UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper
Economic Series 15-06
June 2015
ISSN 22340-5031

Departamento de Economía
Universidad Carlos III de Madrid
C/Madrid, 126 28903 Getafe (Spain)
Fax (34-91) 6249875

The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment

Vincenzo Andrietti[1]

Abstract

I exploit a unique educational policy - implemented in most German states between 2001 and 2007 - that reduced high school duration by one year while keeping its curriculum unaltered to investigate how the resulting increase in learning intensity affected student achievement. Using 2000-2009 PISA data and a difference-in-differences approach, I find robust evidence that the reform significantly improved the reading, mathematics, and science literacy skills acquired by academic-track high school students upon treatment. A more direct estimate of the effects of the increased learning intensity - as measured by the cumulative weekly number of instructional hours delivered in high school grades - corroborates the latter finding. Furthermore, there is some evidence that the effects of the reform differ by gender and grade retention. Finally, I find no evidence of a significant average effect of the reform on high school grade retention, although I do find that the latter increased significantly for boys and for students with a migration background.

***Keywords**:* G8, Learning intensity, Instructional hours, Student achievement, Academic-track high school, Difference-in-Differences

***JEL Classification**:* I21, I28, D04

***Acknowledgments**:* I wish to thank Xuejuan Su, Jan Stuhler, Vincent Hildebrand, and Julio Caceres-Delpiano as well as seminar participants at IQB Berlin, Universidad Carlos III de Madrid, University of Toronto, CEA conference 2015, SOLE-EALE world conference 2015, and EEA conference 2015 for helpful comments and suggestions. I gratefully acknowledge IQB-FDZ for the use of PISA data. This research project was carried out at IQB-Berlin, Universidad Carlos III de Madrid, and University of Toronto – Centre for Industrial Relations and Human Resources. The hospitality of these research institutions is gratefully acknowledged.

[1] Università "G. d'Annunzio" di Chieti e Pescara, Dipartimento di Scienze Filosofiche, Pedagogiche ed Economico-Quantitative (DiSFPEQ), Viale Pindaro 42, 65127 Pescara, Italy. E-mail: vincenzo.andrietti@unich.it.

1 Introduction

High school duration and curricula design are important features of the school system since they shape the workload distribution across grades and the learning intensity (i.e., the amount of workload per unit of time) that students have to cope with, and might affect both the level and the distribution of students' cognitive skills. Optimal design of duration and intensity of high school learning is therefore key to ensure that university-bound students are equipped with the skills they need to succeed in higher education – and ultimately on the labor market – while allocating limited public resources as efficiently as possible.

If high school length were shortened by cutting existing curricula, thereby keeping learning intensity constant, substantial cost savings could be achieved and students could enter the labor market sooner – yet this might leave them without important skills and knowledge or have other adverse consequences on learning or career outcomes. By contrast, shortening high school length while keeping the curriculum unaltered would imply reallocating the total amount of instructional hours required for graduation (and the corresponding curriculum) over a reduced period of time, thereby increasing the intensity of student learning – but potentially also creating an increased burden on students. In this second scenario, understanding the relationship between learning intensity at school and human capital accumulation is important for the optimal design of curricula.

This study contributes to the literature by exploiting a unique educational policy – implemented in most German states between 2001 and 2007 – that reduced high school duration by one year while keeping its curriculum unaltered.[1] By redistributing the instructional hours formerly allotted to the last grade of high school and the corresponding curriculum across the previous grades, the G8 reform led to a higher workload per grade, thereby increasing the intensity of student learning.

Despite the controversial implementation of the G8 reform and the ongoing debate around it, which has already led some states to (partially) switch back to the G9 regime, there is still a lack of evidence on how the reform has affected student achievement. This study contributes to filling this gap by investigating the impact of the G8 reform on the reading, mathematics, and science literacy skills of academic-track ninth-graders, as measured by the Program of International Student Assessment (PISA). This is an important matter, given the impact that these skills have been shown to have both at the micro – e.g., on individual earnings and educational attainment (Heckman, Stixrud, and Urzua, 2006) – and at the macro level – e.g., on economic growth (Hanushek and Wössmann, 2008).

[1] While in the few states – i.e., Berlin and Brandenburg – where the transition to secondary school (tracking) takes place in grade seven the length of high school was reduced from seven to six years, in most states – with tracking taking place in grade five – the reform shortened the length of high school from nine to eight years. In the following, high school length refers to the last eight (or nine) years of school from grade five until graduation, and the reform will be referred to as the G8 reform.

My difference-in-differences (DD) estimation strategy exploits over-time and across-state heterogeneity in the reform implementation to isolate the causal effects of the reform from any other potential confounding factors. I find that the increased learning intensity introduced by the G8 reform had positive and significant effects on student reading, math, and science achievement (within a 0.095-0.145 standard deviations range). I offer graphical evidence that pre-reform outcomes followed plausibly parallel paths for treated and control states and, as a consequence, my results are not just picking up long-running trends in student outcome differences between treated and control states. A variety of falsification and specification tests lend additional support to the common trends assumption, buttressing a causal interpretation of my findings. Further robustness checks are also provided to address specification and selection issues that might be driving my results and/or threatening their internal validity. Overall, the robustness analysis suggests that the treatment assignment can be plausibly considered as good as random, and, as a consequence, the G8 reform can be viewed as a quasi-experiment.

Besides estimating the effects of a G8 treatment variable that discretely switches off and on, I exploit the variation over time and across states of a more direct measure of the increased learning intensity, captured by the cumulative number of weekly hours of instruction provided in high school. The estimation results are in line with my main results: a twenty-hour increase distributed over grades 5-9, or a ten-hour increase distributed over grades 8-9, leads to an average increase in student achievement of 0.08-0.15 standard deviations, respectively, depending on the subject.

Furthermore, to shed further light on the effects of the reform, I estimate additional specifications that explore possible heterogeneous policy effects. I find that the reform effect is driven by girls in reading skills and by high-ability students (i.e., students that did not experience grade retention episodes) in reading, math, and science skills.

Finally, I provide evidence on how the G8 reform has affected the probability of repeating a high school grade for those cohorts that entered high school after the first treated cohort, as an indicator of potential unintended effects of the reform. I find no evidence of a significant average effect of the reform on high school grade retention. However, I do find that the latter increased significantly for boys and for students with a migration background.

This study differs from the existing G8 literature in several dimensions. First, I exploit a large dataset representative of the sixteen German states, covering the time period in which the G8 reform was implemented in most states. Furthermore, my student outcomes are pre-graduation standardized test scores in a wide range of subjects. Finally, and most importantly, by focusing on the achievement of academic-track ninth-graders, I am essentially considering the impact of the increased learning intensity on the performance of students that, by the end of grade 9 post-reform, have received about the same amount of instruction, and covered the same curriculum, than students that have completed two-

thirds of grade 10 pre-reform.[2] Therefore, this study is not about the overall G8 reform effect (i.e., higher intensity and shorter duration), but only focuses on the higher intensity aspect.[3]

Besides its natural policy interest, my study also adds to the existing literature by showing that students benefit from increased instructional time despite the increased burden of a higher learning intensity. This finding lends further support to past studies showing a beneficial impact of additional instructional hours in alternative settings, where the learning intensity is either kept constant or reduced.

The remainder of this work is organized as follows. Section 2 provides background on the German educational system and on the G8 reform. Section 3 reviews the related literature. Section 4 illustrates the empirical strategy. Section 5 describes the data. The main results are presented in Sections 6. Section 7 probes the robustness of the findings. Section 8 reports further results (e.g., heterogeneity analysis, grade retention results). Section 9 concludes.

2 The G8 Reform

Educational policy in the Federal Republic of Germany is under the responsibility of the sixteen federal states. In general, children enroll in primary school at the age of six. They continue on to secondary school after four years.[4] Students are then tracked into three basic types of secondary school, each offering a single educational track geared toward the attainment of a specific school-leaving certificate.[5] The basic-track school (*Hauptschule*) and the middle-track school (*Realschule*) provide schooling through grade 9 or 10, grade

[2]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 and only 430 less 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) and grade 10 (6,890 hours), respectively.

[3]An alternative would be to focus on the overall effect of the G8 reform, for example by comparing test scores of students in a double cohort at the end of their last high school grade (i.e., grade 13 pre-reform, and grade 12 post-reform). By then, these students have different age (G8 students are one year younger) but have received the same amount of instructional hours covering the same curriculum, although distributed over a different time span. Büttner and Thomsen (2015) and Dahmann (2015) – exploiting data on double cohorts from Saxony Anhalt and Baden-Württemberg, respectively – constitute examples of this type of analysis. Similarly, studies that focus on post-graduation outcomes are assessing the overall effect of the reform, rather than its higher learning intensity aspect alone.

[4]Exceptions are the states of Berlin and Brandenburg, where the transition to secondary school (tracking) takes place at the start of grade seven, as opposed to grade five.

[5]Some states also have comprehensive schools (*Gesamtschulen*), which combine the three basic secondary school types in one organizational unit offering multiple educational tracks. In addition, most states offer types of school that bring the lower tracks – i.e., basic- and middle-track – under one educational and organizational umbrella. These schools – classified for statistical purposes as *schularten mit mehrerhen bildungsgängen* (schools with multiple educational tracks) – take usually state-specific names (Lohmar and Eckhardt, 2010).

9 being the minimum attendance requirement in Germany. The highest level of secondary school is academic-track high school (*Gymnasium*), referred to as academic-track because only its successful completion leads to university entrance qualification (*Abitur*).

Up to 2001, the academic-track high school lasted nine years in almost all federal states, resulting in a total of thirteen years of schooling to graduate from high school and qualify for university entrance.[6] However, following a heated debate, and guided by the desire to speed up graduation and increase labor market participation of high school students, starting in 2001 most German states reduced the length of the academic-track by one year. As a consequence, in those states, the overall number of years of schooling required to complete academic-track high school has been reduced from thirteen to twelve.

Figure 1 offers a visual summary of the G8 reform implementation in the different federal states. Figure 1A displays the timing of the reform introduction, as well as the grades initially treated. Although in most states the G8 reform affected only students entering the academic-track – i.e., fifth-graders –, some states[7] extended its applicability to students that entered high school in previous years and currently attending later grades (up to grade nine). Figure 1B indicates the expected graduation year of the first treated cohort in each state. The latter is usually referred to as *double cohort* because it is expected to graduate at the same time as the last G9 cohort.[8] Figure 2 further adds a spatial dimension, displaying the geographical distribution over time of the G8 states. What emerges from the latter figure is that the timing of the G8 reform implementation does not seem to follow a geographical pattern, possibly related to economic and/or school conditions of particular German macro-regions (e.g., northern versus southern, or eastern vs. western states).

Despite the reduction in the length of academic-track high school, the overall curriculum and the academic requirements for obtaining the university entrance qualification were left unaltered. The minimum required instructional time as well as the length of the school year did not change either: At least 265 hours per week still had to be distributed over the remaining eight grades up to graduation. This implies that the number of weekly hours of instruction per grade, and the corresponding curriculum covered, have been increased in G8 compared to G9 academic-track high schools, leading to a higher learning intensity. While a uniform distribution of the overall instructional hours requirement across grades would imply an increase in the weekly number of instructional hours from

[6] Whereas since the Second World War the overall length of Gymnasium in the West German states has been thirteen years, it was set at twelve years in the former East German states. Following reunification, the former East German states – with the exception of Saxony and Thuringia – adapted to West German standards, increasing the overall schooling length to thirteen years.

[7] Saxony-Anhalt in 2003 and Bavaria, Mecklenburg-Vorpommern, and Lower Saxony in 2004.

[8] Students graduating in this cohort might have had particularly strong incentives to study given the increase in competition for admission to university degree programs or jobs. I address this issue in Section 7.

about 29 to 33 in each grade,[9] the actual allocation policy was left up to the federal states. Figure 2 – based on grade-level state-specific data (Kultusministerkonferenz, 1997-2011) – summarizes this distribution, comparing the average number of weekly instructional hours offered by grade under the new (G8) and the old (G9) regime. It reveals that middle grades (7 to 9) experienced the highest increase in additional workload. By contrast, the workload was left almost unaltered in lower grades (5 and 6), while the increase was lower in the upper grades (10 to 12).

Although, in line with its original purpose, the reform allows earlier graduation with the same level of qualification and earlier labor market participation, the debate surrounding the reform has been very controversial, both before and since its implementation. A major issue of the public debate concerns the question of whether it is possible to improve educational performance by increasing the learning intensity in high school. Based on fears that the increase in learning intensity will overburden students, thereby negatively affecting their educational achievement, some states have already announced, or implemented, a (partial) switch back to the old regime.[10]

3 Related Literature

Despite the abundant literature on the effects of various inputs into the education production function, evidence on the effect of instructional time is limited and sometimes conflicting. Several studies exploit between-country variation in instructional time providing a wide range of estimated effects on test scores, from no effect (Lee and Barro, 2001), to small positive and significant effects (Wössman, 2003), to larger positive and significant effects (Lavy, 2010; Rivkin and Schiman, 2013). Most of the recent literature focuses, however, on exploiting within-country, within-state, or within-district exogenous variation of the amount of instructional time arising from educational policies. By altering the amount and/or the timing of instruction and/or the curriculum to be covered, a particular educational policy can either decrease, keep constant, or increase the intensity of learning.

