009**)			
	DEPENDENT VARIABLE:	Control Group (**C1**)		Treatment Group (**T**)		Control Group (**C2**)		Treatment Group (**T**)	
	MATHEMATICS Test Scores in PISA (STDPVMATH3)	Before (1)	After (2)	Before (3)	After (4)	Before (5)	After (6)	Before (7)	After (8)
CONTROL: **Individual Characteristics (IC)**									
i)	female	-0.469***	-0.507***	-0.461***	-0.478***	-0.439***	-0.521***	-0.461***	-0.478***
		(0.050)	(0.059)	(0.049)	(0.059)	(0.044)	(0.054)	(0.049)	(0.059)
	age in years	-0.216***	-0.317***	-0.210***	-0.264***	-0.227***	-0.298***	-0.210***	-0.264***
		(0.052)	(0.054)	(0.045)	(0.070)	(0.048)	(0.057)	(0.045)	(0.070)
ii)	migration background	-0.049	-0.094	-0.114	-0.177**	-0.019	-0.067	-0.114	-0.177**
		(0.078)	(0.074)	(0.076)	(0.089)	(0.062)	(0.052)	(0.076)	(0.089)
	NO German spoken at home	-0.215	-0.309***	-0.091	-0.287**	-0.211*	-0.285**	-0.091	-0.287**
		(0.136)	(0.118)	(0.110)	(0.140)	(0.115)	(0.120)	(0.110)	(0.140)
CONTROL: **Parental Characteristics (PC)**									
iii)	Parental Education: [Base: *ISCED-level (3-4)*]								
	# max. lower sec. educ. (ISCED-level (1-2))	-0.214*	0.176	-0.170**	-0.320**	-0.185	0.055	-0.170**	-0.320**
		(0.114)	(0.189)	(0.078)	(0.133)	(0.119)	(0.182)	(0.078)	(0.133)
	# tertiary educ. (ISCED-level (5-6))	0.008	0.034	0.014	0.008	0.058	0.034	0.014	0.008
		(0.042)	(0.088)	(0.047)	(0.055)	(0.045)	(0.075)	(0.047)	(0.055)
CONTROL: **Socio-Economic Status (SES)**									
iv)	No. of books in household [Base: *101-500*]								
	# max 10 books	-0.065	-0.582***	-0.387**	-0.203	-0.038	-0.284	-0.387**	-0.203
		(0.134)	(0.183)	(0.154)	(0.176)	(0.111)	(0.279)	(0.154)	(0.176)
	# 11-100 books	-0.220***	-0.164**	-0.256***	-0.184***	-0.170***	-0.170***	-0.256***	-0.184***
		(0.045)	(0.077)	(0.063)	(0.059)	(0.055)	(0.062)	(0.063)	(0.059)
	# more than 500 books	0.066	0.172	0.06	0.102	0.100**	0.119	0.06	0.102
		(0.053)	(0.107)	(0.055)	(0.075)	(0.045)	(0.076)	(0.055)	(0.075)
v)	highest ISEI-level of parental jobs	0.003	0.003**	0.001	0.004	0.002	0.004**	0.001	0.004
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
CONTROL: **Family Characteristics (FC)**									
vi)	family structure [Base: *No*]								
	single parent household	-0.019	0.105	0.043	0.211***	-0.06	0.11	0.043	0.211***
		(0.068)	(0.097)	(0.053)	(0.078)	(0.058)	(0.075)	(0.053)	(0.078)
vii)	father: employment [Base: *full-time (FT)*]								
	# part-time (PT)	-0.175	-0.183	-0.024	-0.037	-0.145	-0.208	-0.024	-0.037
		(0.114)	(0.169)	(0.090)	(0.089)	(0.112)	(0.132)	(0.090)	(0.089)
	# unemployed (UE)	-0.076	0.147	-0.23	0.023	-0.024	0.305	-0.23	0.023
		(0.183)	(0.214)	(0.142)	(0.239)	(0.149)	(0.241)	(0.142)	(0.239)
	#out-of-labor force (OLF)	-0.153	-0.052	-0.036	-0.103	-0.105	0.021	-0.036	-0.103
		(0.142)	(0.175)	(0.141)	(0.158)	(0.131)	(0.159)	(0.141)	(0.158)
	mother: employment [Base: *full-time (FT)*]								
	# part-time (PT)	0.036	0.02	0.002	0.001	0.021	0.039	0.002	0.001
		(0.043)	(0.077)	(0.068)	(0.079)	(0.040)	(0.076)	(0.068)	(0.079)
	# unemployed (UE)	-0.03	0.013	-0.019	0.308*	-0.039	-0.014	-0.019	0.308*
		(0.140)	(0.224)	(0.079)	(0.160)	(0.116)	(0.181)	(0.079)	(0.160)
	# out-of-labor force (OLF)	-0.006	0.054	-0.021	0.057	0.024	0.076	-0.021	0.057
		(0.063)	(0.087)	(0.087)	(0.086)	(0.060)	(0.090)	(0.087)	(0.086)
	Constant	3.494***	4.974***	3.650***	4.380***	3.627***	4.625***	3.650***	4.380***
		(0.870)	(0.818)	(0.692)	(1.071)	(0.764)	(0.863)	(0.692)	(1.071)
	Year FE	yes	yes	yes	yes	yes	yes	yes	yes
	Federal States FE	yes	yes	yes	yes	yes	yes	yes	yes
	School FE	yes	yes	yes	yes	yes	yes	yes	yes
	Observations	1854	1159	2168	1467	2326	1458	2168	1467
	R2	0.216***	0.233***	0.257***	0.223***	0.243***	0.222***	0.257***	0.223***
		(0.029)	(0.033)	(0.032)	(0.041)	(0.031)	(0.027)	(0.032)	(0.041)
	R2-adjusted	0.191***	0.203***	0.233***	0.196***	0.220***	0.195***	0.233***	0.196***
		(0.030)	(0.035)	(0.033)	(0.042)	(0.032)	(0.028)	(0.033)	(0.042)

Notes: This table shows the first stage OLS regressions to derive the R^2 as IEOp measure for conducting the DiD estimation approach shown in the second sub-panel of Table 5 (Appendix A.5). 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-***ST** (2003-2009) (Footnote 42).
Columns (1-2) show the first-step regression results for Control Group (**C1**), columns (3-4) for Treatment Group (**T**), both with respect to (**ST**) model (2003-2006 vs. 2009). Columns (5-6) provide first-step regression results for **C2**, columns (7-8) for **T**, both with respect to the (**ST**) model. All regressions take into account *year*, *federal state* and *school* fixed effects. Even columns show results for the *After-reform*, odd ones for the *Before-reform* time period. 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.3.1 on their computation) and provided in parentheses with $***$ denoting significance at the 1%, $**$ at the 5%, $*$ at the 10% and $+$ at the 15% level. *Source:* Author's calculations base on PISA-I-data 2003, 2006, 2009.

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

Subject	*Model Base-**MT** (2003-2012)* - **T** vs. **C** — (**Figure A.4**)					
	with Bundesland-FE			**with School-FE**		
Reading	**C**	**T**	**Δ *(T - C)***	**C**	**T**	**Δ *(T - C)***
Before (2003-2006)	0.180	0.113	-0.067	0.242	0.173	-0.068
	(0.054)	(0.025)	(0.060)	(0.057)	(0.032)	(0.065)
After (2009-2012)	0.131	0.129	-0.002	0.162	0.206	0.044
	(0.033)	(0.022)	(0.040)	(0.033)	(0.023)	(0.040)
Change in R2	-0.049	0.016	**0.065**	-0.079	0.033	**0.112**
	(0.063)	(0.033)	(0.071)	(0.066)	(0.039)	(0.077)
Mathematics	**C**	**T**	**Δ *(T - C)***	**C**	**T**	**Δ *(T - C)***
Before (2003-2006)	0.300	0.158	-0.142	0.353	0.257	-0.097
	(0.059)	(0.022)	(0.063)	(0.060)	(0.032)	(0.068)
After (2009-2012)	0.161	0.168	0.007	0.190	0.232	0.042
	(0.039)	(0.025)	(0.046)	(0.039)	(0.028)	(0.048)
Change in R2	-0.139	0.010	**0.150**	-0.163	-0.025	**0.139**
	(0.071)	(0.033)	(0.078)	(0.071)	(0.043)	*(0.083)*
Science	**C**	**T**	**Δ *(T - C)***	**C**	**T**	**Δ *(T - C)***
Before (2003-2006)	0.295	0.133	-0.161	0.363	0.203	-0.160
	(0.055)	(0.020)	(0.058)	(0.052)	(0.024)	(0.058)
After (2009-2012)	0.129	0.130	0.001	0.173	0.202	0.028
	(0.037)	(0.018)	(0.041)	(0.047)	(0.024)	(0.053)
Change in R2	-0.166	-0.003	**0.162**	-0.189	-0.001	**0.188**
	(0.066)	(0.027)	(0.071)	(0.071)	(0.034)	(0.078)

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.3.1, 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/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 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) - mother/father unemployed (UE) - mother/father out of labor force (OLF)
Compare: Table A.6, Table A.7 and Table A.8 for details on first-step regression for **T**/**C** according to Equation (8).
Source: Author's calculations based on PISA-I-data 2003, 2006, 2009, 2012.

Table A.12: Robustness Checks: Placebo Tests (2003-2006) — Mathematics

Subject	*Model Base-**MT** with school FE (2003-2012)* - **T** vs. **C/C1/C2** — (**Figure A.4**)					
	with R^2 measure			**with $R^2_{adjusted}$ measure**		
Mathematics	**C**	**T**	**Δ *(T - C)***	**C**	**T**	**Δ *(T - C)***
Before (2003)	0.353	0.225	-0.128	0.249	0.189	-0.059
	(0.109)	(0.049)	(0.119)	(0.127)	(0.051)	(0.137)
After (2006)	0.362	0.278	-0.084	0.267	0.250	-0.017
	(0.054)	(0.048)	(0.072)	(0.062)	(0.049)	(0.079)
Change in R2	0.009	0.053	**0.044**	0.018	0.060	**0.042**
	(0.122)	(0.068)	(0.139)	(0.141)	(0.071)	(0.158)
Mathematics	**C1**	**T**	**Δ *(T - C1)***	**C1**	**T**	**Δ *(T - C1)***
Before (2003)	0.200	0.225	0.025	0.162	0.189	0.027
	(0.050)	(0.049)	(0.070)	(0.053)	(0.051)	(0.073)
After (2006)	0.245	0.278	0.034	0.214	0.250	0.036
	(0.035)	(0.048)	(0.059)	(0.037)	(0.049)	(0.062)
Change in R2	0.044	0.053	**0.009**	0.051	0.060	**0.009**
	(0.061)	(0.068)	(0.092)	(0.064)	(0.071)	(0.096)
Mathematics	**C2**	**T**	**Δ *(T - C2)***	**C2**	**T**	**Δ *(T - C2)***
Before (2003)	0.219	0.225	0.007	0.185	0.189	0.004
	(0.042)	(0.049)	(0.064)	(0.043)	(0.051)	(0.067)
After (2006)	0.268	0.278	0.010	0.241	0.250	0.009
	(0.043)	(0.048)	(0.064)	(0.045)	(0.049)	(0.067)
Change in R2	0.049	0.053	**0.004**	0.056	0.060	**0.005**
	(0.060)	(0.068)	(0.091)	(0.063)	(0.071)	(0.095)

Notes: Table entries are R^2 measures of IEOp (Equation (7)). Due to space constraints, only **Placebo Test** results for mathematics test scores are shown. However, as shown in the main results, these scores tend to be good proxies between the reading and science scores. Robust standard errors are in parentheses and were calculated using replication weights following the method as explained in Appendix A.3.1, 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/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) - mother/father unemployed (UE) - mother/father out of labor force (OLF)
Compare: Due to space constraints first-step regressions for **T** vs. **C/C1/C2** have been omitted, but they remain available upon request from the author. The same applies to Placebo Test results for the other two testing domains, science and reading.
Source: Author's calculations based on PISA-I-data 2003, 2006, 2009.