A branch of this literature focuses on analyzing the impact of policies that provide exogenous variation of instructional time by lengthening the school day or the school year (Bellei, 2009; Parinduri, 2014), by shifting state-mandated school start dates and/or test dates (Sims, 2008; Hansen, 2011; Agüero and Beleche, 2013; Carlsson, Dahl, and Rooth, 2015), or by reallocating instructional time to a specific subject (Cortes, Goodman, and

[9] These figures are obtained by dividing the total number of weekly instructional hours (265) by nine or eight grades, respectively.

[10] While a full reversion to the G9 regime has been announced in Lower Saxony, other states – i.e., Bavaria, Baden-Württemberg, Hesse, and North Rhine-Westphalia – have announced or already implemented a partial switch back to allow for a G9 option. Since these reversions do not affect the cohorts of students in my sample, they will not be further discussed here.

Nomi, 2015) while keeping the curriculum unaltered, thereby reducing the intensity of learning. These studies generally find positive (albeit sometimes small) and significant effects of instructional time on standardized test scores or on other educational or labor market outcomes (e.g., grade repetition, educational attainment, wages) (Parinduri, 2014).[11] These findings are consistent with the idea that students might benefit from a decrease in learning intensity, with the additional hours of instruction used by teachers to cover the same curricular content in more depth, i.e., with more opportunities for practice and review, and to provide additional support to slow learners.

Another strand of the literature has explored the benefits of an additional year of secondary schooling. Given that an additional year of schooling represents a direct increase of both instructional time and curriculum to be covered, there is no change in the intensity of learning. A number of studies (see Angrist and Krueger, 1991; Oreopoulos, 2006, among others) use either reforms (i.e., changes in compulsory schooling laws) or variables (e.g., quarter of birth) affecting the minimum legal number of years of schooling as an instrument, offering estimates that can be interpreted as the benefit of an extra year of school for potential dropouts. A few studies have also analyzed the benefits of an additional year for university-bound students. Morin (2013) and Krashinsky (2014) exploit a reform that reduced high school duration (by one year: from five to four years) as well as the corresponding curriculum in the Canadian province of Ontario, finding that the reform significantly lowered (between 2 and 5-8 percentage points, respectively) the university performance of the affected cohorts.[12]

By contrast, there is still a paucity of evidence on the effects of educational policies that lead to increased learning intensity. In a seminal study, Pischke (2007) analyze the effects of a reform that introduced an earlier start of the academic year in 1960s Germany by shortening two contiguous academic years (1966-1967). Similar to the G8 reform, the reform dramatically changed the amount of instructional time for some students in school at the time without directly affecting the curriculum, thereby increasing the learning intensity (i.e., the same curriculum had to be covered in a shorter time for the grades affected). This change increased grade repetition in primary school and lowered enrollment in academic tracks, but it had no adverse effect on earnings and employment outcomes of the affected cohorts.

A small but growing number of studies have recently started to contribute to this literature by exploiting the G8 reform. A first set of studies (Thile, Thomsen, and Bütt-

[11] A related literature exploits exogenous variation provided by natural events, i.e., changes in the number of school days missed due to inclement weather (Marcotte, 2007; Marcotte and Hemelt, 2008; Marcotte and Hansen, 2010), finding that more time in school before tests improves student performance on state-wide exams.

[12] This reform can be considered to some extent as the "reverse" of the typical compulsory schooling law change, and these findings interpreted as the value-added (in terms of university grades) of an extra year of high-school for university-bound students.

ner, 2014; Meyer and Thomsen, 2013, 2014; Büttner and Thomsen, 2015) exploit data on the double graduation cohort from Saxony-Anhalt, focusing on the overall effect of the G8 reform on graduation and/or post-graduation outcomes. While Thile, Thomsen, and Büttner (2014) do not find any significant reform effect on adolescent personality development, Meyer and Thomsen (2013) and Büttner and Thomsen (2015) find a significant delay in university enrollment among female students and a significant negative effect on final achievement in mathematics for both genders, respectively. However, Meyer and Thomsen (2014) do not find significant reform effects on university students' motivation and drop-out rates. Despite their valuable contributions, the nature of the data employed in these studies poses limits to their internal and external validity (i.e., their findings might be confounded by time-/state-specific factors and pure maturation effects, or might be driven by the increased competition over post-graduation resources arising in the double cohort).

A second set of studies use more representative data and an identification strategy that exploits the variation in the implementation of the G8 reform over time and across states. Dahmann and Anger (2014), Dahmann (2015), and Huebener and Marcus (2015) focus on pre-graduation and/or graduation outcomes. Using 17-year-olds samples from SOEP survey data, Dahmann and Anger (2014) and Dahmann (2015) find that the increased learning intensity introduced by the G8 reform affected some aspects of adolescent personality and improved boys' crystallized intelligence, respectively. Using administrative data, Huebener and Marcus (2015) find that the reform led to a significant increase of grade repetition rates only in grades ten to twelve, i.e. in the final three years prior to graduation, but did not affect the graduation rate.

By contrast, Dörsam and Lauber (2015) and Meyer, Thomsen, and Schneider (2015) analyze the overall effect of the G8 reform on post-graduation outcomes. Using university register data, Dörsam and Lauber (2015) find negative and zero effects on university grades for treated students belonging to the double cohorts of Bavaria and Baden-Württemberg, respectively, and positive effects – driven by female and higher-ability students – for the following cohorts of treated students in both states. Using data on high school graduates from all German states, Meyer, Thomsen, and Schneider (2015) find that the reform reduced university enrollment in the first year after high school graduation (and beyond for males), and increased the probability of spending one year abroad or in voluntary service.

4 Empirical strategy

4.1 Identification strategy

Under certain conditions, the G8 reform allows a quasi-random assignment of academic-track high school students to a treatment and a control group. Students in the control

group are those who entered academic-track high school prior to the reform or who attended grades not affected by the reform when the latter was extended to higher grades. These students would graduate after nine years and would receive a total of 265 weekly hours of instruction distributed over nine grades. By contrast, the treatment group consists of students who entered academic-track high school after the implementation of the reform or, where the reform applicability was broadened to higher grades, students who entered high school previously and were currently attending those grades. These students would graduate after only eight years and, by the end of high school, receive the same total number of weekly instructional hours covering the same curricular content, although compressed into a shorter time.

The staggered implementation (over time and across states) of the reform is exploited for identification purposes. Student cohorts attending high school in a treated state experienced an increase in learning intensity as compared to previous cohorts that were not affected by the reform in the same state, or compared to cohorts attending high school in other control states. Time and state variation thus make it possible to isolate the effect of the reform from other confounding factors, within a difference-in-differences (DD) approach.[13]

My main DD model is captured by the equation:

$$zscore_{ist} = \beta_0 + \beta_1 G8_{st} + \alpha X_{ist} + \delta_s + \gamma_t + \varepsilon_{ist}, \tag{1}$$

where $zscore_{ist}$ is the (standardized) PISA reading, math, or science score measured in year t for an academic-track student i in state s. $G8_{st}$ is the G8 reform indicator which equals one if a student observed in year t and in state s belongs to the cohort treated by the G8 reform in that state, and zero otherwise. This is my main variable of interest, as its coefficient β_1 measures the impact of the reform on the treated group after covariates adjustment. X_{ist} is a vector of student-, school-, and state-level variables. δ_s and γ_t represent state and time fixed effects, respectively. The state (time) fixed effects control for unobserved factors that differ across states and not over time (over time and not across states). ε_{ist} is an individual-specific error term.

Equation (1) represents the main specification of the DD model employed to estimate the G8 reform effects. However, to account for state-specific increases in learning intensity introduced by the G8 reform, I also estimate a slightly different version of equation (1):

$$zscore_{ist} = \beta_0 + \beta_1 Hours_{st} + \alpha X_{ist} + \delta_s + \gamma_t + \varepsilon_{ist}, \tag{2}$$

[13]A drawback of the DD approach is that it does not control for state-specific stocks, which might similarly affect all students in a state, for example, due to changes in primary school. This issue is addressed in Section 7. Among the multiple robustness checks carried out, I exploit a third source of exogenous variation offered by the fact that only academic-track students, but not middle-track students, were exposed to the G8 reform. Middle-track students can therefore be used as an additional control group in a difference-in difference-in-differences (DDD) approach.

where $Hours_{st}$ is a state- and time-varying variable indicating the total number of weekly instructional hours provided in high school grades 5 (or higher) to 9.

4.2 Treatment definition

Tables 1 and 2 are used to define the G8 treatment status of the PISA cohorts included in my dataset for the purpose of estimating the impact of the G8 reform on standardized test scores.[14] Table 1 displays the timing of G8 adoption in the different federal states (column 1), the grades initially treated (column 2), and the year of academic-track enrollment (tracking year) for the cohorts attending those grades (column 3). Reported in bold in columns 2 and 3 are those grades and tracking years that are used – together with the relevant tracking calendars displayed in Table 2 – to define the treatment status (i.e., the $G8_{st}$ treatment dummy included in equation (1)) of the 2000-2009 PISA cohorts of ninth-graders included in my sample. The treatment status (T for treatment, C for control) of each PISA cohort is displayed in columns 4 (PISA 2000) to 7 (PISA 2009).[15] The former East states of Saxony and Thuringia are excluded from the main sample because they kept the G8 regime after reunification.[16]

The treatment assignment is problematic in the case of Hesse, where the G8 reform was introduced for the cohort of 2004 fifth-graders only in 10% of the academic-track schools. Given the low probability of treatment assignment, I keep Hesse in the sample assuming that the PISA 2009 cohort of ninth-graders – tracked in 2004 – was not affected by the G8 reform.[17] Furthermore, the PISA 2006 cohorts of academic-track ninth graders in the states of Saxony-Anhalt and Mecklenburg-Vorpommern were already in grade 7 and grade 8, respectively, when they were assigned to G8. As a consequence, the treatment they were exposed to in the following grades was different than the standard: It was shorter in length, but higher in intensity, given that the instructional hours previously delivered in the last G9 grade (grade 13) had to be distributed over fewer remaining grades (from 7 (8) to 12). Although I consider these cohorts as treated in equation (1), in Section 7 I also test the robustness of my results to the exclusion of these states.

[14] A different treatment definition is used for the high school grade retention analysis, as explained in Section 8.

[15] For example, in Bavaria the reform – introduced in 2004 – affected high school students attending grades five and six. As shown in Table 2, ninth-graders from Bavaria assessed in PISA 2009 entered high school in 2004 and were therefore treated by the G8 reform. By contrast, those assessed in earlier PISA cycles entered high school before 2003 and are therefore assigned to the control group.

[16] Essentially, the treatment effect could not be separately identified for these states because they are always treated – i.e., never switched to treatment – during my observation period. However, if the assumption that the treatment leads only to a jump (level difference) but not to a change in test score trends (slope difference) is valid, once the level difference is captured by state fixed effects, these states can also serve as additional controls. In Section 7, I assess the robustness of my results to the inclusion of Saxony and Thuringia as control states.

[17] However, excluding the state of Hesse from the estimation sample does not affect my results. These results are available upon request.

The reallocation of instructional hours across grades following the G8 reform was left to the states. According to Figure 3, the additional G8 workload is highest in grades 7 to 9. In order to capture state- and grade-specific increases in learning intensity, I construct three variables indicating the cumulative number of weekly instructional hours delivered in high school – in grades 5 to 9, 7 to 9, and 8 to 9, respectively – that will then be alternatively used in estimating equation (2).[18] These variables are derived from a dataset including the state-specific number of weekly instructional hours by grade, school type, and school year,[19] and are then merged by state and year to the PISA pooled dataset.

4.3 Threats to identification

Several potential threats to internal validity arise when estimating the DD models described in Section 4.1. The key identifying assumption is, however, that, in the absence of treatment, the difference in outcomes between treatment and comparison groups is constant over time (common trends assumption). Accordingly, a disadvantage of my identification strategy is that any state-specific shock contemporaneous to the G8 reform will bias my estimates. I address this concern in a number of ways.

First, although the common trends assumption is not directly testable, I exploit the availability of data for two (or three) cohorts of ninth-graders that entered high school before the policy change to enhance my regression results with graphical evidence on how PISA test scores deviate in treated and control states with the advent of the G8 reform, compared with their pre-reform trends. Figure 4 allows inspection of the common trends assumption for states that switched to treatment at different times, by separately comparing test score trends in control states[20] and: i) in states that switched to treatment in 2006[21] (4A, 4B, and 4C); ii) in states whose PISA cohorts were treated in 2009[22] (4D, 4E, and 4F). Figure 4 suggests that pre-reform trends are quite similar for the treatment and comparison groups, but that there are significant changes in their relative outcomes

[18] There are several reasons for the use of different measures of learning intensity. First, in some states – i.e., Berlin and Brandenburg – high school tracking takes place in grade 7 rather than in grade 5. Furthermore, in other states – i.e., Saxony-Anhalt and Mecklenburg-Vorpommern – the first G8 cohorts experienced a higher treatment intensity, although limited to grades 7 (8) to 9. Finally, the state- and grade-specific weekly number of instructional hours scheduled in 1995 and 1996 – when the PISA 2000 cohorts of ninth-graders were in fifth and sixth grade, respectively – is not available, and has to be retrieved from the 1997 school year. It is therefore important to assess the sensitivity of the estimates to the use of these different, although partially overlapping, measures of learning intensity.

[19] The dataset contains the figures provided in the *Wochenplichtstunden der Schüler nach Schularten und Ländern. Grundstuden Im Schuljahr 1997/1998 - 2011/2012* series (Kultusministerkonferenz, 1997-2011).

[20] As illustrated in Table 1, control states are those that did not switch to treatment during my observation period (i.e., Hesse, North Rhein-Westfalia, Rheinland-Palatinate, and Schleswig-Holstein).

[21] Mecklenburg-Vorpommern, Saarland, and Saxony-Anhalt.

[22] Baden-Württemberg, Bavaria, Berlin, Brandenburg, Bremen, Hamburg, and Lower Saxony.

after the reform kicked in. Furthermore, the graphical evidence is particularly convincing when the treatment group is limited to states that switched to treatment in the last available period (2009), and for which a longer pre-reform period is available.

More generally, the plausibility of common trends rests on the choice of a comparable control group: In the following Section, I analyze pre-reform differences in observables among treatment and control states, as well as compositional changes over time. Finally, to increase the confidence in my identification strategy, and to make sure that my results are not just picking up long-running trends in differences between treated and control states, I run a battery of specification checks and falsification tests. The robustness analysis is presented in Section 7, after the main results.

5 Data

5.1 Data and sample restrictions

The empirical analysis is based on data from the first four PISA cycles (2000, 2003, 2006, and 2009), as well as on data from administrative sources.[23]

While PISA is conducted by the OECD in a number of countries sampling 15-year-old students, independent of grade, national extensions of the study (PISA-E) were conducted in Germany in 2000, 2003, and 2006. About 45-50,000 students were assessed in each PISA-E cycle, with the original PISA samples enlarged by the addition of grade (9) and age (15)-based samples. The aim of these national extensions was to provide a sample large enough to allow comparisons between the different German federal states. PISA-E was discontinued in 2009. However, PISA 2009 was also enlarged with a grade (9)-based sample. Although the latter extension is smaller than the PISA-E samples it still represents a large sample, with about 9,500 ninth-graders assessed, and one that maintains representativeness of the federal states because of its stratified design by state and school type. Moreover, it represents the only PISA 2009 sample that can be used for the purpose of this study.[24] The empirical analysis undertaken in this study is therefore based on a dataset that pools grade-9 samples from PISA-E 2000, 2003, 2006, and PISA 2009.

The main sample includes all ninth-graders enrolled in academic-track high schools, with a valid test score assessment.[25] PISA tests cover three different domains (reading,

[23]Either provided by the *Kulturministerkonferenz* (The Standing Conference of the Ministers of Education and Cultural Affairs of the Federal States) or by the *Statistische Ämter des Bundes und der Länder* (the Statistical Office of the Federal States).

[24]Due to confidentiality agreements, the standard PISA 2009 age-15 sample released by OECD does not provide identifiers for German federal states, and therefore cannot be used in this analysis.

[25]Grade- and age-based samples are largely overlapping, as both include 15-year-old ninth-graders – i.e., in principle, students that enrolled in primary school in the year they turned 6 and did not experience grade retention. Grade-9 samples further include students younger or older than 15. By contrast, age-15 samples also include students in grades lower or higher than the ninth. While studies that focus

mathematics, and science), assessing a range of relevant skills and competencies that should reflect how well young adults are prepared to analyze, reason, and communicate their ideas effectively.[26] Each PISA domain is tested using a broad sample of tasks with differing levels of difficulty to represent a coherent and comprehensive indicator of the continuum of students' abilities.[27]

An issue related to the pooled nature of my data regards the comparability of the student performance measures defined in the different PISA assessments. Reading is indeed the only domain whose assessment is directly comparable across all four cycles. This is because reading was the major domain in 2000, and all subsequent reading assessments were measured on the same scale until 2009, when reading was again the major domain. By contrast, mathematics and science were the main domains in 2003 and 2006, respectively, and between 2000 and 2003 (2006) the mathematics (science) test underwent major revisions. However, under the plausible assumption that the degree to which the assessments differ is orthogonal to the timing of the introduction of the G8 reform, the DD estimator employed in this study – which is not a simple before-after estimator of the treated students, but also takes into account the time trend in the control group – does not require comparability of assessments across cycles. I therefore consider PISA reading, mathematics, and science (standardized) test scores over the period 2000-2009 as my main outcome variables.

5.2 Control variables

Three groups of variables, defined at the student-, school, and state-level, are employed as controls in the empirical analysis. They capture between-states compositional differences that may be correlated with G8 adoption.[28] Student-level controls include a set

on international comparisons should be based on age-based samples, to avoid possible distortions that country-specific entry ages and grade-repetition rules might cause in grade-based samples (Fuchs and Wössmann, 2007), this is not necessarily the case for a within-country analysis. For Germany, for example, in the period under study the federal states had the same entry-age and grade-repetition rules. Accordingly, a possible increase of grade retention induced by the G8 reform should be spread similarly across grade (in age-based samples) and age (in grade-based samples). If the reform increased grade repetition, the student age distribution in a grade-based sample would have more mass at higher ages. By contrast, in an age-based sample, the student grade distribution would have more mass at lower grades. Given the small fraction of grade repeaters in the student population, grade- and age-based samples are therefore expected to deliver qualitatively similar results.

[26] PISA tests are paper-and-pencil tests lasting a total of two hours for each student. Test items include both multiple-choice items and questions requiring the students to formulate their own responses.

[27] Using item response theory, PISA maps performance in each subject on a scale with an international mean of 500 and a standard deviation of 100 test-score points across the OECD countries included in the study. The scores are averages of plausible values, which are drawn from a distribution of values that a student with the given amount of correct answers could achieve as a test score. See OECD (2012).

[28] I am aware that variables that may be endogenous to the treatment are not necessarily good controls, as they may capture part of or bias the treatment effect. Nonetheless, I find interesting to validate my empirical strategy by assessing its robustness to additional specifications.

of demographic and socio-economic characteristics. Among the demographic characteristics, besides a dummy indicating female students and a quadratic age term (in months) that controls for potential age/maturation effects, a grade retention dummy is included to control for different schooling experiences.

The socio-economic characteristics include an indicator for the number of books at home, two indicators for parents' highest educational level (ISCED), as well as the Highest International Socio-Economic Index (HISEI), which uses the higher of the two parents' ISEI scores or the only available parent's ISEI score.[29] I also derive controls for students' migration background, namely three dummy variables indicating whether the student was born in Germany, whether she speaks German at home, and whether at least one of her parents was born in Germany.[30]

School-level controls include the total number of enrolled students, the proportion of girls enrolled, the student-teacher ratio, the percentage of government funding received, as well as dummy variables indicating urban schools – i.e., schools located in a community of more than 100,000 inhabitants – and privately run schools. Moreover, although PISA does not provide objective measures of the school financial situation, school resources are proxied by the school principals' subjective assessments of whether a lack of instructional material or a lack of computers hindered instruction at their school.

State-level controls include additional factors that characterize the school system and the economic environment in which the students obtained their education, and that are potentially related to the G8 reform: the share of academic-track schools and the share of students enrolled (overall, as well as in the seventh grade) in these schools, the share of academic-track all-day schools and the share of students enrolled in these schools, and the GDP per capita.

5.3 Descriptive statistics

The main sample includes about 26,500 academic-track high school students whose skills were assessed by PISA at some point during the period 2000-2009. Table 3 displays means of outcome and control variables for states that introduced G8 at some point during my observation period, as well as treated-control states' mean differences by PISA cycle, were PISA 2000 and 2003 cohorts are pre-reform periods in all states. My identification strategy relies on comparing the change in PISA scores before and after the reform for ninth-graders attending academic-track high schools in treatment and comparison states. While there are generally positive but statistically insignificant differences in test scores

[29] In each PISA cycle, parents' occupational data were obtained by asking open-ended questions. The responses were coded to four-digit ISCO codes and then mapped to the International Socio-Economic Index of Occupational Status. Higher ISEI scores indicate higher levels of occupational status.

[30] Other important background variables such as whether the student attended pre-primary education or currently lives in a single-parent household are not available in all cycles, and cannot therefore be controlled for.

between treatment and comparison groups for pre-reform cohorts and for the first post-reform period (2006), these differences widen substantially in the last available post-reform period (2009), becoming statistically significant as well. In a DD framework, this pattern is suggestive of significant effects of the G8 reform on treatment states' student outcomes.

Furthermore, significant pre-reform observable differences or, most importantly, substantial changes over time in observable differences might call my empirical strategy into question by suggesting unobserved compositional pre-reform differences or changes over time, respectively. Table 3 shows instead that treatment and comparison groups have similar characteristics across PISA cohorts, and, most importantly, that changes over time in the relative characteristics of the two groups are quite small. This provides support for the validity of my identification strategy, which relies on the assumption that the groups systematically differ only in the treatment assignment. Nonetheless, to further address these compositional concerns, the empirical analysis focuses on a main specification that controls for student, school, and state characteristics.

6 Main Results

The results of estimating equations (1) and (2) under the baseline specification – i.e., including only state and year fixed effects – are reported in column (1) of Tables 4 and 5 for reading (panel A), math (panel B), and science (panel C), respectively. I use plausible values in all analyses that involve test scores,[31] constructing my dependent variable as the average of five (standardized) plausible values provided for each domain and PISA cycle.[32] Standard errors are clustered on the state level (rather than on the state-year level) to avoid the assumption of within-cluster time-independent errors (Bertrand, Duflo, and Mullainathan, 2004).[33] In all instances, final sample weights are used to take into account the complex survey nature of PISA data, where schools are sampled within strata, and students are then sampled within schools according to an age- and/or grade-based criteria (OECD, 2012).

The baseline specification results obtained estimating equation (1) – reported in co-

[31] Plausible values are imputed values that resemble individual test scores and have approximately the same distribution as the latent trait being measured. They represent random draws from an empirically derived distribution of proficiency values that are conditional on the observed values of the assessment items and the background variables. See OECD (2012).

[32] The estimation procedure recommended by OECD (OECD, 2012) involves the calculation of the required statistic five times, one for each set of PISA plausible values. The final estimate is the arithmetic average of the five estimates. In an OLS setting, this is equivalent to estimating a regression with the average of (standardized) plausible values as dependent variable.

[33] Although this approach may lead to over-rejection of the null hypotheses when the number of clusters is small (Cameron and Miller, 2015), this does not seem to be an issue in my setting: The p-values computed with state-level clustering or applying the wild t-bootstrap procedure proposed by Cameron, Gelbach, and Miller (2008) on state clustered data – reported in Table A.1 – produce similar inferential results.

lumn (1) of Table 4 – indicate that the G8 reform had a positive and significant effect on reading, math, and science achievements of academic-track ninth-graders in treated states. In those states, the reform increased PISA standardized scores of the same order of magnitude (within a 0.12-0.138 standard deviations range). The results obtained estimating the baseline version of equation (2) for different treatment intensities – reported in column (1) of Table 5 – are in line with the former results: A twenty hour increase distributed over grades 5-9, or a ten hour increase distributed over grades 8-9, led to an average increase in student achievement of 0.08-0.14 standard deviations, respectively, depending on the subject.

If the assignment to increased learning intensity was quasi-random, estimating equations (1) and (2) under the baseline specification would be sufficient to identify the causal effects of the G8 reform. In this situation, there would be no need to add further controls to the baseline specification for identification purposes, given that the distribution of treatment and control groups' characteristics would be fully balanced. However, the exogenous variation offered by natural experiments is not always as clean as that provided by randomized experiments. Including student-, school-, and state-level variables may therefore help to control for confounding trends. On these premises, extending the baseline specification to include further controls also represents an important specification check.

To adjust for compositional changes over time between the observations in the different groups, I progressively add three sets of control variables to the baseline specification.[34] The results obtained estimating equations (1) and (2) under these enlarged specifications are presented in columns (2) to (4) of Tables 4 and 5, respectively. Specification (2) adds student-level characteristics (demographics and socio-economic background) to the baseline specification in column (1). Specification (3) further adds school-level controls, to account for support measures that might have been implemented at the school level following the reform (e. g., the recruitment of more teachers, which is captured by the student-teacher ratio). Finally, specification (4) also includes state-level controls. The purpose of the latter specification is to gauge the sensitivity of my estimates to state-specific school and economic conditions.