Table A.13: Difference-in-Difference Results: Overview - Model Base-**MT** - Control Group C

(1) Outcome	(2) Treatment	(3) Control	(4) Model	(5) Control set	(6) R^2-DD-BL	(7) $R^2adjusted$-DD_BL	(8) R^2-DD_SF	(9) $R^2adjusted$-DD_SF
read	T	C	Base-MT	1	0.0486019	0.0418389	0.1026374	0.0986496
read	T	C	Base-MT	2	0.0508262	0.0413408	0.10818	0.1020977
read	T	C	Base-MT	3	0.0492137	0.0362789	0.1089052	0.1000562
read	T	C	Base-MT	4	0.0573307	0.0453275	0.116741	0.1091316
read	T	C	Base-MT	5	0.0347353	0.0203472	0.1036062	0.0938706
read	T	C	Base-MT	6	0.0734411	0.053641	0.1333018	0.1196875
read	T1	C	Base-MT	1	0.0552201	0.0485851	0.1045078	0.1004037
read	T1	C	Base-MT	2	0.053747	0.0443369	0.1076478	0.1013548
read	T1	C	Base-MT	3	0.0515521	0.0386309	0.1092806	0.1001985
read	T1	C	Base-MT	4	0.0594724	0.047456	0.1171495	0.1092804
read	T1	C	Base-MT	5	0.0391092	0.0246972	0.1052288	0.0952423
read	T1	C	Base-MT	6	0.0775819	0.0576589	0.134443	0.1204617
read	T2	C	Base-MT	1	0.0294624	0.0224113	0.075922	0.070706
read	T2	C	Base-MT	2	0.0205508	0.0105245	0.0749819	0.06732
read	T2	C	Base-MT	3	0.0195176	0.0057722	0.0762089	0.065553
read	T2	C	Base-MT	4	0.026705	0.0137465	0.0832847	0.0737119
read	T2	C	Base-MT	5	0.0044039	-0.011066	0.071847	0.0600673
read	T2	C	Base-MT	6	0.0440509	0.0226172	0.1031965	0.087029
math	T	C	Base-MT	1	0.1374442	0.1326837	0.1173572	0.1150981
math	T	C	Base-MT	2	0.1522048	0.1457725	0.1318355	0.1284723
math	T	C	Base-MT	3	0.1525054	0.1440416	0.1334687	0.1284267
math	T	C	Base-MT	4	0.1627268	0.155587	0.1438371	0.1403698
math	T	C	Base-MT	5	0.1517936	0.1433778	0.1421455	0.1376176
math	T	C	Base-MT	6	0.1613632	0.1488317	0.1465036	0.1384381
math	T1	C	Base-MT	1	0.1485623	0.1439196	0.126169	0.1239403
math	T1	C	Base-MT	2	0.1587099	0.1523483	0.1377001	0.1342778
math	T1	C	Base-MT	3	0.1573647	0.148909	0.1397943	0.1346712
math	T1	C	Base-MT	4	0.1678983	0.1607435	0.1502135	0.1466463
math	T1	C	Base-MT	5	0.1596974	0.1512595	0.1496798	0.1450618
math	T1	C	Base-MT	6	0.168932	0.1562866	0.1531902	0.1449399
math	T2	C	Base-MT	1	0.1225039	0.1174748	0.1161204	0.1132698
math	T2	C	Base-MT	2	0.1265655	0.1196356	0.1252417	0.1210203
math	T2	C	Base-MT	3	0.1239317	0.1146941	0.1247105	0.1185743
math	T2	C	Base-MT	4	0.1338416	0.1257928	0.1343879	0.1297064
math	T2	C	Base-MT	5	0.1232632	0.1138137	0.1334823	0.1276546
math	T2	C	Base-MT	6	0.1350299	0.1209538	0.1401597	0.1304406
science	T	C	Base-MT	1	0.1478232	0.1432816	0.1925849	0.1923788
science	T	C	Base-MT	2	0.1586137	0.1526844	0.1937428	0.1924779
science	T	C	Base-MT	3	0.1603698	0.1524458	0.188899	0.1859919
science	T	C	Base-MT	4	0.1743826	0.1680937	0.201756	0.2008021
science	T	C	Base-MT	5	0.1563525	0.1485939	0.1925775	0.1904328
science	T	C	Base-MT	6	0.1814693	0.1706351	0.2123226	0.2083999
science	T1	C	Base-MT	1	0.148146	0.1436911	0.1885749	0.1881179
science	T1	C	Base-MT	2	0.1568542	0.1509604	0.1887328	0.1871525
science	T1	C	Base-MT	3	0.156604	0.1486444	0.1836784	0.1803993
science	T1	C	Base-MT	4	0.1703102	0.1639529	0.1961935	0.1948287
science	T1	C	Base-MT	5	0.1541647	0.14632	0.1878349	0.1852712
science	T1	C	Base-MT	6	0.178593	0.1675509	0.2065061	0.2020138
science	T2	C	Base-MT	1	0.1289499	0.1241082	0.1688373	0.1674972
science	T2	C	Base-MT	2	0.1311111	0.1246503	0.1658215	0.1631425
science	T2	C	Base-MT	3	0.131187	0.122473	0.1598517	0.1552819
science	T2	C	Base-MT	4	0.1426166	0.1353887	0.1703004	0.167499
science	T2	C	Base-MT	5	0.1249573	0.1161332	0.1617619	0.1576607
science	T2	C	Base-MT	6	0.1522492	0.1398239	0.1838509	0.1775085

Notes: This table shows *T/T1/T2 vs. C* in **Model Base-MT** for all 3 test score domains and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$ (compare also Appendix A.4 and Section 4.1). Note that column (6) shows the DiD results with *federal states* fixed effects, (7) shows the same but using adjusted R^2 as IEOp measure. Column (8) shows the DiD results with *school* fixed effects and (9) using additionally adjusted R^2 IEOp measures.

Table A.14: Difference-in-Difference Results: Overview - Model Base-**ST** - Control Group C

(1) Outcome	(2) Treatment	(3) Control	(4) Model	(5) Control set	(6) R^2-DD-BL	(7) R^2*adjusted*-DD_BL	(8) R^2-DD_SF	(9) R^2*adjusted*-DD_SF
read	T	C	Base-ST	1	0.115189	0.114791	0.111681	0.111725
read	T	C	Base-ST	2	0.121095	0.121916	0.118177	0.119875
read	T	C	Base-ST	3	0.117737	0.12046	0.114672	0.11858
read	T	C	Base-ST	4	0.121297	0.125021	0.119677	0.124813
read	T	C	Base-ST	5	0.103633	0.10752	0.111402	0.117025
read	T	C	Base-ST	6	0.13003	0.139194	0.131452	0.143227
read	T1	C	Base-ST	1	0.10786	0.107649	0.104675	0.104539
read	T1	C	Base-ST	2	0.1121	0.113107	0.110076	0.111575
read	T1	C	Base-ST	3	0.107721	0.110617	0.106849	0.110579
read	T1	C	Base-ST	4	0.11114	0.114999	0.11136	0.116273
read	T1	C	Base-ST	5	0.095784	0.099826	0.104421	0.109862
read	T1	C	Base-ST	6	0.121562	0.130887	0.123486	0.13507
read	T2	C	Base-ST	1	0.074232	0.073239	0.095674	0.094884
read	T2	C	Base-ST	2	0.073427	0.073553	0.097302	0.098033
read	T2	C	Base-ST	3	0.072164	0.074112	0.093102	0.096023
read	T2	C	Base-ST	4	0.075347	0.07814	0.097446	0.101448
read	T2	C	Base-ST	5	0.056112	0.059006	0.089524	0.094007
read	T2	C	Base-ST	6	0.082682	0.090642	0.110492	0.121088
math	T	C	Base-ST	1	0.051357	0.050464	0.031523	0.030167
math	T	C	Base-ST	2	0.073721	0.073815	0.053254	0.053582
math	T	C	Base-ST	3	0.086096	0.087854	0.060405	0.062834
math	T	C	Base-ST	4	0.085608	0.088037	0.063293	0.066655
math	T	C	Base-ST	5	0.0792	0.082129	0.066116	0.070348
math	T	C	Base-ST	6	0.067613	0.072198	0.050883	0.057212
math	T1	C	Base-ST	1	0.042805	0.042077	0.024931	0.023404
math	T1	C	Base-ST	2	0.062891	0.063151	0.04483	0.044959
math	T1	C	Base-ST	3	0.073168	0.075077	0.051544	0.053782
math	T1	C	Base-ST	4	0.072468	0.075007	0.053327	0.056438
math	T1	C	Base-ST	5	0.070316	0.073397	0.058665	0.062723
math	T1	C	Base-ST	6	0.059527	0.0643	0.043525	0.049722
math	T2	C	Base-ST	1	0.013281	0.011798	0.037425	0.035791
math	T2	C	Base-ST	2	0.02857	0.027986	0.053992	0.053944
math	T2	C	Base-ST	3	0.039033	0.040013	0.057108	0.0591
math	T2	C	Base-ST	4	0.038951	0.040461	0.058802	0.061596
math	T2	C	Base-ST	5	0.032046	0.034009	0.061858	0.065531
math	T2	C	Base-ST	6	0.023221	0.026675	0.048596	0.054453
science	T	C	Base-ST	1	0.097082	0.096825	0.10027	0.100811
science	T	C	Base-ST	2	0.101857	0.102772	0.102155	0.104122
science	T	C	Base-ST	3	0.093722	0.096029	0.098357	0.102221
science	T	C	Base-ST	4	0.098091	0.10139	0.106631	0.111915
science	T	C	Base-ST	5	0.08564	0.089083	0.102506	0.108355
science	T	C	Base-ST	6	0.098285	0.105394	0.113678	0.124412
science	T1	C	Base-ST	1	0.089691	0.089613	0.093076	0.093426
science	T1	C	Base-ST	2	0.094175	0.095279	0.095024	0.096815
science	T1	C	Base-ST	3	0.083815	0.086304	0.09062	0.094315
science	T1	C	Base-ST	4	0.087784	0.091224	0.098081	0.103143
science	T1	C	Base-ST	5	0.078259	0.081878	0.096034	0.101724
science	T1	C	Base-ST	6	0.091031	0.098355	0.106575	0.117172
science	T2	C	Base-ST	1	0.065456	0.064626	0.087105	0.086897
science	T2	C	Base-ST	2	0.063994	0.064257	0.085631	0.08676
science	T2	C	Base-ST	3	0.057436	0.059037	0.080508	0.083504
science	T2	C	Base-ST	4	0.059844	0.062283	0.086441	0.09068
science	T2	C	Base-ST	5	0.047085	0.04963	0.083111	0.087918
science	T2	C	Base-ST	6	0.061208	0.067272	0.095825	0.105506