Overall, the parameter estimates of the reform effect are very similar in all the specifications. This implicitly validates the use of the G8 reform as a quasi-natural experiment, as student-, school-, and state-level characteristics that may be correlated with student achievement do not appear to be correlated with the reform, and their omission would

[34] As with any survey data set, the samples collected in each cycle contain missing values in some background variables. Given that, for most of them, the missing rate is relatively low in the pooled sample (below 5 percent), this issue is addressed in the empirical analysis by recoding the missing values to zero and including in the estimated models dummy variables indicating the presence of missing values in each of the affected variables when the latter are included in the specification. In any case, the empirical analysis conducted after dropping missing values on the relevant background variables leads to similar results, available upon request from the author.

not significantly bias its baseline estimated impact. Nonetheless, as the reform effects obtained estimating specification (4) are somewhat smaller than the ones obtained from the baseline specification and in order to improve the precision of my estimates, I use the former as the main specification to conduct the remaining empirical analysis.[35]

7 Robustness

In this section, I assess the sensitivity of the main results to multiple robustness checks demonstrating that the effects of the increased learning intensity introduced by the G8 reform is very similar across different specifications and samples, and qualitatively the same. The results are reported in Tables 6 and 7. For comparability purposes, both tables report in column (1) the results obtained estimating equation (1) in its main specification, i.e., including student-, school-, and state-level controls. The robustness checks on the main specification are reported in columns (2) to (10) of Table 6 and in columns (2) to (7) of Table 7.[36]

7.1 Falsification tests

My DD approach identifies the G8 reform effects under the assumption of a common time trend in treatment and comparison groups in the absence of the reform, i.e., there are no unobserved variables that change over time resulting in differential effects on test scores of students that were treated by the G8 reform and students that were not. Equivalently, the treatment must be the only reason why treatment and control group trends deviate in the post-reform period. The main concern is therefore that the reform effects reflect differential time trends in the outcomes of interest between treatment and comparison states, rather than a true policy impact. This might be the case if, for example, another policy reform was implemented during my observation period, affecting treatment and control groups differently. To the best of my knowledge, no such policy occurred in the post-reform period.

While a direct test of the common trends assumption is not possible, given the unobservability of the treatment counterfactual, graphical and regression based evidence might be used to corroborate its validity. A simple way to enhance the graphical evidence already presented in Section 4.1 by partially testing the plausibility of the common trends assumption is a placebo treatment test in the years preceding the actual treatment that can show deviations from the common trend in pre-treatment years. I run this test on the

[35]The empirical analysis based on estimating baseline models leads to similar results, available upon request.

[36]The G8 policy effect estimates reported in Table 6 refer to the $G8_{st}$ dummy in the models estimated in columns (1), (3), (4), and (7) - (10), and to the relevant interaction terms in the models estimated in the other columns, respectively, as described in this Section.

2000-2003 PISA sample, with 2003 considered as the treatment period, by estimating a model that includes a G8 reform dummy, a post-reform dummy, and an interaction term. The coefficient estimated on the latter – column (2) of Table 6 – represents the G8 policy effect. Consistent with the graphical evidence, it turns out to be insignificant. Furthermore, I provide an additional placebo test, based on the idea that the achievement of basic- and middle-track students in treated states should not be significantly affected by the G8 reform, as they were not directly exposed to it. The insignificance of the G8 dummy estimated coefficients – in column (3) – also confirms this expectation. Taken together, the evidence proceeding from these falsification tests, as well as from earlier visual inspections of treatment/comparison groups trends and analysis of compositional differences between groups over time, corroborates the validity of the common trends assumption.

7.2 State-specific linear trends

Besley and Burgess (2004) show that allowing for differential time trends in a DD regression may destroy otherwise large and statistically significant treatment effects. Column (4) reports the results obtained when the main specification is augmented by state-specific linear time trends. The idea is to use the pre-reform data to extrapolate the time trend of each state into the post-reform periods. This allows treatment and comparison states to follow different secular trends in a limited but potentially revealing way.[37] Despite being estimated less precisely, as I am now exploiting deviations from pre-existing state trends to pin down the G8 reform effects, the results strongly support the picture provided by my main specification, in terms of both the economic magnitude and statistical significance of the reform effects.

7.3 Difference-in-difference-in-differences

As an alternative way to control for both state-specific trends and regional shocks potentially correlated with the G8 policy, I exploit the fact that the latter was implemented at different points in time across different states and affected academic-track students but not middle-track students. Adding middle-track students as an additional control group leads to a difference-in-difference-in-differences (DDD) model that makes use of the outcome change of middle-track students to control for state-specific shocks potentially correlated with the policy.[38]

[37] See Angrist and Pischke (2009).

[38] Besley and Case (2000) discuss the conditions under which DD and DDD estimators deliver unbiased estimates, emphasizing that the latter are crucially dependent on the quality of the control group chosen. In the German three-track educational system, middle-track students represent, among the students that were not affected by the G8 reform, the group that is most closely comparable to academic-track students. It seems therefore plausible to assume that academic- and middle-track students are comparable, i. e. respond similarly to state-specific shocks. Although this is an untestable assumption, the similarity of DD and DDD estimates and the DDD falsification test carried out in what follows supports this assumption,

The model is captured by the following baseline equation:

$$zscore_{iast} = \beta_0 + \beta_1 G8_{st} + \beta_2 Atrack_{ist} + \beta_3 G8_{st} \times Atrack_{ist} + \delta_{sa} + \gamma_{ta} + \lambda_{st} + \varepsilon_{iast}, \quad (3)$$

where s indexes state, t indexes time, and a indexes track. $Atrack_{ist}$ is a dummy taking the value 1 for academic-track students in state s and time t, and 0 for middle-track students. The parameters δ_{sa}, γ_{ta}, and λ_{st} are, respectively, state-by-track, time-by-track, and state-by-time fixed effects.[39] The state-by-track effects account for state-specific factors that vary across tracks but are fixed over time. These include, for example, fixed-differences across states in terms of educational policies and local labor market opportunities. The time-by-track effects account for time varying and track-specific factors that are common across states. The state-by-time effects account for time-varying state-specific factors that have a common effect across tracks. In its main specification, the model also includes a vector of student-, school-, and state-level controls, as well as its interaction with the academic-track dummy. The coefficient β_3 represents the impact of the G8 reform on the achievement of academic-track students versus middle-track students in treated states relative to control states.

The results obtained estimating equation (3) – in column (5) – confirm my main findings, both in terms of the economic and statistical significance of the G8 policy effects. Furthermore, the latter have the same order of magnitude – within a 0.11-0.14 standard deviation range, depending on the subject – then the ones estimated under the main specification. The similarity of the DD and DDD results indicates that test score trends in the treated states are similar to those in the control states.

The interpretation of the DDD coefficient as a causal effect of the reform relies on a weaker assumption: in the absence of the reform, the difference in outcomes between academic- and middle-track students would have developed similarly in treated and control states. Nonetheless, as this assumption is not testable, I carry out an additional placebo test using a period in which the G8 reform was not yet affecting my cohorts. Similar to the first DD placebo test, I pretend that academic-track ninth-graders were treated in all states in 2003. The estimation sample is limited in this design to the PISA 2000-2003 pooled data, and the model includes an interaction term between the reform dummy, a post-reform dummy, and the academic-track dummy. The coefficients estimated on the latter term – in column (6) – are insignificant for all subjects, suggesting that the policy effects estimated with DDD are not confounded by systematic differences in trends between treatment and comparison groups.

---

lending further credibility to my DD results. Similar results, available upon request, are obtained when including also basic-track students in the additional control group.

[39] Note that state and time fixed effects included in the DD model are now absorbed by the vector of state-specific time effects, λ_{st}.

7.4 Selection and non-compliance issues

Further internal validity threats may arise from self-selection or from non-compliance issues. First, the reform could have changed the distribution of students across school types within a German state in response to the introduction of the reform. Weaker students that would have enrolled in the academic track offered by comprehensive schools could easily avoid the reform by switching to a lower track within the same school. Or, weaker students that would have enrolled in the academic track in the pre-reform period might rather prefer to enroll in other secondary schools (either lower tracks or comprehensive schools) after the reform. In both cases, I might find a positive reform effect even if the reform had no direct effect on student achievement. While the former case was addressed at the sample selection stage, excluding from the sample students enrolled in comprehensive schools, I address the latter possibility estimating equation (1) with academic-track attendance (vs. attendance of other types of secondary schools) as dependent variable. The statistical and economic insignificance of the G8 reform coefficient – column (7) – suggests that the reform effects do not proceed from a change in the distribution of students across school types.

Moreover, since the reform was introduced in an entire state at one time, avoiding the reform while staying in the academic track – i. e., self-selecting into the control group – would require moving to a different state, an unlikely possibility considering the high costs associated with residential mobility. A more plausible scenario is that students from treated states living at the border of control states would avoid the reform by attending high school in the control states. While PISA data do not offer information on student residence, it is likely that the number of cross-border commuters is very small. However, to investigate this possibility I add to my main specification a dummy indicating those states that, at the time when they implemented the reform, were bordered by states that still offered the G9 regime,[40] as well as its interaction term with the G8 reform dummy. The G8 policy effects – reported in column (8) – are robust to this specification.

Further concerns arise from non-compliance to the treatment that might have affected those states where the G8 reform was announced – and therefore anticipated – before its implementation. Although in principle students in the first G8 cohorts or in the last G9 cohort might have tried to switch to G9 – by skipping a grade – or to G8 – by voluntary repeating a grade, respectively, it is very unlikely that they actually did so, as in either case they would end up graduating in their original cohort. Moreover, these concerns only apply to students belonging to the double cohort. In the next subsection, I assess whether my results are driven by the peculiarity of this cohort.

7.5 First G8 (double) cohorts

The first cohort that experienced the G8 regime in each state was considered part of a *double graduating cohort* because it was expected to graduate at the same time as the last cohort graduating under the G9 regime. Each double graduating cohort was approximately twice as

[40] These include Berlin, Brandenburg, Bremen, Hesse, North Rhine-Westfalia, Rhineland-Palatinate, and Saxony-Anhalt.

large as earlier or later cohorts[41] and was therefore subject to much stronger competition for post-graduation resources (jobs, admission to university degree programs, etc.).[42] Anecdotal evidence says that parents were worried about the consequences on the future academic and labor market outcomes of their children possibly deriving from this increased competition. At the same time, it might be that students experienced this increasing pressure as an incentive to work harder. This increased competition/pressure should not be a major cause of concern in my setting, given that the first treated cohorts assessed by PISA at the end of their ninth grade are still three years apart from graduation. Nonetheless, it is interesting to check whether my results are driven by these cohorts. To this end, I add to my main specification a dummy variable that equals one if a student belongs to the first G8 cohort in her state. The results of estimating the model under this specification – in column (9) – indicate that the reform effects for treated students that are not in the first G8 cohorts, although estimated less precisely, preserve their economical and statistical significance. This suggests that, although my results should still be interpreted as short-run effects of the increased learning intensity introduced by the G8 reform, they are not driven by peculiarities pertaining to the first treated cohorts.

7.6 Centralized Exit Examinations

I also want to check whether other educational reforms taking place during my observation period could be driving the results. A reform of the German high school system that has received considerable attention in the recent literature is the existence of Centralized Exit Examinations (CEEs).[43] While CEEs were introduced long before the start of my observation period in some federal states, most of the remaining states introduced CEE between 2005 and 2008. Since my empirical strategy exploits variation over time and state, other policy changes occurring at different times – like the introduction of CEEs, which targeted high school student populations statewide rather than specific cohorts – should not prevent identification of the G8 reform effects. Moreover, Jürges, Schneider, Senkbeil, and Carstensen (2012) provide evidence that CEEs do not matter significantly either for students in academic-track or for literacy tests like the ones analyzed in this study. However, it may still be the case that the introduction of CEEs affected students exposed and not exposed to the G8 reform in different ways. To assess this possibility, I add to the main specification a dummy indicating those states that introduced CEEs during my observation period[44] as well as its interaction with the G8 reform dummy. The results of estimating this model – in column (10) – suggest that the G8 effects on reading and science scores preserve their economic and statistical significance in states that did not introduce CEE

[41] Exceptions are the states of Hesse, where the reform implementation was staggered across schools over a three year period, and Rhineland Palatinate, where the G8 has been implemented only in selected schools so far. Both states are in the control group states during my observation period.

[42] See Morin (2015a) and Morin (2015b) for an analysis of the effects of the increased competition arising from the Ontario's double cohort on the earnings of high school graduates, and on university grades, respectively.

[43] See, among others, Jürges, Schneider, Senkbeil, and Carstensen (2012).

[44] These include Berlin, Brandenburg, Bremen, Hamburg, Hesse, Lower Saxony, North Rhine-Westfalia, and Schleswig Holstein.

during my observation period. By contrast, the effects on mathematics scores are reduced in size, and estimated less precisely.

7.7 Further robustness analysis

I conclude this Section by reporting, in Table 7, the results obtained estimating equation (1) on different samples. This additional exercise has two main purposes. On the one hand, it aims at assessing how robust my results are to a smaller treatment group or to a larger control group. On the other hand, it aims at gauging their sensitivity to more restrictive sample selection criteria.

The picture emerging from Table 7 indicates that the qualitative nature of my main results is not significantly affected by the timing of the treatment – 2006, in column (2), vs. 2009, in column (3) –, by the exclusion from the treatment group of the states whose cohorts were at some point exposed to a different treatment intensity – column (4) –, or by the inclusion of additional control states, i.e. those that were already adopting G8 at the start of the observation period – column (5). Similarly, when excluding students with a grade retention episode – column (6) – or whose birth year does not correspond to the cut-off years of their cohort[45] – column (7) –, the estimated effects preserve their order of magnitude and statistical significance.

8 Further results

8.1 Heterogeneity

The estimates reported in Table 4 show the average effects of the reform for the overall population of academic-track ninth-graders, indicating that treated students tend to score significantly better in reading, math, or science tests. However, students' characteristics – such as gender, parental education and migration background, and ability – may affect their capacity to deal with the increased learning intensity introduced by the G8 reform. To shed further light on the effects of the reform, I estimate additional specifications that explore possible heterogeneous policy effects by adding to the main specification an interaction term between each of the categories considered (gender, parental education and migration background, and grade retention) and the G8 dummy. The coefficients estimated on the reform dummy and on its interaction with the category considered are reported in columns (1) to (4) of Table 8.

[45] In Germany, children enroll in primary school in the year they turn six according to a cut-off rule defined by the month of birth. Although state-specific cut-off dates have been recently introduced, for the cohorts employed in this study the cut-off date was June 30 in the states of the former West Germany and May 31 in the former East. Children who turned 6 on or before June 30 (May 31) were admitted to primary school in that school year, while those turning 6 after June 30 (May 31) were admitted to primary school one year later. However, there is some flexibility embedded in this rule. Before children are admitted to primary school, they have to pass a basic maturity test. Children who are old enough to enter school but do not pass this test are admitted to primary school one year later (late enrollment). Children who are born after the cut-off date (but before December 31) may be admitted to school upon parental request provided they pass the maturity test (early enrollment). Therefore, rather than excluding students whose birth year and month are outside the range provided by the standard cut-off rule, I exclude only students whose birth year does not correspond to the cut-off years of their cohort (flexible cutoff). Similar results – available upon request – are obtained when the sample is selected based on the standard cut-off date.

The first distinction I consider is gender. As a consequence of behavioral and developmental diversity, boys and girls of the same age may have responded differently to the increased learning intensity introduced by the G8 reform. In particular, girls may have developed a wider set of non-cognitive skills – i.e., attitudes, behaviors, and strategies such as motivation, perseverance, and self-control – that might allow them a better adaptation to the new learning environment (Spinath, Eckert, and Steinmayr, 2014). The results – reported in column (1) – partially confirm this hypothesis, suggesting that the positive and significant effect of the increased learning intensity on standardized reading scores is mostly driven by girls. Given that girls in my sample outperform boys in terms of reading skills even before the introduction of the reform, this finding is consistent with the hypothesis that the effects of a more intensive instruction are heterogeneous based on initial skill differences, with students equipped with higher existing (reading) skills benefiting from higher returns on the latter (Cuhna and Heckman, 2007). By contrast, I do not find evidence of heterogeneous reform effects by gender in math or in science, where boys tend to outperform girls.

Further distinctions are by parental education and migration background. The performance of students with less educated parents, or with migrant parents, might have been negatively affected by the reform, possibly because of a lack of parental support in dealing with the increased learning intensity. However, it may also be the case that those same students benefited from longer school days and/or from increased support from their peer groups. The results – reported in column (2) and (3), respectively – provide little evidence that the reform significantly enlarged inequality arising from socio-economic or migration background.

Finally, in column (4) I consider heterogeneous reform effects by grade retention. The latter can be viewed as a low-ability proxy. Low-ability students are particularly vulnerable to the reform as they are most at risk of experiencing difficulties in adjusting to the new learning environment. It is reasonable to expect the effects of a more intensive instruction to be heterogeneous based on initial skill differences, with the most harmful – or less beneficial – effects on the students with lower existing skills, i.e., those that benefit from lower returns on the existing skills (Cuhna and Heckman, 2007). The evidence is consistent with this hypothesis: The estimated differential reform effect for low-ability students is negative and significant in all subjects, suggesting that the average reform effects are essentially driven by high-ability students.[46]

8.2 Grade retention in high school

Grade retention represents an important cost to the educational system. It may also serve as an indicator of student ability to deal with the increased learning intensity introduced by the G8 reform. It is therefore important to provide additional pieces of evidence on possibly unintended effects of the reform, documenting its effects on the probability of repeating a high school grade.

[46]This finding points to important heterogeneous reform effects, as lower-ability students appear to be less capable of coping with the higher per-grade curriculum requirements introduced by the G8 reform. See Andrietti and Su (2015) for a theoretical and empirical analysis of the distributional impact of the G8 reform on student achievement.

To this end, I estimate the following linear probability DD model:

$$Repeat\_high_{ist} = \beta_0 + \beta_1 G8r_{st} + \alpha X_{ist} + \delta_s + \gamma_t + \varepsilon_{ist}, \tag{4}$$

where $Repeat\_high_{st}$ equals one if a student experienced grade retention during high school, and zero otherwise, and $G8r_{st}$ equals one if a student entered high school *after the first treated cohort*, and zero otherwise. In states where multiple grades switched to G8 at the same time, i.e., Bavaria, Mecklenburg-Vorpommern, Niedersachsen, and Saxony-Anhalt, the first treated cohort corresponds to the highest grade initially treated. By contrast, the initially treated cohorts that I observe in these states – PISA 2006 in Saxony-Anhalt and Mecklenburg-Vorpommern; PISA 2009 in Bavaria and Lower Saxony – do not correspond to the highest grade treated.[47] Assigning these cohorts to treatment might therefore be problematic because grade retention found in these cohorts could have happened in grades that were not yet exposed to treatment (i.e., grades five to seven in Mecklenburg-Vorpommern, grades five and six in Saxony-Anhalt, and grade five in Bavaria and Lower-Saxony). To avoid this contamination issue, I drop these cohorts from the sample.[48] I first estimate equation (4) in its main specification. Then, I estimate additional specifications that explore possible heterogeneous policy effects by gender, parental education, and parental migration background. Table 9 reports the estimated coefficients on the reform dummy and on its interaction with the category considered. The first finding – in column (1) – is that there is no significant evidence of a G8 reform effect on the probability of repeating a grade in high school. Although the estimated coefficient is positive, it is also both economically (about 1 percentage point, or 10% of the baseline) and statistically insignificant. This is consistent with the evidence provided by Huebener and Marcus (2015) that the reform did not affect repetition rates in grades seven to nine. Furthermore, the heterogeneity analysis – in columns (2) to (4) – reveals that the probability of repeating a grade in high school is significantly higher after the reform for boys and for students whose parents have a migration background. By contrast, it is significantly lower for girls.

9 Conclusions

Quantifying the benefits of additional instructional time for university-bound students is key in answering important policy questions about the design of high school curricula, and in particular, about the duration and the intensity of learning. However, there is still a paucity of evidence on the effects of educational policies that lead to increased learning intensity.

[47] In Saxony-Anhalt and Mecklenburg-Vorpommern, the G8 reform – introduced in 2003 and 2004, respectively – affected grades five to nine. This means that PISA 2006 academic-track ninth-graders switched to G8 when they were in grade seven (Saxony-Anhalt) or eight (Mecklenburg-Vorpommern), and that earlier cohorts were treated since their eight or ninth grade. Similarly, in Bavaria and Lower-Saxony, the reform – introduced in 2004 – affected contemporaneously grades five and six. This implies that, although the PISA 2009 cohorts in Bavaria and Lower Saxony were treated since grade five, an earlier cohort was treated since grade six.

[48] Similar results – available upon request – are obtained dropping all the PISA cohorts from these states from the sample.

Germany's G8 reform offers a clean natural experiment to evaluate the impacts of an increase in learning intensity (i.e., a per-grade increase in instructional hours covering additional curriculum) on student achievement. Using pooled cross-sectional PISA data from the period 2000-2009 and a quasi-experimental approach, I offer the first comprehensive evaluation of the effects of this increased learning intensity on the reading, math, and science literacy skills of academic-track ninth-graders. Overall, the main concerns raised in the ongoing public G8 vs. G9 debate have only a limited basis in my empirical findings. First, the G8 reform did not harm student achievement, as usually claimed by its detractors. By contrast, I provide robust evidence that the reform led to significant improvements in reading, math, and science literacy skills. However, I do find some evidence of heterogenous reform effects: Girls benefit from the reform significantly more then boys on their reading scores, while high-ability students drive the reform effects in all subjects. Furthermore, I find no evidence of a significant average effect of the reform on high school grade retention, although I do find that the latter increased significantly for boys and for students with a migration background.

By analyzing the effects of higher learning intensity alone on student pre-graduation outcomes my study complements the existing G8 literature, which focuses mostly on estimating the overall reform effects (i.e., higher intensity and shorter high school duration) on graduation and/or post-graduation outcomes. It therefore represents an important contribution to the current "G8 vs. G9" debate in Germany. At minimum, my results suggest that more evidence should be gathered before switching back to G9. In particular, the availability of data over a longer time period on different student pre- and post-graduation outcomes would make it possible to evaluate the G8 reform once the main actors involved in the educational system – students, teachers, parents, and schools – have fully adjusted their behavior to the new regime after the initial transitional period.

References

Agüero, J. M., and T. Beleche (2013): "Test-Mex: Estimating the effects of school year length on student performance in Mexico," *Journal of Development Economics*, 103, 353–316. 6

Andrietti, V., and X. Su (2015): "Education curriculum and student achievement: Theory and evidence," Mimeo. 23

Angrist, J., and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. 18

Angrist, J. D., and A. B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014. 7

Bellei, C. (2009): "Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile," *Economics of Education Review*, 28, 629–640. 6

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?," *The Quarterly Journal of Economics*, 119(1), 249–275. 15

Besley, T., and R. Burgess (2004): "Can labor regulation hinder economic performance? Evidence from India," *The Quarterly Journal of Economics*, 119(1), 91–134. 18

Besley, T., and A. Case (2000): "Unnatural experiments? Estimating the incidence of endogenous policies," *The Economic Journal*, 110(november), F672–F694. 18

Büttner, B., and S. L. Thomsen (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. 4, 8

Cameron, C. A., J. G. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 90, 414–427. 15, 43

Cameron, C. A., and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 50(2), 317–372. 15

Carlsson, M., G. B. Dahl, and D. Rooth (2015): "The effect of schooling on cognitive skills," *The Review of Economics and Statistics*, 97(3), 533–547. 6

Cortes, K. E., J. S. Goodman, and T. Nomi (2015): "Intensive math instruction and educational attainment: Lon-run impacts of double-dose algebra," *Journal of Human Resources*, 50(1), 108–158. 6

Cuhna, F., and J. J. Heckman (2007): "The technology of skill formation," *American Economic Review*, 97(2), 31–47. 23

Dahmann, S. (2015): "How does education improve cognitive skills? Instructional time versus timing of instruction," SOEP papers on Multidisciplinary Panel Data Research 769, DIW Berlin. 4, 8

Dahmann, S., and S. Anger (2014): "The impact of education on personality. Evidence from a German high school reform," SOEP papers on Multidisciplinary Panel Data Research 658, DIW Berlin. 8

Dörsam, M., and V. Lauber (2015): "The effect of a compressed high school curriculum on university performance," Discussion paper, University of Kostanz. 8

Fuchs, T., and L. Wössmann (2007): "What accounts for international differences in student performance? A re-examination using PISA data," *Empirical Economics*, 32(2), 433–464. 13

Hansen, B. (2011): "School year length and student performance: Quasi-experimental evidence," Working paper, SSRN. 6

Hanushek, E., and L. Wössmann (2008): "The role of cognitive skills in economic development," *Journal of Economic Literature*, 46, 607–668. 2

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor Economics*, 24(43), 411–482. 2

Huebener, M., and J. Marcus (2015): "Moving up a gear: the impact of compressing instructional time into fewer years of schooling," Discussion paper 1450, DIW Berlin. 8, 24

Jürges, H., K. Schneider, M. Senkbeil, and K. Carstensen (2012): "Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy," *Economics of Education Review*, 31, 56–65. 21

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. 7

Kultusministerkonferenz (1997-2011): *Wochenpflichtstunden der Schülerinnen und Schüler - Statistiks 1997 bis 2011.* 6, 11

Lavy, V. (2010): "Do differences in school's instruction time explain international achievement gaps in maths, science and language? Evidence from developed and developing countries," Working Paper 16227, NBER. 6

Lee, R., and R. Barro (2001): "School quality in a cross-section of countries," *Economica*, 68(272), 465–488. 6

Lohmar, B., and T. Eckhardt (2010): *The education system in the Federal Republic of Germany 2008: A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe.* Bonn: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs. 4

Marcotte, D. (2007): "Schooling and test scores: A mother natural experiment," *Economics of Education Review*, 26, 629–640. 7

Marcotte, D., and B. Hansen (2010): "Time for school?," *Education Next*, 3(3), 316–338. 7

Marcotte, D., and S. Hemelt (2008): "Unscheduled school closing and student performance," *Education Finance and Policy*, 3(3), 316–338. 7

Meyer, T., and S. L. Thomsen (2013): "How important is secondary school duration for post-school education decisions? Evidence from a natural experiment," Discussion Paper 6, NIW. 8

——— (2014): "Are 12 years of schooling sufficient preparation for tertiary education? Evidence from the reform of secondary school duration in Germany," Discussion Paper 8, NIW. 8

Meyer, T., S. L. Thomsen, and H. Schneider (2015): "New evidence on the effects of the shortened school duration in the German states: An avaluation of post-secondary education decisions," Discussion paper, NIW. 8