Notes: This table shows *T/T1/T2 vs. C* in **Model Base-ST** for all 3 test score domains and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$ (compare also Table 4 and Section 4.1). Note that column (6) shows the DiD results with *federal states* fixed effects, (7) shows the same but using adjusted R^2 as IEOp measure. Column (8) shows the DiD results with *school* fixed effects and (9) using additionally adjusted R^2 IEOp measures.

Table A.15: Difference-in-Difference Results: Overview - Model Base-**ST** - Control Group C1

(1) Outcome	(2) Treatment	(3) Control	(4) Model	(5) Control set	(6) R^2-DD-BL	(7) R^2*adjusted*-DD_BL	(8) R^2-DD_SF	(9) R^2*adjusted*-DD_SF
read	T	C1	Base-ST	1	0.010973	0.011505	0.016437	0.018523
read	T	C1	Base-ST	2	0.01901	0.019878	0.024988	0.027575
read	T	C1	Base-ST	3	-0.00342	-0.00234	-0.0023	-5.1E-05
read	T	C1	Base-ST	4	-0.00295	-0.00186	-0.00091	0.001399
read	T	C1	Base-ST	5	-0.00433	-0.00303	0.002639	0.005235
read	T	C1	Base-ST	6	-0.00354	-0.00154	0.000885	0.004162
read	T1	C1	Base-ST	1	0.003644	0.004363	0.009431	0.011337
read	T1	C1	Base-ST	2	0.010015	0.011069	0.016887	0.019275
read	T1	C1	Base-ST	3	-0.01344	-0.01218	-0.01012	-0.00805
read	T1	C1	Base-ST	4	-0.01311	-0.01188	-0.00922	-0.00714
read	T1	C1	Base-ST	5	-0.01218	-0.01073	-0.00434	-0.00193
read	T1	C1	Base-ST	6	-0.01201	-0.00985	-0.00708	-0.004
read	T2	C1	Base-ST	1	-0.02998	-0.03005	0.00043	0.001682
read	T2	C1	Base-ST	2	-0.02866	-0.02848	0.004113	0.005734
read	T2	C1	Base-ST	3	-0.04899	-0.04869	-0.02387	-0.02261
read	T2	C1	Base-ST	4	-0.0489	-0.04874	-0.02314	-0.02197
read	T2	C1	Base-ST	5	-0.05185	-0.05155	-0.01924	-0.01778
read	T2	C1	Base-ST	6	-0.05089	-0.05009	-0.02008	-0.01798
math	T	C1	Base-ST	1	-0.01862	-0.01826	-0.06333	-0.06315
math	T	C1	Base-ST	2	-0.00817	-0.00754	-0.0525	-0.05193
math	T	C1	Base-ST	3	-0.01696	-0.01612	-0.06427	-0.06384
math	T	C1	Base-ST	4	-0.0187	-0.01789	-0.06419	-0.06381
math	T	C1	Base-ST	5	-0.02194	-0.021	-0.05933	-0.05871
math	T	C1	Base-ST	6	-0.02162	-0.02015	-0.05894	-0.05793
math	T1	C1	Base-ST	1	-0.02717	-0.02665	-0.06992	-0.06991
math	T1	C1	Base-ST	2	-0.019	-0.0182	-0.06092	-0.06055
math	T1	C1	Base-ST	3	-0.02989	-0.0289	-0.07313	-0.07289
math	T1	C1	Base-ST	4	-0.03184	-0.03092	-0.07416	-0.07402
math	T1	C1	Base-ST	5	-0.03082	-0.02973	-0.06678	-0.06634
math	T1	C1	Base-ST	6	-0.02971	-0.02805	-0.0663	-0.06542
math	T2	C1	Base-ST	1	-0.0567	-0.05693	-0.05742	-0.05753
math	T2	C1	Base-ST	2	-0.05332	-0.05337	-0.05176	-0.05157
math	T2	C1	Base-ST	3	-0.06402	-0.06396	-0.06756	-0.06757
math	T2	C1	Base-ST	4	-0.06536	-0.06546	-0.06868	-0.06887
math	T2	C1	Base-ST	5	-0.06909	-0.06912	-0.06359	-0.06353
math	T2	C1	Base-ST	6	-0.06601	-0.06568	-0.06123	-0.06069
science	T	C1	Base-ST	1	-0.00581	-0.00538	-0.00808	-0.00657
science	T	C1	Base-ST	2	0.003451	0.004171	0.000448	0.002385
science	T	C1	Base-ST	3	-0.02413	-0.02336	-0.03234	-0.03102
science	T	C1	Base-ST	4	-0.02222	-0.02144	-0.02975	-0.02836
science	T	C1	Base-ST	5	-0.02045	-0.01951	-0.02322	-0.02152
science	T	C1	Base-ST	6	-0.01694	-0.01543	-0.02213	-0.01987
science	T1	C1	Base-ST	1	-0.0132	-0.01259	-0.01528	-0.01396
science	T1	C1	Base-ST	2	-0.00423	-0.00332	-0.00668	-0.00492
science	T1	C1	Base-ST	3	-0.03403	-0.03309	-0.04008	-0.03892
science	T1	C1	Base-ST	4	-0.03253	-0.03161	-0.0383	-0.03713
science	T1	C1	Base-ST	5	-0.02783	-0.02671	-0.02969	-0.02815
science	T1	C1	Base-ST	6	-0.0242	-0.02247	-0.02924	-0.02711
science	T2	C1	Base-ST	1	-0.03743	-0.03757	-0.02125	-0.02049
science	T2	C1	Base-ST	2	-0.03441	-0.03434	-0.01608	-0.01498
science	T2	C1	Base-ST	3	-0.06041	-0.06036	-0.05019	-0.04973
science	T2	C1	Base-ST	4	-0.06047	-0.06055	-0.04994	-0.0496
science	T2	C1	Base-ST	5	-0.059	-0.05896	-0.04262	-0.04196
science	T2	C1	Base-ST	6	-0.05402	-0.05355	-0.03999	-0.03878

Notes: This table shows *T/T1/T2 vs. C1* in **Model Base-ST** for all 3 test score domains and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$ (compare also Table 5 and Section 4.1). Note that column (6) shows the DiD results with *federal states* fixed effects, (7) shows the same but using adjusted R^2 as IEOp measure. Column (8) shows the DiD results with *school* fixed effects and (9) using additionally adjusted R^2 IEOp measures.

Table A.16: Difference-in-Difference Results: Overview - Model Base-**ST** - Control Group C2

(1) Outcome	(2) Treatment	(3) Control	(4) Model	(5) Control set	(6) R^2-DD-BL	(7) $R^2adjusted$-DD_BL	(8) R^2-DD_SF	(9) $R^2adjusted$-DD_SF
read	T	C2	Base-ST	1	0.014695	0.015222	0.02057	0.022444
read	T	C2	Base-ST	2	0.034424	0.035185	0.039258	0.041688
read	T	C2	Base-ST	3	0.010918	0.01172	0.017893	0.019953
read	T	C2	Base-ST	4	0.011301	0.012165	0.018712	0.020863
read	T	C2	Base-ST	5	0.011957	0.01294	0.022806	0.02516
read	T	C2	Base-ST	6	0.011061	0.012372	0.019702	0.022318
read	T1	C2	Base-ST	1	0.007366	0.008079	0.013564	0.015258
read	T1	C2	Base-ST	2	0.025429	0.026377	0.031157	0.033388
read	T1	C2	Base-ST	3	0.000903	0.001877	0.01007	0.011952
read	T1	C2	Base-ST	4	0.001145	0.002144	0.010395	0.012323
read	T1	C2	Base-ST	5	0.004108	0.005246	0.015825	0.017997
read	T1	C2	Base-ST	6	0.002593	0.004064	0.011736	0.01416
read	T2	C2	Base-ST	1	-0.02626	-0.02633	0.004563	0.005603
read	T2	C2	Base-ST	2	-0.01324	-0.01318	0.018383	0.019846
read	T2	C2	Base-ST	3	-0.03465	-0.03463	-0.00368	-0.0026
read	T2	C2	Base-ST	4	-0.03465	-0.03472	-0.00352	-0.0025
read	T2	C2	Base-ST	5	-0.03556	-0.03557	0.000928	0.002142
read	T2	C2	Base-ST	6	-0.03629	-0.03618	-0.00126	0.000179
math	T	C2	Base-ST	1	-0.02231	-0.02197	-0.02839	-0.02774
math	T	C2	Base-ST	2	-0.00217	-0.00166	-0.01059	-0.00948
math	T	C2	Base-ST	3	-0.00737	-0.00679	-0.01748	-0.01648
math	T	C2	Base-ST	4	-0.00951	-0.0089	-0.01836	-0.01735
math	T	C2	Base-ST	5	-0.00795	-0.00727	-0.01207	-0.01085
math	T	C2	Base-ST	6	-0.01357	-0.01278	-0.01725	-0.01602
math	T1	C2	Base-ST	1	-0.03086	-0.03035	-0.03499	-0.0345
math	T1	C2	Base-ST	2	-0.013	-0.01233	-0.01901	-0.01811
math	T1	C2	Base-ST	3	-0.0203	-0.01957	-0.02634	-0.02554
math	T1	C2	Base-ST	4	-0.02265	-0.02193	-0.02832	-0.02757
math	T1	C2	Base-ST	5	-0.01683	-0.016	-0.01952	-0.01848
math	T1	C2	Base-ST	6	-0.02166	-0.02067	-0.0246	-0.02351
math	T2	C2	Base-ST	1	-0.06038	-0.06063	-0.02249	-0.02211
math	T2	C2	Base-ST	2	-0.04732	-0.04749	-0.00985	-0.00912
math	T2	C2	Base-ST	3	-0.05443	-0.05463	-0.02078	-0.02022
math	T2	C2	Base-ST	4	-0.05616	-0.05648	-0.02285	-0.02241
math	T2	C2	Base-ST	5	-0.0551	-0.05539	-0.01633	-0.01567
math	T2	C2	Base-ST	6	-0.05796	-0.0583	-0.01953	-0.01878
science	T	C2	Base-ST	1	-0.00145	-0.00101	0.00946	0.011064
science	T	C2	Base-ST	2	0.016339	0.016956	0.024027	0.026056
science	T	C2	Base-ST	3	-0.00047	0.000147	0.002207	0.003794
science	T	C2	Base-ST	4	0.001246	0.001935	0.003558	0.005233
science	T	C2	Base-ST	5	0.006424	0.007236	0.011134	0.013081
science	T	C2	Base-ST	6	0.00672	0.007789	0.010526	0.012724
science	T1	C2	Base-ST	1	-0.00884	-0.00822	0.002266	0.003679
science	T1	C2	Base-ST	2	0.008657	0.009463	0.016896	0.018749
science	T1	C2	Base-ST	3	-0.01038	-0.00958	-0.00553	-0.00411
science	T1	C2	Base-ST	4	-0.00906	-0.00823	-0.00499	-0.00354
science	T1	C2	Base-ST	5	-0.00096	3.05E-05	0.004661	0.00645
science	T1	C2	Base-ST	6	-0.00053	0.00075	0.003423	0.005484
science	T2	C2	Base-ST	1	-0.03307	-0.03321	-0.0037	-0.00285
science	T2	C2	Base-ST	2	-0.02152	-0.02156	0.007504	0.008694
science	T2	C2	Base-ST	3	-0.03676	-0.03684	-0.01564	-0.01492
science	T2	C2	Base-ST	4	-0.037	-0.03717	-0.01663	-0.016
science	T2	C2	Base-ST	5	-0.03213	-0.03222	-0.00826	-0.00736
science	T2	C2	Base-ST	6	-0.03036	-0.03033	-0.00733	-0.00618

Notes: This table shows *T/T1/T2 vs. C2* in **Model Base-ST** for all 3 test score domains and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$ (compare also Table 5 and Section 4.1). Note that column (6) shows the DiD results with *federal states* fixed effects, (7) shows the same but using adjusted R^2 as IEOp measure. Column (8) shows the DiD results with *school* fixed effects and (9) using additionally adjusted R^2 IEOp measures.

Table A.17: Robustness Check: DiD Results - Model Base-**MT** - Control Group Ch

(1) Outcome	(2) Treatment	(3) Control	(4) Model	(5) Control set	(6) R^2-DD-BL	(7) R^2*adjusted*-DD_BL	(8) R^2-DD_SF	(9) R^2*adjusted*-DD_SF
read	T	Ch	Base-MT	1	-0.0227491	-0.0146742	0.0601538	0.0706621
read	T	Ch	Base-MT	2	-0.0279907	-0.0146276	0.0508613	0.0668377
read	T	Ch	Base-MT	3	-0.0518645	-0.0325074	0.0147625	0.0365238
read	T	Ch	Base-MT	4	-0.0508516	-0.0279142	0.0162766	0.0419544
read	T	Ch	Base-MT	5	-0.0534799	-0.0268254	0.0251033	0.0552038
read	T	Ch	Base-MT	6	-0.0708992	-0.0308939	0.0070143	0.0511753
read	T1	Ch	Base-MT	1	-0.0161309	-0.0079281	0.0620241	0.0724162
read	T1	Ch	Base-MT	2	-0.0250699	-0.0116315	0.0503291	0.0660949
read	T1	Ch	Base-MT	3	-0.0495261	-0.0301554	0.0151379	0.0366661
read	T1	Ch	Base-MT	4	-0.0487099	-0.0257857	0.0166851	0.0421032
read	T1	Ch	Base-MT	5	-0.0491059	-0.0224755	0.0267258	0.0565755
read	T1	Ch	Base-MT	6	-0.0667583	-0.026876	0.0081554	0.0519494
read	T2	Ch	Base-MT	1	-0.0418886	-0.0341019	0.0334383	0.0427186
read	T2	Ch	Base-MT	2	-0.0582661	-0.0454439	0.0176632	0.03206
read	T2	Ch	Base-MT	3	-0.0815606	-0.0630141	-0.0179337	0.0020206
read	T2	Ch	Base-MT	4	-0.0814772	-0.0594952	-0.0171796	0.0065348
read	T2	Ch	Base-MT	5	-0.0838113	-0.0582387	-0.006656	0.0214004
read	T2	Ch	Base-MT	6	-0.1002893	-0.0619177	-0.0230911	0.0185168
math	T	Ch	Base-MT	1	-0.0383542	-0.030447	-0.113262	-0.1077581
math	T	Ch	Base-MT	2	-0.0382803	-0.0250819	-0.1175293	-0.1073031
math	T	Ch	Base-MT	3	-0.0630841	-0.0441709	-0.135462	-0.1199216
math	T	Ch	Base-MT	4	-0.0569075	-0.0343625	-0.1259881	-0.1067127
math	T	Ch	Base-MT	5	-0.0340091	-0.0070513	-0.0894577	-0.065028
math	T	Ch	Base-MT	6	-0.040216	0.0008469	-0.0979307	-0.0605138
math	T1	Ch	Base-MT	1	-0.0272361	-0.0192111	-0.1044502	-0.098916
math	T1	Ch	Base-MT	2	-0.0317752	-0.0185061	-0.1116647	-0.1014976
math	T1	Ch	Base-MT	3	-0.0582248	-0.0393035	-0.1291364	-0.1136771
math	T1	Ch	Base-MT	4	-0.051736	-0.029206	-0.1196117	-0.1004362
math	T1	Ch	Base-MT	5	-0.0261053	0.0008304	-0.0819234	-0.0575838
math	T1	Ch	Base-MT	6	-0.0326471	0.0083018	-0.0912441	-0.054012
math	T2	Ch	Base-MT	1	-0.0532945	-0.0456559	-0.1144988	-0.1095864
math	T2	Ch	Base-MT	2	-0.0639195	-0.0512189	-0.124123	-0.1147551
math	T2	Ch	Base-MT	3	-0.0916578	-0.0735185	-0.1442201	-0.129774
math	T2	Ch	Base-MT	4	-0.0857927	-0.0641568	-0.1354374	-0.1173761
math	T2	Ch	Base-MT	5	-0.0625395	-0.0366154	-0.0981209	-0.074991
math	T2	Ch	Base-MT	6	-0.0665493	-0.027031	-0.1042747	-0.0685113
science	T	Ch	Base-MT	1	-0.0503647	-0.0418693	-0.0230783	-0.014873
science	T	Ch	Base-MT	2	-0.0663383	-0.0525347	-0.0438829	-0.0311266
science	T	Ch	Base-MT	3	-0.0790938	-0.0596141	-0.0769183	-0.0596133
science	T	Ch	Base-MT	4	-0.0766676	-0.0537081	-0.0709693	-0.0500218
science	T	Ch	Base-MT	5	-0.0712979	-0.0444684	-0.0517354	-0.0264631
science	T	Ch	Base-MT	6	-0.0730955	-0.0321738	-0.0509291	-0.011915
science	T1	Ch	Base-MT	1	-0.0500419	-0.0414598	-0.0270883	-0.0191339
science	T1	Ch	Base-MT	2	-0.0680978	-0.0542588	-0.0488929	-0.036452
science	T1	Ch	Base-MT	3	-0.0828596	-0.0634155	-0.0821389	-0.0652059
science	T1	Ch	Base-MT	4	-0.08074	-0.0578488	-0.0765318	-0.0559953
science	T1	Ch	Base-MT	5	-0.0734857	-0.0467423	-0.0564781	-0.0316247
science	T1	Ch	Base-MT	6	-0.0759718	-0.035258	-0.0567455	-0.0183011
science	T2	Ch	Base-MT	1	-0.069238	-0.0610427	-0.0468258	-0.0397547
science	T2	Ch	Base-MT	2	-0.0938409	-0.0805688	-0.0718042	-0.060462
science	T2	Ch	Base-MT	3	-0.1082766	-0.0895869	-0.1059656	-0.0903233
science	T2	Ch	Base-MT	4	-0.1084335	-0.086413	-0.1024249	-0.083325
science	T2	Ch	Base-MT	5	-0.1026931	-0.0769291	-0.082551	-0.0592352
science	T2	Ch	Base-MT	6	-0.1023156	-0.062985	-0.0794007	-0.0428064

Notes: This table shows *T/T1/T2 vs. Ch* in **Model Base-MT** for all 3 test score domains and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$ (compare also Section 4.1). Note that column (6) shows the DiD results with *federal states* fixed effects, (7) shows the same but using adjusted R^2 as IEOp measure. Column (8) shows the DiD results with *school* fixed effects and (9) using additionally adjusted R^2 IEOp measures.