Morin, L. P. (2013): "Estimating the benefit of high school for university-bound students: Evidence of subject-specific human capital accumulation," *Canadian Journal of Economics*, 46(2), 441–468. 7

——— (2015a): "Cohort size and youth earnings: Evidence from a quasi-experiment," *Labour Economics*, 32, 99–111. 21

——— (2015b): "Do men and women respond differently to competition? Evidence from a major education reform," *Journal of Labor Economics*, 33(2), 443–491. 21

OECD (2012): *PISA 2009 technical report*. OECD Publishing. 13, 15

Oreopoulos, P. (2006): "Estimating average and local average treatment effects of education when compulsory schooling laws really matter," *American Economic Review*, 96, 152–175. 7

Parinduri, R. A. (2014): "Do children spend too much time in schools? Evidence from a longer school year in Indonesia," *Economics of Education Review*, 41, 89–104. 6, 7

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. 7

Rivkin, S. G., and J. C. Schiman (2013): "Instruction time, classroom quality, and academic achievement," Working Paper 19464, NBER. 6

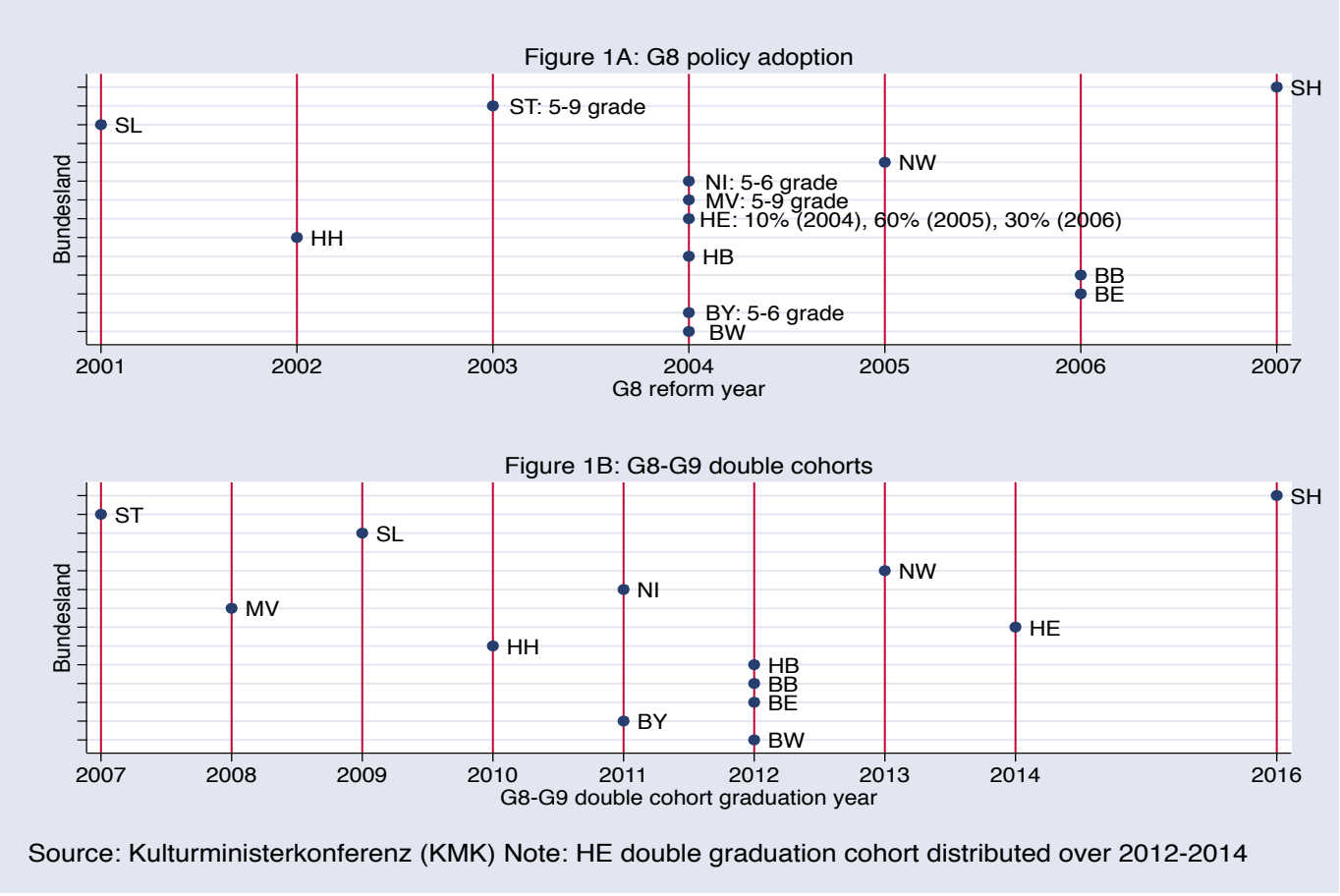
Sims, D. P. (2008): "Strategic responses to school accountability measures: It's all in the timing," *Economics of Education Review*, 27, 58–68. 6

Spinath, B., C. Eckert, and R. Steinmayr (2014): "Gender differences in school success: what are the roles of students' intelligence, personality, and motivation," *Educational Research*, 56(22), 230–243. 23

Thile, H., S. L. Thomsen, and B. Büttner (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. 7, 8

Wössman, L. (2003): "Schooling resources, educational institutions and student performance: The international evidence," *Oxford Bullettin of Economics and Statistics*, 65(2), 117–170. 6

Fig. 1. Timing of the G8 reform implementation



Source: Kulturministerkonferenz (KMK) Note: HE double graduation cohort distributed over 2012-2014

Legenda
BW: Baden-Württemberg
BY: Bavaria
BE: Berlin
BB: Brandenburg
HB: Bremen
HH: Hamburg
HE: Hessen
MV: Mecklenburg-Vorpommern
NI: Lower Saxony
NW: North Rehin-Westfalia
SL: Saarland
ST: Saxony-Anhalt
SH: Schleswig-Holstein

Fig. 2. Map of the G8 reform implementation timing

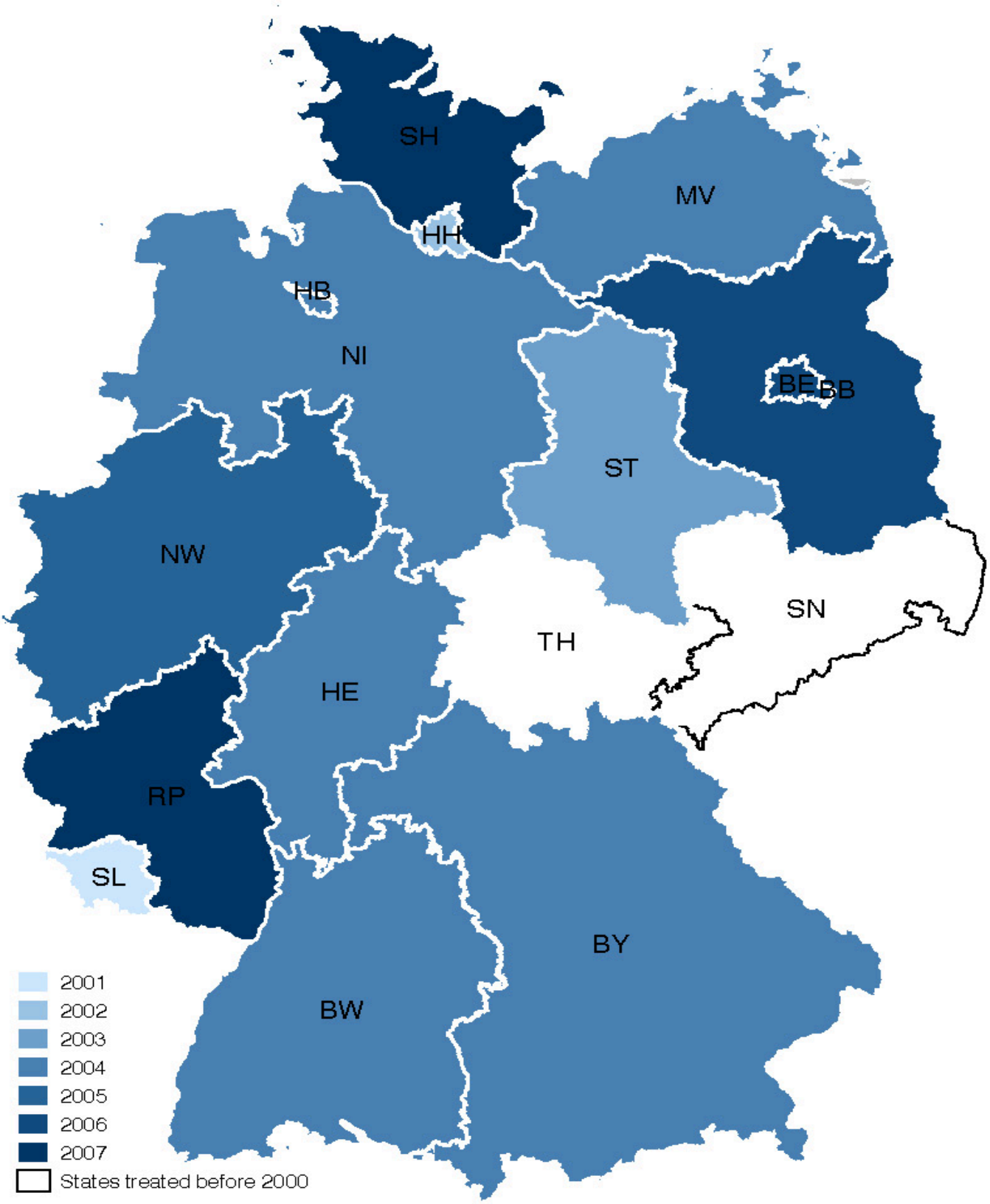
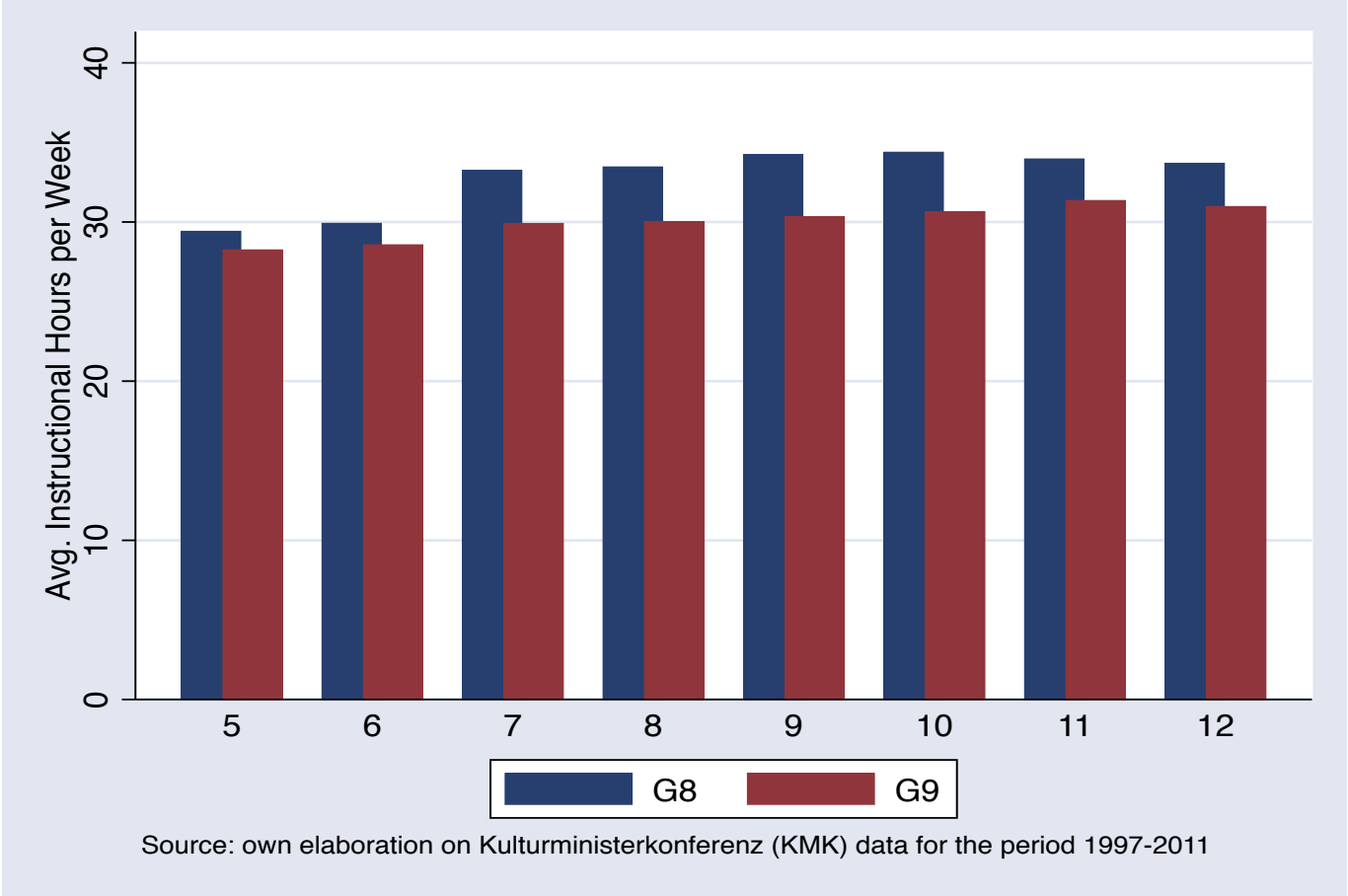