Table A.18: Robustness Check: DiD Results - Model Base-**ST** - Control Group Ch

(1) Outcome	(2) Treatment	(3) Control	(4) Model	(5) Control set	(6) R^2-DD-BL	(7) $R^2adjusted$-DD_BL	(8) R^2-DD_SF	(9) $R^2adjusted$-DD_SF
read	T	Ch	Base-ST	1	-0.13167	-0.10788	-0.10471	-0.08823
read	T	Ch	Base-ST	2	-0.1349	-0.09595	-0.101	-0.06869
read	T	Ch	Base-ST	3	-0.18792	-0.13463	-0.16528	-0.11825
read	T	Ch	Base-ST	4	-0.18752	-0.13006	-0.16377	-0.11234
read	T	Ch	Base-ST	5	-0.1899	-0.12251	-0.15552	-0.09344
read	T	Ch	Base-ST	6	-0.2207	-0.11429	-0.18673	-0.08511
read	T1	Ch	Base-ST	1	-0.139	-0.11502	-0.11172	-0.09541
read	T1	Ch	Base-ST	2	-0.1439	-0.10476	-0.1091	-0.07699
read	T1	Ch	Base-ST	3	-0.19793	-0.14447	-0.17311	-0.12625
read	T1	Ch	Base-ST	4	-0.19768	-0.14008	-0.17209	-0.12088
read	T1	Ch	Base-ST	5	-0.19775	-0.13021	-0.1625	-0.1006
read	T1	Ch	Base-ST	6	-0.22917	-0.12259	-0.19469	-0.09327
read	T2	Ch	Base-ST	1	-0.17263	-0.14943	-0.12072	-0.10507
read	T2	Ch	Base-ST	2	-0.18257	-0.14432	-0.12187	-0.09054
read	T2	Ch	Base-ST	3	-0.23349	-0.18098	-0.18685	-0.14081
read	T2	Ch	Base-ST	4	-0.23347	-0.17694	-0.186	-0.1357
read	T2	Ch	Base-ST	5	-0.23742	-0.17103	-0.1774	-0.11646
read	T2	Ch	Base-ST	6	-0.26805	-0.16284	-0.20769	-0.10725
math	T	Ch	Base-ST	1	-0.04886	-0.02013	-0.12096	-0.10231
math	T	Ch	Base-ST	2	-0.04189	0.005881	-0.11049	-0.07318
math	T	Ch	Base-ST	3	-0.06617	0.003817	-0.1406	-0.0817
math	T	Ch	Base-ST	4	-0.05855	0.017325	-0.13096	-0.06622
math	T	Ch	Base-ST	5	-0.03955	0.050224	-0.09572	-0.01637
math	T	Ch	Base-ST	6	-0.0551	0.090802	-0.11763	0.014335
math	T1	Ch	Base-ST	1	-0.05741	-0.02851	-0.12755	-0.10908
math	T1	Ch	Base-ST	2	-0.05272	-0.00478	-0.11891	-0.08181
math	T1	Ch	Base-ST	3	-0.0791	-0.00896	-0.14946	-0.09075
math	T1	Ch	Base-ST	4	-0.0717	0.004295	-0.14092	-0.07644
math	T1	Ch	Base-ST	5	-0.04843	0.041492	-0.10317	-0.024
math	T1	Ch	Base-ST	6	-0.06318	0.082904	-0.12499	0.006846
math	T2	Ch	Base-ST	1	-0.08693	-0.05879	-0.11506	-0.09669
math	T2	Ch	Base-ST	2	-0.08704	-0.03995	-0.10975	-0.07282
math	T2	Ch	Base-ST	3	-0.11324	-0.04402	-0.1439	-0.08543
math	T2	Ch	Base-ST	4	-0.10521	-0.03025	-0.13545	-0.07128
math	T2	Ch	Base-ST	5	-0.0867	0.002105	-0.09997	-0.02119
math	T2	Ch	Base-ST	6	-0.09949	0.045279	-0.11992	0.011576
science	T	Ch	Base-ST	1	-0.0953	-0.06736	-0.0527	-0.03236
science	T	Ch	Base-ST	2	-0.10149	-0.0555	-0.05884	-0.02053
science	T	Ch	Base-ST	3	-0.15535	-0.09345	-0.12988	-0.07546
science	T	Ch	Base-ST	4	-0.1455	-0.07842	-0.12239	-0.06264
science	T	Ch	Base-ST	5	-0.1485	-0.07076	-0.10819	-0.0368
science	T	Ch	Base-ST	6	-0.17429	-0.05204	-0.13762	-0.02242
science	T1	Ch	Base-ST	1	-0.10269	-0.07457	-0.0599	-0.03974
science	T1	Ch	Base-ST	2	-0.10918	-0.063	-0.06597	-0.02784
science	T1	Ch	Base-ST	3	-0.16526	-0.10317	-0.13762	-0.08336
science	T1	Ch	Base-ST	4	-0.15581	-0.08859	-0.13094	-0.07141
science	T1	Ch	Base-ST	5	-0.15588	-0.07797	-0.11466	-0.04343
science	T1	Ch	Base-ST	6	-0.18154	-0.05908	-0.14472	-0.02966
science	T2	Ch	Base-ST	1	-0.12693	-0.09956	-0.06587	-0.04627
science	T2	Ch	Base-ST	2	-0.13936	-0.09402	-0.07536	-0.0379
science	T2	Ch	Base-ST	3	-0.19163	-0.13044	-0.14773	-0.09418
science	T2	Ch	Base-ST	4	-0.18375	-0.11753	-0.14258	-0.08388
science	T2	Ch	Base-ST	5	-0.18705	-0.11021	-0.12758	-0.05724
science	T2	Ch	Base-ST	6	-0.21137	-0.09016	-0.15547	-0.04133

Notes: This table shows *T/T1/T2 vs. Ch* in **Model Base-ST** for all 3 test score domains and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$ (compare also Section 4.1). Note that column (6) shows the DiD results with *federal states* fixed effects, (7) shows the same but using adjusted R^2 as IEOp measure. Column (8) shows the DiD results with *school* fixed effects and (9) using additionally adjusted R^2 IEOp measures.

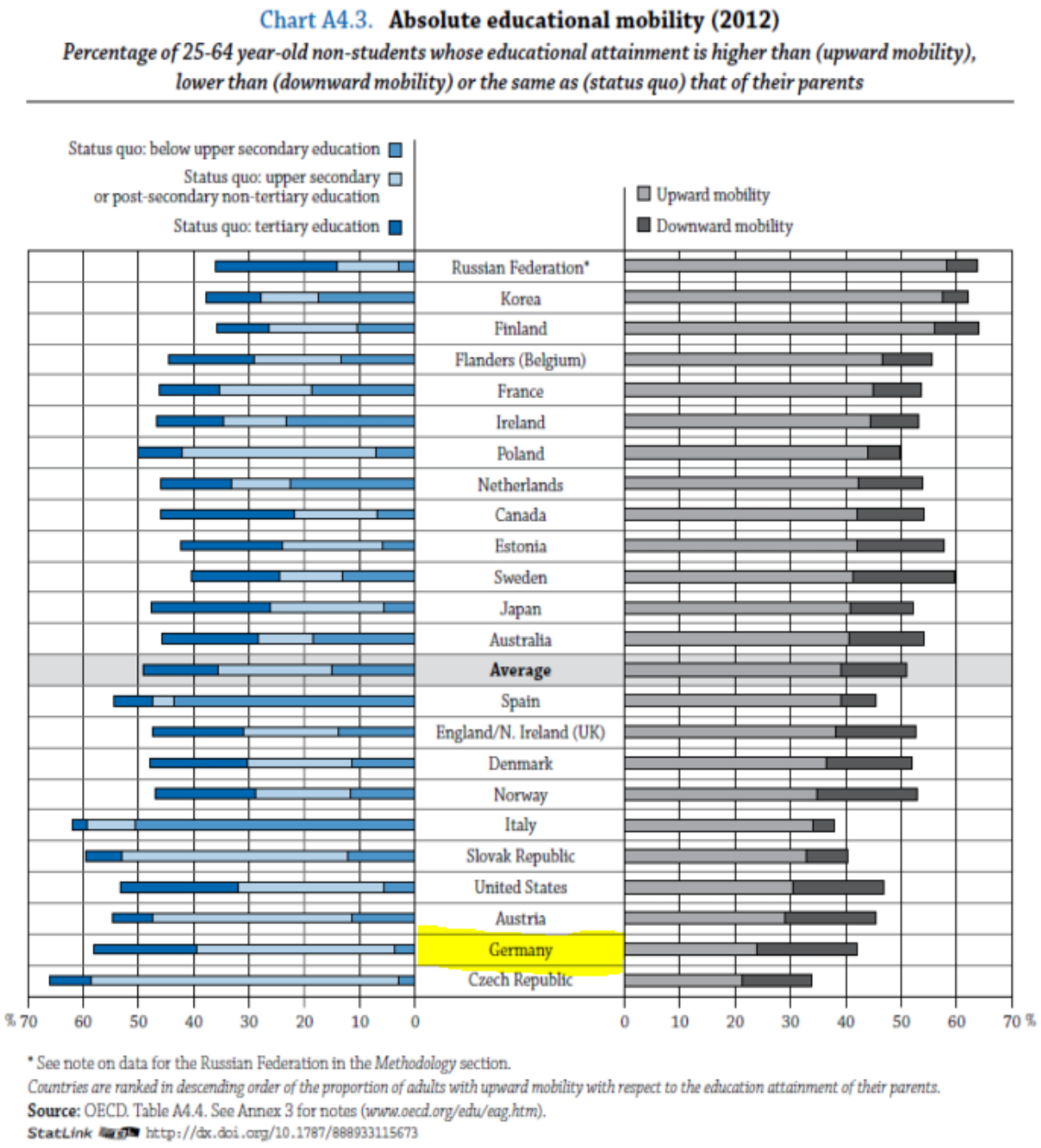
Table A.19: Robustness Check: Overview of DiD Results - Model Full-**MT** - Control Group C

(1) Outcome	(2) Treatment	(3) Control	(4) Model	(5) Control set	(6) R^2-DD-BL	(7) R^2*adjusted*-DD_BL	(8) R^2-DD_SF	(9) R^2*adjusted*-DD_SF
read	T	C	Full-MT	1	0.033046	0.036931	0.093264	0.091588
read	T	C	Full-MT	2	0.032047	0.038399	0.09349	0.094568
read	T	C	Full-MT	3	0.025697	0.035358	0.089567	0.094009
read	T	C	Full-MT	4	0.031441	0.042861	0.093703	0.100081
read	T	C	Full-MT	5	0.009385	0.022148	0.083888	0.091618
read	T	C	Full-MT	6	0.029813	0.049689	0.097794	0.113363
read	T1	C	Full-MT	1	0.035779	0.039219	0.088497	0.088696
read	T1	C	Full-MT	2	0.031723	0.037524	0.086996	0.089778
read	T1	C	Full-MT	3	0.025085	0.034073	0.083867	0.089847
read	T1	C	Full-MT	4	0.030305	0.040983	0.088291	0.096144
read	T1	C	Full-MT	5	0.010393	0.022364	0.079617	0.088778
read	T1	C	Full-MT	6	0.030198	0.049014	0.092885	0.109591
read	T2	C	Full-MT	1	0.00766	0.010701	0.060214	0.05935
read	T2	C	Full-MT	2	-0.00179	0.003551	0.055756	0.057333
read	T2	C	Full-MT	3	-0.00762	0.000854	0.051719	0.056415
read	T2	C	Full-MT	4	-0.00316	0.006971	0.055108	0.061616
read	T2	C	Full-MT	5	-0.02463	-0.01324	0.046732	0.054531
read	T2	C	Full-MT	6	-0.003	0.015122	0.062835	0.078121
math	T	C	Full-MT	1	0.051045	0.053245	0.060234	0.045001
math	T	C	Full-MT	2	0.061606	0.06553	0.07126	0.05832
math	T	C	Full-MT	3	0.050375	0.056285	0.063509	0.052661
math	T	C	Full-MT	4	0.060182	0.067332	0.073002	0.06388
math	T	C	Full-MT	5	0.049744	0.057805	0.071991	0.063996
math	T	C	Full-MT	6	0.057976	0.070598	0.073331	0.070357
math	T1	C	Full-MT	1	0.061997	0.063882	0.06555	0.054744
math	T1	C	Full-MT	2	0.067896	0.071386	0.073521	0.0648
math	T1	C	Full-MT	3	0.055193	0.060522	0.066284	0.059448
math	T1	C	Full-MT	4	0.064788	0.071281	0.076159	0.070984
math	T1	C	Full-MT	5	0.056959	0.064306	0.076386	0.072303
math	T1	C	Full-MT	6	0.064858	0.076467	0.077088	0.077748
math	T2	C	Full-MT	1	0.030857	0.0323	0.04556	0.034304
math	T2	C	Full-MT	2	0.031603	0.034538	0.051558	0.042223
math	T2	C	Full-MT	3	0.019254	0.0239	0.043209	0.035603
math	T2	C	Full-MT	4	0.027894	0.033629	0.05175	0.045695
math	T2	C	Full-MT	5	0.017998	0.024511	0.051604	0.046557
math	T2	C	Full-MT	6	0.027428	0.037931	0.054375	0.053881
science	T	C	Full-MT	1	0.055852	0.058481	0.112282	0.10136
science	T	C	Full-MT	2	0.06769	0.072282	0.113213	0.104678
science	T	C	Full-MT	3	0.071157	0.07811	0.112054	0.106061
science	T	C	Full-MT	4	0.078049	0.086378	0.118382	0.114141
science	T	C	Full-MT	5	0.064415	0.073681	0.112392	0.109258
science	T	C	Full-MT	6	0.085998	0.100831	0.128497	0.131994
science	T1	C	Full-MT	1	0.057771	0.06004	0.107569	0.101061
science	T1	C	Full-MT	2	0.067756	0.071878	0.10737	0.103079
science	T1	C	Full-MT	3	0.069246	0.07559	0.105562	0.103533
science	T1	C	Full-MT	4	0.075568	0.083212	0.111904	0.111548
science	T1	C	Full-MT	5	0.063609	0.072126	0.106835	0.107539
science	T1	C	Full-MT	6	0.084183	0.09796	0.121978	0.128986
science	T2	C	Full-MT	1	0.032707	0.034542	0.08496	0.078063
science	T2	C	Full-MT	2	0.037865	0.041455	0.082288	0.077407
science	T2	C	Full-MT	3	0.037788	0.043486	0.077657	0.074821
science	T2	C	Full-MT	4	0.042629	0.049558	0.082498	0.081223
science	T2	C	Full-MT	5	0.029119	0.036853	0.076974	0.076686
science	T2	C	Full-MT	6	0.052592	0.065355	0.095818	0.101716