Fig. 3. G8 vs. G9: average instructional hours per week, by grade



Source: own elaboration on Kulturministerkonferenz (KMK) data for the period 1997-2011

Table 1. G8 treatment status of PISA cohorts across states and PISA cohorts

State	G8 adoption	Grades treated	Tracking year	PISA cohorts 2000	2003	2006	2009
Baden-Württemberg (BW)	2004	**5**	**2004**	C	C	C	T
Bavaria (BY)	2004	6	2003				
		5	**2004**	C	C	C	T
Berlin (BE)	2006	**7**	**2006**	C	C	C	T
Brandenburg (BB)	2006	**7**	**2006**	C	C	C	T
Bremen (HB)	2004	**5**	**2004**	C	C	C	T
Hamburg (HH)	2002	**5**	**2002**	C	C	C	T
Hesse (HE)*	2004	**5**	**2004**	C	C	C	C
Mecklenburg-Vorpommern (MV)**	2004	9	2000				
		8	**2001**	C	C	T**	
		7	2002				
		6	2003				
		5	**2004**				T
Lower Saxony (NI)	2004	6	2003				
		5	**2004**	C	C	C	T
North Rhine-Westfalia (NW)	2005	**5**	**2005**	C	C	C	C
Rhineland-Palatinate (RP)***	2007	**5**	**2007**	C	C	C	C
Saarland (SL)	2001	**5**	**2001**	C	C	T	T
Saxony (SN)******	-	-	-	-	-	-	-
Saxony-Anhalt (ST)**	2003	9	1999				
		8	2000				
		7	**2001**	C	C	T**	T
		6	2002				
		5	2003				
Schleswig-Holstein (SH)	2007	**5**	**2007**	C	C	C	C
Thuringia (TH)******	-	-	-	-	-	-	-

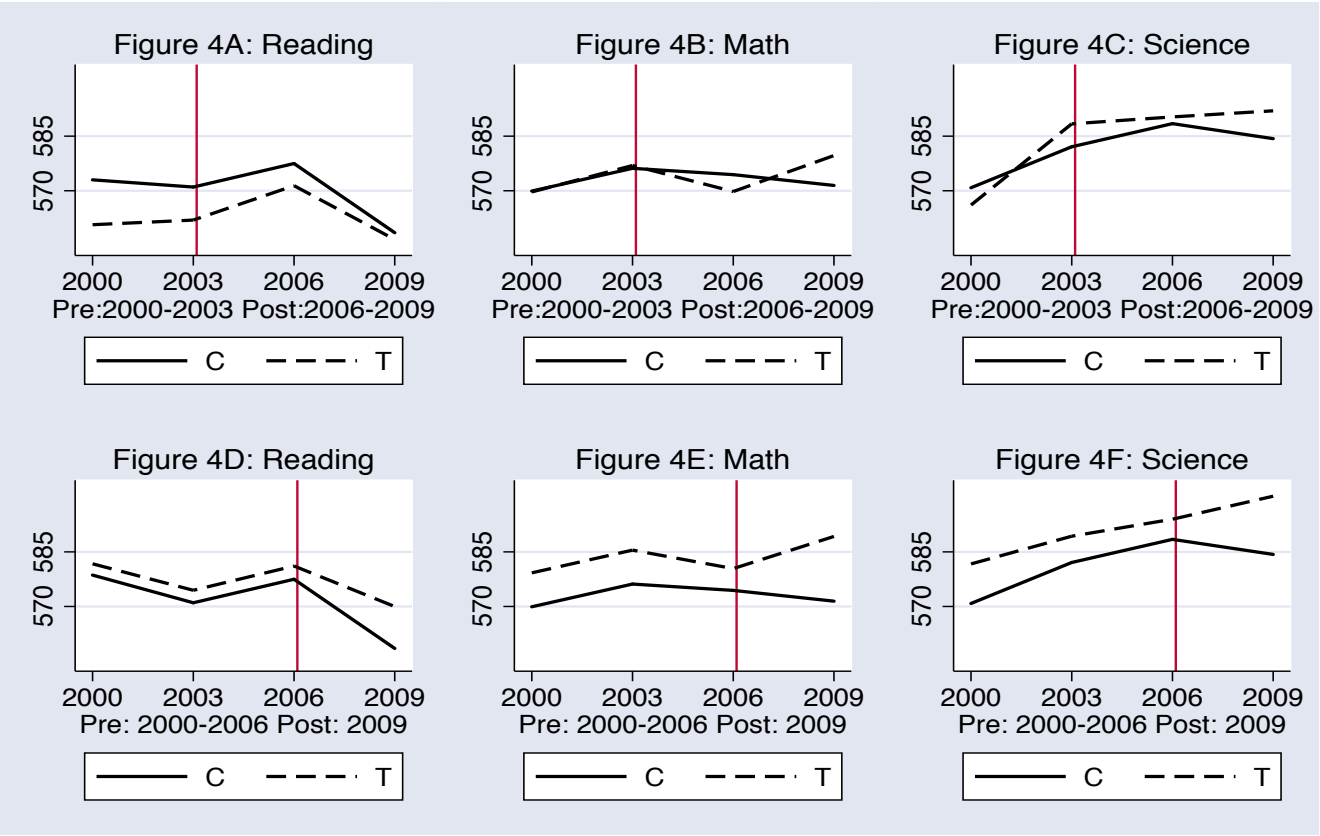
Notes: Column 1 indicates the year when the G8 reform was adopted. Column 2 reports the grades (cohorts) initially treated. Column 3 reports the tracking year of the cohorts initially treated, i.e., the academic year in which they entered academic-track high school. Figures in columns 2 and 3 are reported in bold when relevant to define the treatment status of PISA cohorts. T and C indicate treatment and control group, respectively. * In Hesse the G8 reform was introduced gradually: 10%, 60%, and 30% of schools were affected in 2004, 2005, and 2006, respectively. ** The PISA 2006 cohorts In Mecklenburg-Vorpommern and Saxony-Anhalt entered academic-track high school in 2001 (see Table 2), and were therefore treated only in grades 8 to 9 and 7 to 9, respectively. *** In Rhineland-Palatinate the reform has only been introduced in selected schools so far. **** After reunification, Saxony and Thuringia kept the G8 regime that was typical of academic-track high schools in former East states. *Source*: Kulturministerkonferenz (KMK).

Table 2. Academic-track high school enrollment by PISA cohort

	Grade attended by year						**PISA 2000**									
Grade 8			5	6	7	8										
Grade 9		**5**	**6**	**7**	**8**	**9**	**PISA c.**									
Grade 10	5	6	7	8	9	10	**test**									
Year	1994	1995	1996	1997	1998	1999	2000									
	Grade attended by year								**PISA 2003**							
Grade 8					5	6	7	8								
Grade 9				**5**	**6**	**7**	**8**	**9**	**PISA c.**							
Grade 10			5	6	7	8	9	10	**test**							
Year			1997	1998	1999	2000	2001	2002	2003							
	Grade attended by year											**PISA 2006**				
Grade 8								5	6	7	8					
Grade 9							**5**	**6**	**7**	**8**	**9**	**PISA c.**				
Grade 10						5	6	7	8	9	10	**test**				
Year						2000	2001	2002	2003	2004	2005	2006				
	Grade attended by year															**PISA 2009**
Grade 8											5	6	7	8		
Grade 9										**5**	**6**	**7**	**8**	**9**	**PISA c.**	
Grade 10									5	6	7	8	9	10	**test**	
Year									2003	2004	2005	2006	2007	2008	2009	

Notes: Grade attended by academic year for the PISA cohorts of academic-track ninth-graders indicated in bold. In Berlin and Brandenburg tracking takes place in grade 7.

Fig. 4. Trends in (non-standardized) scores: states treated in 2006 (2009) vs. control states



Source: Computations on PISA 2000-2009 pooled data (Final student weights used)

Legenda

All Figures:
Vertical lines separate pre- from post-reform periods
Control states (C): Hesse, North Rhein-Westfalia, Rheinland-Palatinate, and Schleswig-Holstein

Figures 4A, 4B, and 4C:
States treated in 2006 (and 2009) (T): Mecklenburg-Vorpommern, Saarland, and Saxony-Anhalt

Figures 4D, 4E, and 4F:
States treated only in 2009 (T): Baden-Württemberg, Bavaria, Berlin, Brandenburg, Bremen, Hamburg, and Lower Saxony

Table 3. Descriptives: treated means, and treated-control mean differences by PISA cohort

	PISA cohort							
	Pre-reform				Post-reform 1		Post-reform 2	
	2000		2003		2006		2009	
	T	T-C	T	T-C	T	T-C	T	T-C
PISA scores (p.v. avg.)								
Reading	578.76	0.13	572.66	1.65	579.99	2.52	568.73	10.66**
Math	577.86	7.79	584.27	8.11	579.18	4.76	588.34	16.83**
Science	579.51	8.53	589.17	7.08	593.55	5.11	599.52	16.02**
Grade retention								
Grade repeated	0.09	-0.02	0.06	-0.03*	0.12	-0.02	0.08	-0.01
Grade repeated in high school	0.08	-0.01	0.06	-0.03*	0.10	-0.01	0.07	-0.01
Treatment intensity								
Instruction hours: gr. 5-9	146.90	0.17	147.95	1.66	150.17	4.51**	163.46	17.75**
Instruction hours: gr. 7-9	90.08	0.69	90.28	1.34	92.12	3.16**	102.82	13.75**
Instruction hours: gr. 8-9	60.14	0.35	60.42	0.85	61.68	2.11**	68.92	13.74**
Student-level controls:								
Demographics								
Female	0.56	0.023	0.56	0.015	0.54	0.016	0.50	-0.024
Age (in months)	185.50	0.27	186.76	0.018	183.28	2.10	184.21	-0.159
Socio-economic backgr.								
Parents' ISCED 3-4	0.37	-0.019	0.35	0.002	0.24	-0.012	0.29	-0.01
Parents' ISCED 5-6	0.59	0.023	0.58	-0.002	0.73	0.012	0.61	-0.007
Parents' ISEI	57.76	0.518	56.00	-0.69	59.78	0.44	57.44	0.27
Books in house: >100	0.51	-0.023	0.28	0.027	0.70	-0.001	0.69	0.035
Only child	0.15	-0.005	0.17	-0.001	0.18	0.03	0.20	0.022
Kid born in foreign country	0.06	-0.03	0.04	- 0.019**	0.05	0.007	0.03	-0.019**
Parents born in foreign c.try	0.14	-0.067	0.10	-0.05**	0.14	-0.017	0.06	-0.014
No Deutsch spoken at home	0.03	-0.018	0.03	-0.011**	0.04	0.007	0.04	-0.017
School-level controls:								
Urban school: city>100k	0.32	-0.059	0.32	-0.066	0.28	-0.004	0.26	-0.002
Private school	0.06	-0.021	0.11	-0.053	0.11	-0.072*	0.02	-0.02
% of government funding	91.34	3.31	92.48	2.26	84.66	-2.65	98.12	8.49**
School enrollment	746.78	-110.60	736.37	-168.92**	855.05	-81.47*	943.27	-70.22
%of girls enrolled	51.12	1.92	51.76	-0.116	49.24	-0.74	51.06	- 0.167
Student-teacher ratio	11.64	-0.56	13.52	-3.67**	15.17	-1.8**	16.19	-0.97
Lack of computers	0.43	-0.07	0.36	0.057	0.25	0.006	0.48	0.26*
Lack of textbooks	0.13	-0.046	0.49	0.122	0.26	-0.089	0.22	-0.098
State-level controls:								
% of a-track schools	15.68	-10.43**	16.33	-10.05**	20.04	-6.88**	20.78	-7.41**
% of a-track students	35.21	-4.1*	37.51	-1.65	42.91	0.41	43.78	-1.72
% of grade 7 a-track students	32.23	0.001	34.07	1.76	36.84	1.38	39.53	1.21
% of a-track all-day schools	11.11	2.65	14.33	5.36	28.77	6.15	45.76	15.01
% of a-track all-day students	3.79	-0.27	5.42	1.18	14.66	6.65	24.01	14.57*
GDP per capita (,000)	25.03	-0,93	26.00	-0.90	28.86	0.18	29.45	0.31
Observations: Treated	5,091		4,455		6,585		1,697	
Observations: Control	2,217		1,774		3,511		1,171	

Notes: Mean differences estimated by OLS regressions weighted using final student weights. Standard errors clustered on state. ** and * indicate significance at 5 and 10 percent levels, respectively. The samples include ninth-graders in academic-track schools from each PISA cohort with a valid assessment in reading. Control states are those states whose cohorts are not treated during the observation period: Hesse, North Rhein-Westfalia, Rheinland-Palatinate, and Schleswig-Holstein. Treated states are those states whose PISA cohorts are treated since post-reform period 1 (Mecklenburg-Vorpommern, Saarland, and Saxony-Anhalt) or in post-reform period 2 (Baden-Württemberg, Bavaria, Berlin, Brandenburg, Bremen, Hamburg, and Lower Saxony). Saxony and Thuringia (always G8) are excluded from the samples.

Table 4. Main results: different specifications

	Baseline specification (1)	Student-level controls (2)	School-level controls (3)	State-level controls (4)
Panel A: Reading				
G8	0.138**	0.134**	0.123**	0.130**
	(0.033)	(0.022)	(0.024)	(0.032)
Observations		26,501		
Panel B: Math				
G8	0.120*	0.097*	0.080*	0.095**
	(0.065)	(0.051)	(0.039)	(0.036)
Observations		23,244		
Panel C: Science				
G8	0.120**	0.106**	0.107**	0.145**
	(0.031)	(0.027)	(0.034)	(0.037)
Observations		23,243		

Notes: Dependent variables in panel A, B, and C: standardized PISA scores in reading, mathematics, and science, respectively. Specification (1) is the baseline specification. Specifications (2) to (4) add student-, school-, and state-level controls, respectively. Specification (4) is the main specification. OLS regressions weighted using final student weights. Standard errors clustered on state reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The samples in panel A, B, and C include ninth-graders in academic-track high schools from the pooled PISA 2000-2009 dataset with a valid assessment in either reading, math, or science, respectively.

Table 5. Main results: different specifications

	Baseline specification (1)	Student-level controls (2)	School-level controls (3)	State-level controls (4)
Panel A. Reading				
Instruction hours grades 5-9	0.007*	0.007**	0.006**	0.007**
	(0.003)	(0.002)	(0.002)	(0.002)
Instruction hours grades 7-9	0.009**	0.008**	0.008**	0.010**
	(0.003)	(0.002)	(0.002)	(0.002)
Instruction hours grades 8-9	0.014**	0.013**	0.012**	0.015**
	(0.004)	(0.003)	(0.002)	(0.004)
Observations		26,501		
Panel B: Math				
Instruction hours grades 5-9	0.004	0.003	0.002	0.004**
	(0.004)	(0.003)	(0.003)	(0.002)
Instruction hours grades 7-9	0.007*	0.005	0.004*	0.005**
	(0.004)	(0.003)	(0.002)	(0.002)
Instruction hours grades 8-9	0.011**	0.008*	0.006**	0.009**
	(0.005)	(0.004)	(0.003)	(0.003)
Observations		23,244		
Panel C: Science				
Instruction hours grades 5-9	0.005**	0.004**	0.004**	0.007**
	(0.002)	(0.002)	(0.002)	(0.002)
Instruction hours grades 7-9	0.005*	0.004	0.004	0.008**
	(0.003)	(0.002)	(0.002)	(0.003)
Instruction hours grades 8-9	0.008**	0.007*	0.007*	0.014**
	(0.004)	(0.004)	(0.004)	(0.004)
Observations		23,243		

Notes: Dependent variables in panel A, B, and C: standardized PISA scores in reading, math, and science, respectively. Each panel reports estimated coefficients (and standard errors) for three alternative definitions of the regressor of interest in equation (2): cumulative instructional weekly hours in grades 5-9, 7-9, and 8-9, respectively. Specification (1) is the baseline specification. Specifications (2) to (4) add student-, school-, and state-level controls, respectively. Specification (4) is the main specification. OLS regressions weighted using final student weights. Standard errors clustered on state reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The samples in panel A, B, and C include ninth-graders in academic-track schools from the pooled PISA 2000-2009 dataset with a valid assessment in either reading, math, or science, respectively.

Table 6. Robustness checks: main specification

	Main spec. (1)	DD Placebo (2000-2003) (2)	DD Placebo (low-tracks) (3)	State trends (4)	DDD model (5)	DDD Placebo (2000-2003) (6)	Academic track (7)	G9 borders states (8)	First G8 cohort (9)	New CEE states (10)
Panel A: Reading										
G8 policy effect	0.130**	-0.024	0.031	0.114*	0.142**	-0.006	0.013	0.106**	0.110**	0.123**
	(0.032)	(0.045)	(0.031)	(0.060)	(0.046)	(0.078)	(0.024)	(0.036)	(0.037)	(0.049)
Observations	26,501	13,537	42,695	26,501	52,243	26,813	88,934	26,501	26,501	26,501
Panel B: Math										
G8 policy effect	0.095**	0.00	0.058	0.116*	0.116*	-0.084		0.101*	0.102*	0.044
	(0.036)	(0.035)	(0.035)	(0.065)	(0.059)	(0.103)		(0.057)	(0.055)	(0.046)
Observations	23,244	10,280	37,291	23,244	45,626	20,196		23,244	23,244	23,244
Panel C: Science										
G8 policy effect	0.145**	-0.023	0.049	0.166**	0.113*	-0.027		0.162**	0.138**	0.116**
	(0.037)	(0.030)	(0.031)	(0.070)	(0.060)	(0.089)		(0.044)	(0.043)	(0.036)
Observations	23,243	10,279	37,273	23,243	45,643	20,213		23,243	23,243	23,243

Notes: Dependent variables in all columns – except column (7) – of panel A, B, and C: standardized PISA scores in reading, math, and science, respectively. Dependent variable in column (7): attendance of academic-track high school (vs. other types of secondary schools). All estimated models based on the main specification – i.e., specification (4) in Table 4. G8 policy effect indicates the coefficient on the $G8_{st}$ dummy in the models estimated in columns (1), (3), (4), and (7) - (10), and the coefficient on the relevant interaction terms in the models estimated in the other columns, as explained in Section 7. OLS regressions weighted using final student weights. Standard errors clustered on state reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The sample used in columns (1), (4), and (8) - (10) includes academic-track ninth-graders with a valid assessment in either reading, math, or science (panel A, B, and C, respectively) from the pooled PISA 2000-2009 dataset. The sample used in column (2) includes academic-track ninth-graders from the pooled PISA 2000-2003 dataset. The sample used in column (3) includes ninth-graders in basic- and middle-track schools from the pooled 2000-2009 PISA dataset. The sample used in column (5) includes ninth-graders in middle- and academic-track schools from the pooled PISA 2000-2009 dataset. The sample used in column (6) includes ninth-graders in middle- and academic-track schools from the pooled PISA 2000-2003 dataset. The sample used in column (7) includes the full sample of ninth-graders from the pooled PISA 2000-2009 dataset.

Table 7. Robustness checks: main specification estimated on different samples

	Sample 1 main (1)	Sample 2 T 2006 (2)	Sample 3 T 2009 (3)	Sample 4 without MV, ST (4)	Sample 5 with SN, TH (5)	Sample 6 without repeaters (6)	Sample 7 flexible cutoff (7)
Panel A: Reading							
G8 policy effect	0.130**	0.313**	0.107**	0.099**	0.095**	0.171**	0.146**
	(0.032)	(0.040)	(0.031)	(0.032)	(0.032)	(0.036)	(0.032)
Observations	26,501	13,272	21,902	23,428	30,101	22,359	25,261
Panel B: Math							
G8 policy effect	0.095**	0.188**	0.087*	0.088**	0.097**	0.106**	0.105**
	(0.036)	(0.076)	(0.040)	(0.039)	(0.033)	(0.040)	(0.035)
Observations	23,244	11,551	19,378	20,660	26,310	19,464	22,183
Panel C: Science							
G8 policy effect	0.145**	0.214**	0.126**	0.113**	0.106**	0.170**	0.150**
	(0.037)	(0.039)	(0.041)	(0.108)	(0.038)	(0.037)	(0.035)
Observations	23,243	11,536	19,389	20,671	26,307	19,451	22,173

Notes: Dependent variables in all columns: standardized PISA scores in reading, math, and science, respectively. All estimated models based on the main specification. OLS regressions weighted using final student weights. Standard errors clustered on state reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The sample used in column (1) is the main sample. It includes ninth-graders in academic-track high schools with a valid assessment in either reading, math, or science from the pooled PISA 2000-2009 dataset. It excludes students from states that were already G8 at the beginning of the observation period (I.e., Saxony and Thuringia). The sample used in column (2) includes in the treatment group only those states that switched to treatment in 2006 (Mecklenburg-Vorpommern, Saarland, and Saxony-Anhalt). The sample used in column (3) includes in the treatment group only those states that switched to treatment in 2009 (Baden-Württemberg, Bavaria, Berlin, Brandenburg, Bremen, Hamburg, and Lower Saxony). Control states (C) in columns (1) - (3) are Hesse, North Rhine-Westfalia, Rhineland-Palatinate, and Schleswig-Holstein. The sample used in column (4) excludes from the main sample Mecklenburg-Vorpommern (MV) and Saxony-Anhalt (ST). The sample used in column (5) sample add Saxony (SN) and Thuringia (TH) as control states to the main sample. The sample used in column (6) excludes from the main sample students that experienced grade retention. The sample used in column (7) excludes from the main sample students born outside the school entry cutoff years of their PISA cohort.

Table 8. Heterogeneous effects: main specification

	Female gender (1)	**High educ. parents (2)**	**Migrant parents (3)**	**Grade retention (4)**
Panel A: Reading				
G8	0.054	0.135**	0.130**	0.146**
	(0.037)	(0.037)	(0.034))	(0.032)
Interaction	0.145**	-0.008	-0.005	-0.188**
	(0.042)	(0.046)	(0.154)	(0.077)
Observations		26,501		
Panel B: Math				
G8	0.095**	0.106**	0.100**	0.107**
	(0.038)	(0.031)	(0.037)	(0.038)
Interaction	0.001	-0.016	-0.080	-0.132*
	(0.030)	(0.033)	(0.153)	(0.047)
Observations		23,244		
Panel C: Science				
G8	0.135**	0.163**	0.152**	0.163**
	(0.036)	(0.038)	(0.039)	(0.038)
Interaction	0.020	-0.029	-0.106	-0.200**
	(0.041)	(0.038)	(0.140)	(0.049)
Observations		23,243		

Notes: Dependent variable in all columns of panel A, B, and C: standardized PISA scores in reading, math, and science, respectively. All estimated models based on the main specification – i.e., specification (4) in Table 4 – and include an interaction term between the column category dummy and the G8 dummy. OLS regressions weighted using final student weights. Standard errors clustered on state reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. Panel A, B, and C samples include ninth-graders in academic-track high schools from the pooled PISA 2000-2009 dataset with a valid assessment in either reading, math, or science, respectively.

Table 9. Probability of high school retention

		Heterogeneous effects, by:		
	Main spec. (1)	**Female gender (2)**	**High educ. parents (3)**	**Migrant parents (4)**
G8 policy effect	0.010	0.045*	0.015	0.003
	(0.019)	(0.025)	(0.033)	(0.025)
Interaction		-0.065**	-0.007	0.090**
		(0.021)	(0.042)	(0.026)
Observations		22,572		

Notes: Dependent variable equals one if a grade was repeated in high school, zero otherwise. Results reported in column (1) based on the main specification, i.e., specification (4) in Table 4. Results reported in column (2) - (4) obtained estimating models that add to the main specification an interaction term between the column category dummy and the G8 dummy. OLS regressions weighted using final student weights. Standard errors clustered on state reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The sample includes ninth-graders in academic-track schools with non missing values on the dependent variable, and with a valid assessment in reading.

Appendix

Table A.1. Main results: p-values obtained under different procedures

	Baseline specification (1)	Student-level controls (2)	School-level controls (3)	State-level controls (4)
Panel A: Reading				
G8 policy effect	0.138	0.134	0.123	0.130
P-value: clustering on state	0.001**	0.000**	0.000**	0.001**
P-value: wild cluster bootstrap	0.016**	0.002**	0.000**	0.000**
Observations		26,501		
Panel B: Math				
G8 policy effect	0.120	0.097	0.080	0.095
P-value: clustering on state	0.088*	0.078*	0.063*	0.019**
P-value: wild cluster bootstrap	0.146	0.124	0.068*	0.048**
Observations		23,244		
Panel C: Science				
G8 policy effect	0.120	0.106	0.107	0.145
P-value: clustering on state	0.002**	0.001**	0.007**	0.002**
P-value: wild cluster bootstrap	0.002**	0.002**	0.000**	0.006**
Observations		23,243		

Notes: Dependent variables in panel A, B, and C: standardized PISA scores in reading, mathematics, and science, respectively. Specification (1) is the baseline specification. Specifications (2) to (4) add student-, school-, and state-level controls, respectively. Specification (4) is the main specification. OLS regressions weighted using final student weights. P-values are reported below the estimated coefficients, and are computed according to two different computation procedures. The first procedure clusters on state, which is the default option in this study. The second procedure is based on the wild cluster bootstrap-t procedure (1000 replications, residuals estimated under H0, Rademacher weights used) described in Cameron, Gelbach, and Miller (2008). ** and * indicate significance at 5 and 10 percent levels, respectively. Panel A, B, and C samples include ninth-graders in academic-track high schools from the pooled PISA 2000-2009 dataset with a valid assessment in either reading, math, or science, respectively.