Notes: This table shows *T/T1/T2 vs. C* in **Model Full-MT** for all 3 test score domains and for each version adding all 6 control sets from $1 = [(i) + (ii)]$ until $6 = [(i) + (ii) + (iii) + (iv) + (v) + (vi) + (vii)]$ (compare also Section 4.1). Note that column (6) shows the DiD results with *federal states* fixed effects, (7) shows the same but using adjusted R^2 as IEOp measure. Column (8) shows the DiD results with *school* fixed effects and (9) using additionally adjusted R^2 IEOp measures.

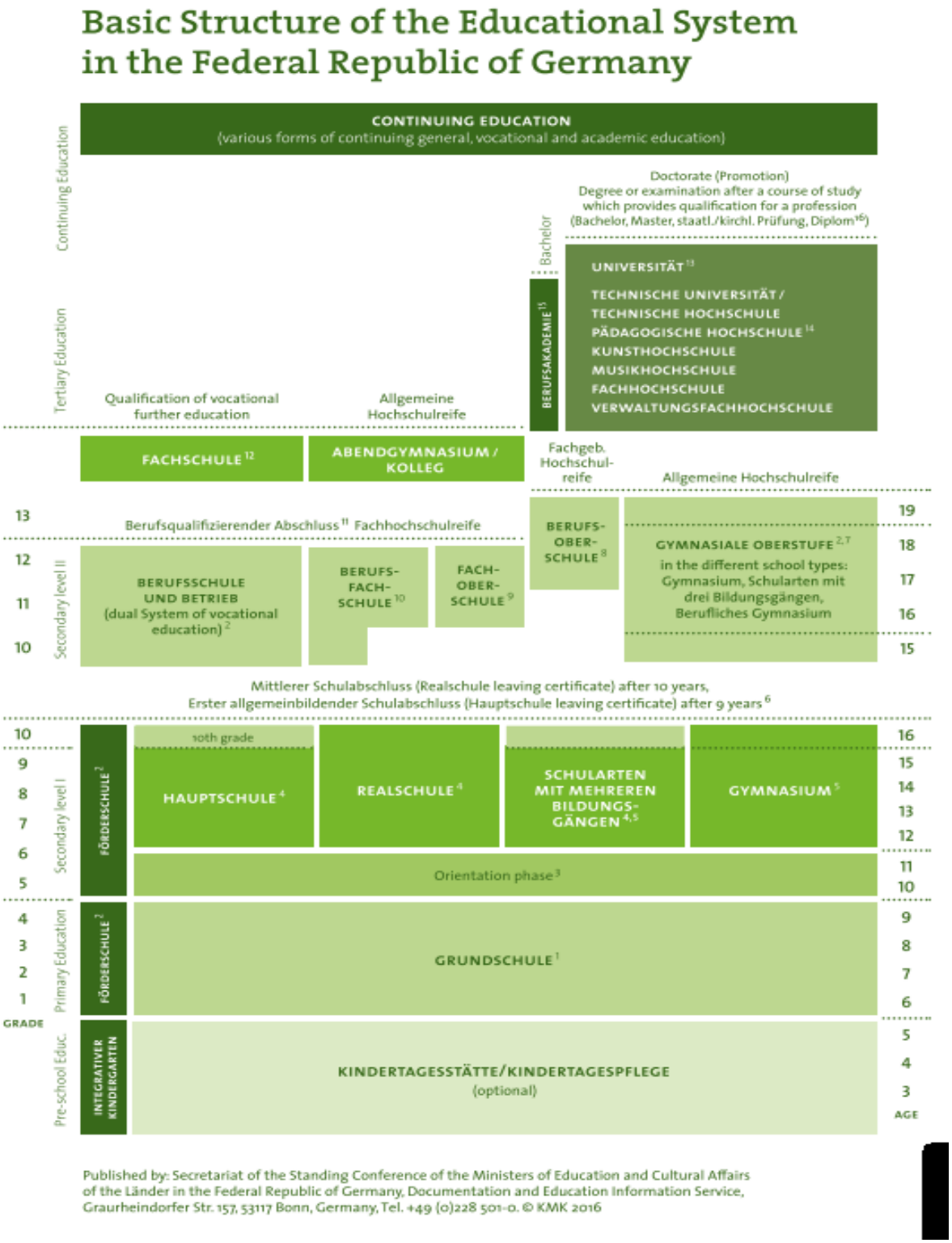
A.5 Supplementary Figures

Figure A.1: Absolute educational mobility (2012)



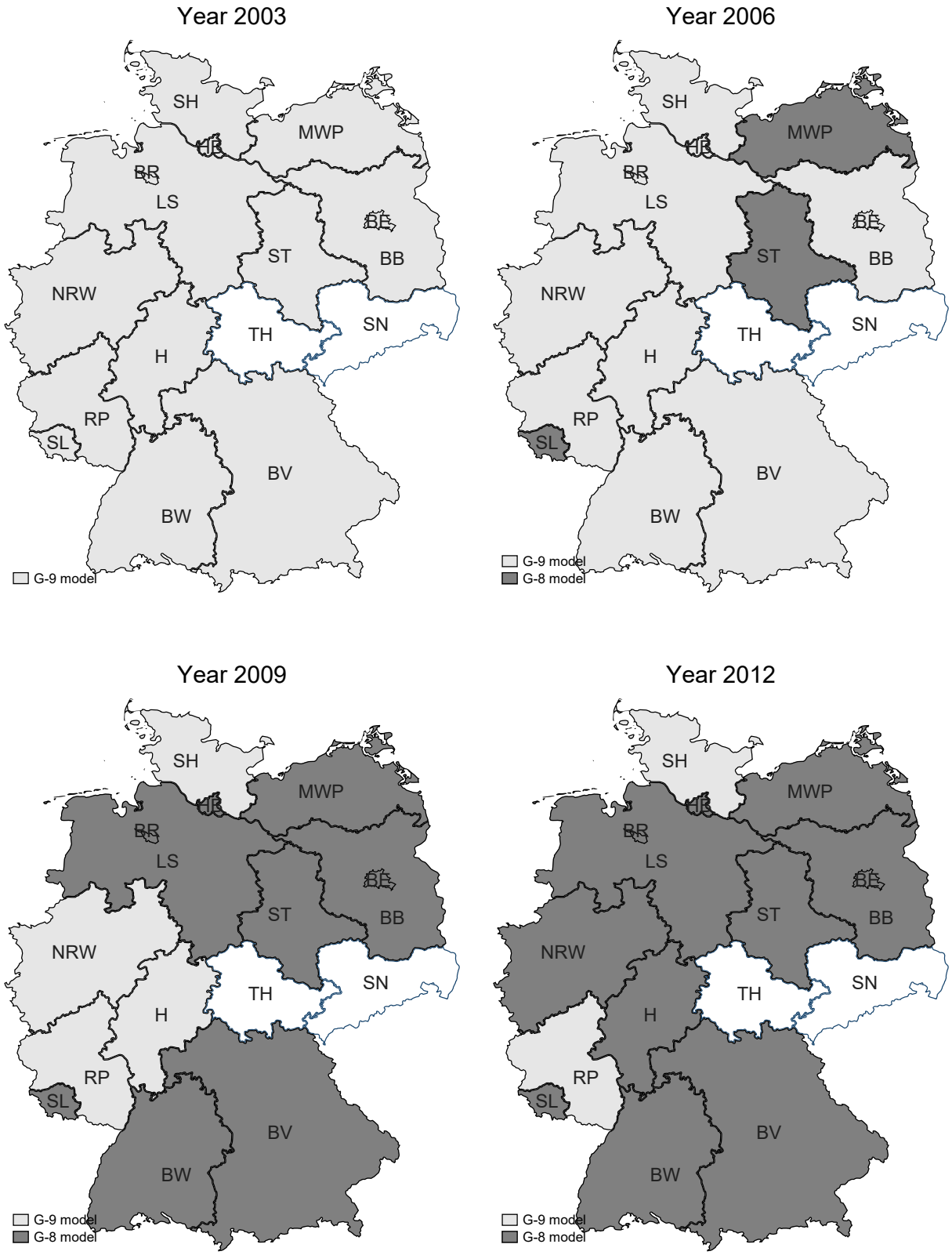
Notes: This figure illustrates absolute educational mobility. It shows the percentage of 25-64 year-old non-students whose educational attainment is higher (upward mobility) or lower (downward mobility) or the same as (status quo) that of their parents as measured in 2012 by the OECD.
Source: Figure taken from OECD (2013b; *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed*).

Figure A.2: Structure of the German educational system



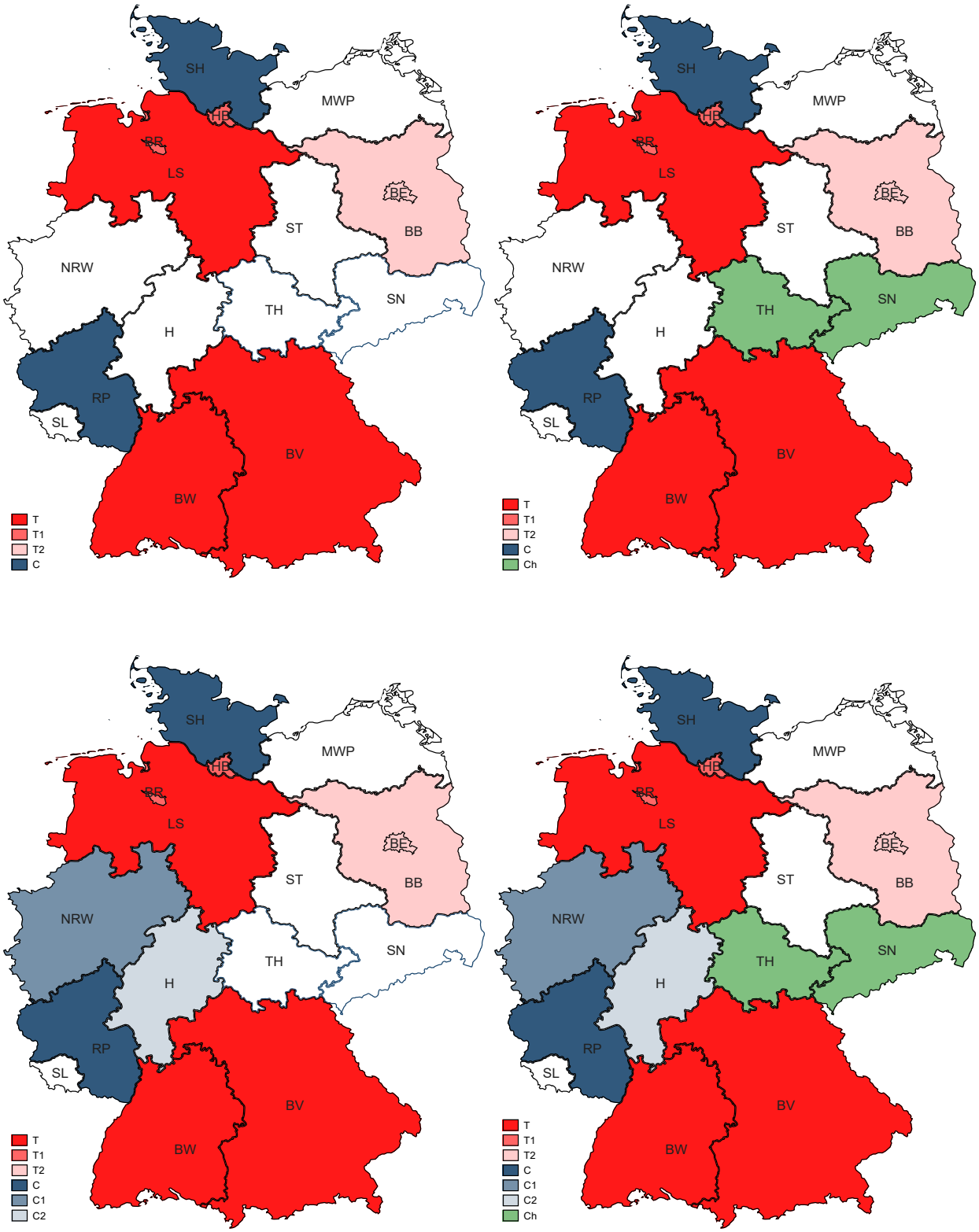
Notes: This figure illustrates the basic structure of the German education system. For more details on the German educational system, see Standing Conference of Education Ministers (2016).
Source: Figure taken from Standing Conference of Education Ministers (2016): Basic Structure of the Education System in the Federal Republic of Germany.

Figure A.3: Overview of G-8-Reform across federal states for students tested in PISA (2003-2012)



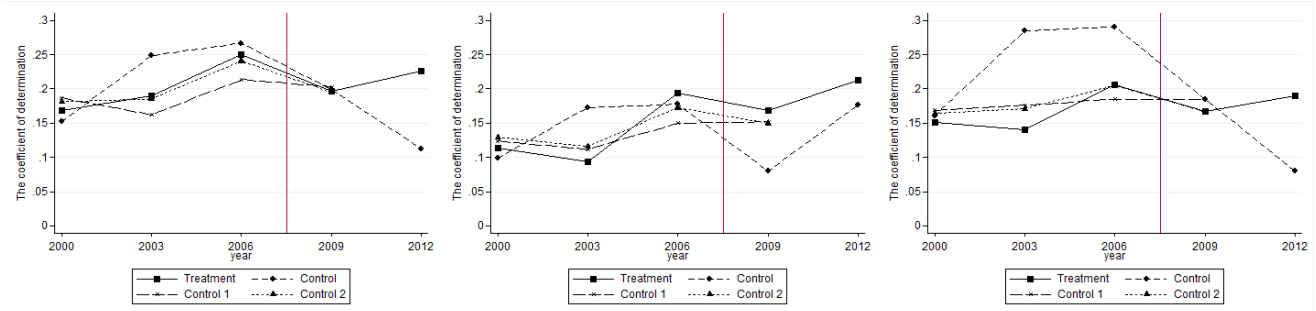
Notes: This figure illustrates whether 9^{th} graders attending a Gymnasium tested in a PISA-test year (2003, 2006, 2009, 2012) were still taught in a G-9 model (light grey) or were already in a reformed G-8 model (dark grey).

Figure A.4: Overview of the Treatment/Control Group setting



Notes: The left-hand figures show the main Treatment/Control Group comparisons for the *medium-term (2003-2012)* in the top and for the *short-term (2003-2009)* model in the bottom panel with main *Control Groups* ***C/C1/C2*** and main *Treatment Groups* ***T/T1/T2***. The right-hand side shows the same settings including *Control Group* ***Ch***.

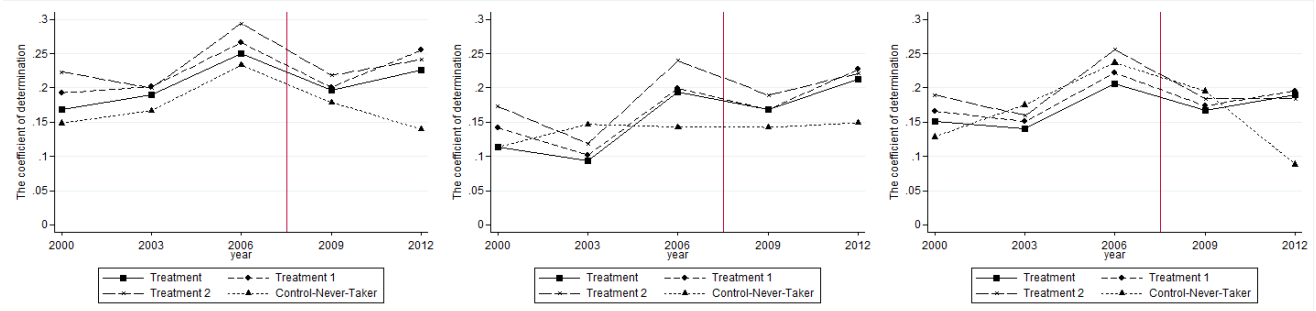
Figure A.5: DiD Graphs of IEOp measure for main Treatment/Control Groups



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

Notes: This figure shows the DiD graphs for all three test domains confirming the parallel trend assumption to hold. *Treatment* is the main treatment group T, *Control* is the main control group C, *Control1* is the main additional short-term model control group C1 and *Control2* is an extended short-term model control group C2.
Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009 and 2012.

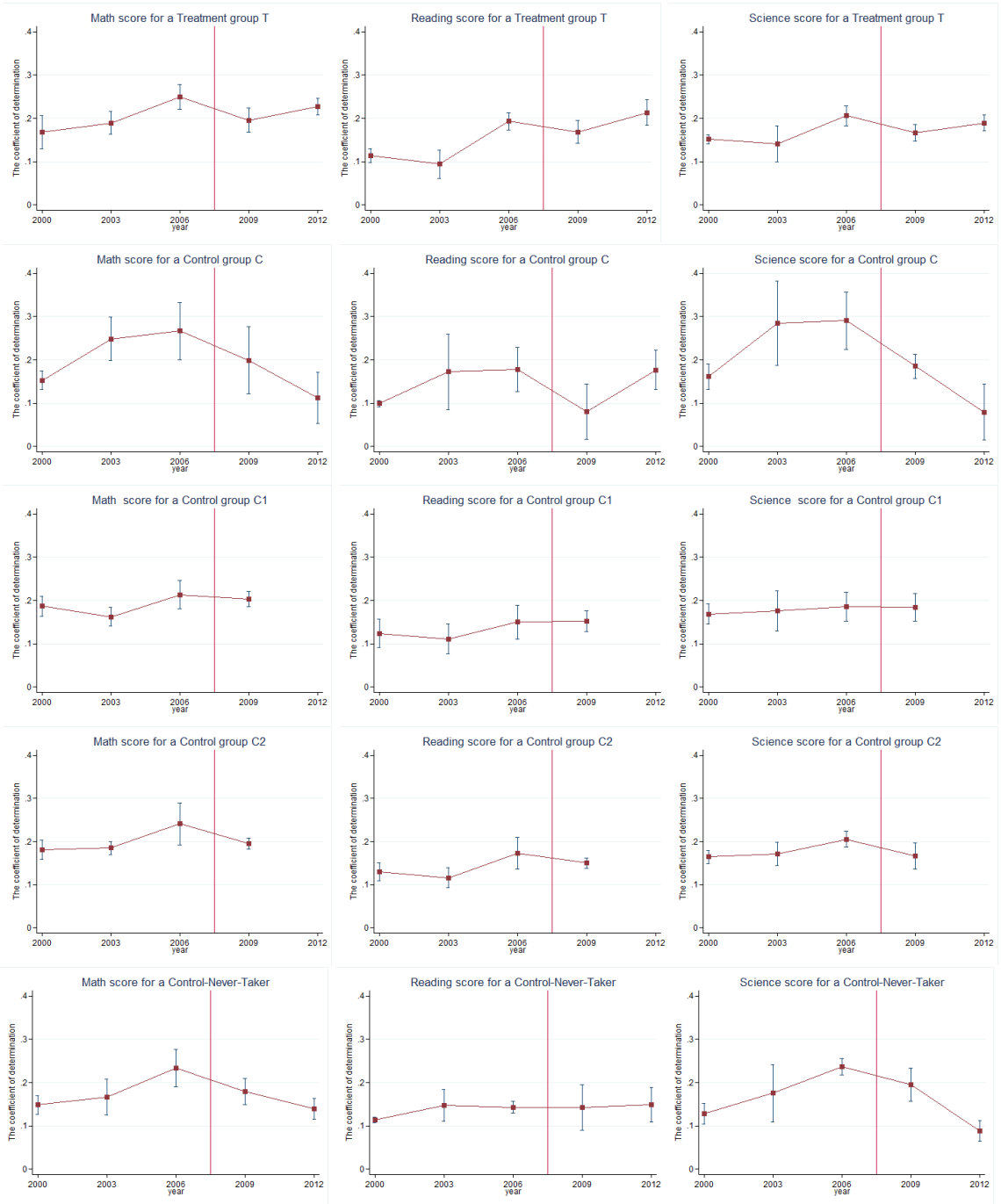
Figure A.6: Robustness - DiD Graphs of IEOp measure for enlarged Treatment/Control Groups



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

Notes: This figure shows the DiD graphs for all three test domains confirming the parallel trend assumption to hold, even when enlarging the treatment group to include instead of three federal states (main Treatment Group T), five (Treatment T1 Extension Group) or seven federal states (Treatment T2 Extension Group) and when comparing it to the never-changing control group (Control-Never-Taker (C-NT)) consisting of four federal states (cf. Section 4).
Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009 and 2012.

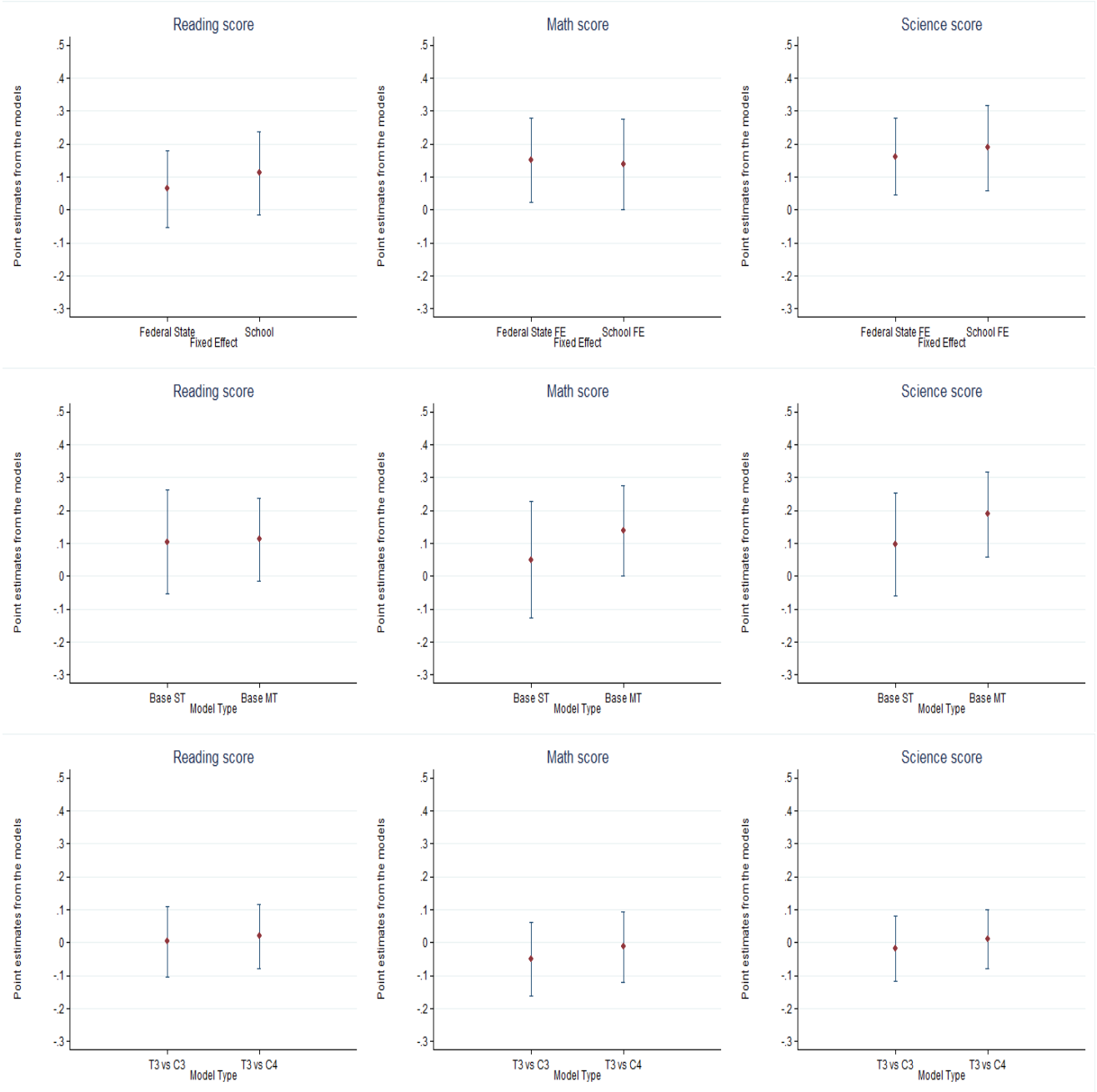
Figure A.7: IEOp measure for main Treatment/Control Groups over time period (2000-2012)



Notes: This figure shows the IEOp measure ($R2_{adjusted}$) with 95% confidence intervals over the whole time period. Standard errors to construct confidence intervals are calculated according to Appendix A.3.1.

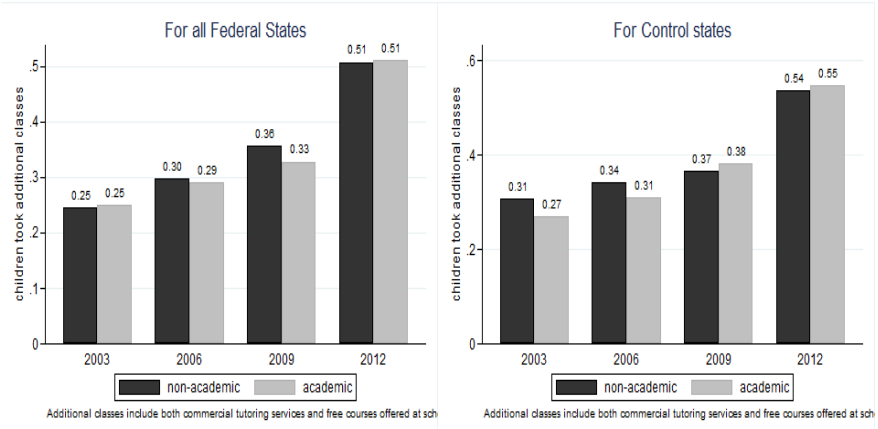
Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009 and 2012.

Figure A.8: Graphical Illustration of Main Results



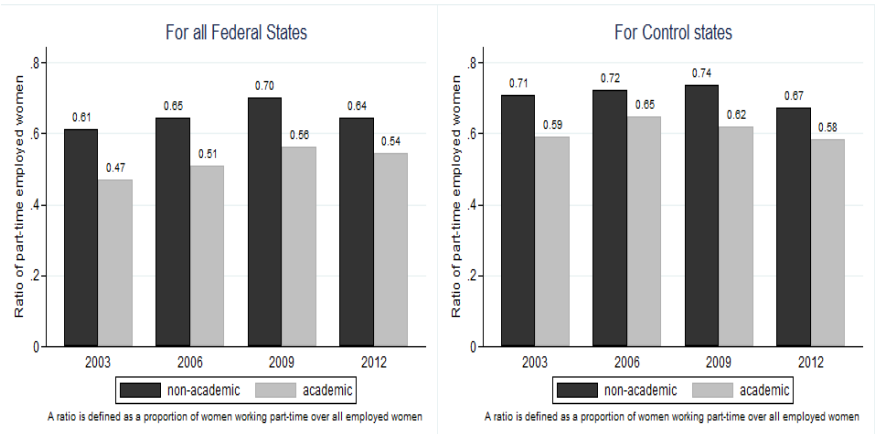
Notes: This figure shows the DiD estimates based on all three test domains and illustrates the main results as shown in the main text in Section 5. The first row of graphs corresponds to Appendix A.4. It also confirms that the Difference-in-Differences strategy works in terms of potential concerns of sorting on the school level, as the main regression patterns remain unchanged whether controlling for school fixed effects or only federal states fixed effects. The second row of graphs corresponds to Table 4. It shows that the medium-term results are statistically significant and larger than the short-term results. Finally, the last row of results corresponds to Table 5. Note that control group C3 corresponds to C1 in the main text and control group C4 to C2 in the main text. It confirms that the main results also hold for enlarged control groups, lending some external validity to the results for the whole of Germany. The IEOp measure was estimated as explained in Section 4.1 and standard errors were calculated according to Appendix A.3.1.
Source: Author's own calculations based on PISA 2000, 2003, 2006, 2009 and 2012.

Figure A.9: Potential Mechanism: Extra Tuition



Notes: This figure shows the percentage of tested students indicating that they took extra classes beyond official school lessons. This mostly includes paid extra tuition. The black bars correspond to students growing up in non-academic households, whereas the grey bars show results for students from academic households, i.e. growing up with at least one parent who has a university diploma (ISCED-level is greater than 5 or 6). The first panel shows that there was an upward trend in the demand for extra classes/tuition between 2003 and 2012 across all federal states. The second panel shows that in treatment states, the increase in extra tuition has been stronger for students from academic than non-academic households in the post-reform period from 2009 to 2012. This indicates that this differential adjustment with respect to extra-tuition depending on a student's parental educational background may explain the observed patterns in the main results. As in control groups, no such differential response in years 2009 and 2012 can be observed. Data are based on responses in both parental and student questionnaires.

Figure A.10: Potential Mechanism: Time Investment by Mothers



Notes: This figure shows the percentage of tested students whose mothers work part-time. The black bars correspond to students growing up in households in which the mother only works part-time and has more time to help them with school work, whereas the grey bars show results for students from households in which the mother is working full-time. The first panel shows that there is an upward trend of mothers working part-time between 2003 and 2009 across all federal states. Only afterwards do mothers increasingly work full-time. The second panel shows that in treatment states, the increase in mothers working part-time remained constant in 2012 for students from academic rather than non-academic households in the post-reform period from 2009 to 2012. This indicates that a differential adjustment with respect to maternal time investment in their school-age children depending on their educational background may explain the observed patterns in the main results. As in the control groups, no such differential response can be observed. Data are based on responses from both parental and student questionnaires.