DIW BERLIN

1561

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2016

Increased Instruction Hours and the Widening Gap in Student Performance

Mathias Huebener, Susanne Kuger and Jan Marcus

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Increased instruction hours and the widening gap in student performance

Mathias Huebener,[a*] Susanne Kuger,[b] Jan Marcus[a,c]

[a] *DIW Berlin,* [b] *DIPF Frankfurt,* [c] *University of Hamburg*

March 2016

Abstract

Do increased instruction hours improve the performance of all students? Using PISA scores of students in ninth grade, we analyse the effect of a German education reform that increased weekly instruction hours by two hours (6.5 percent) over almost five years. In the additional time, students are taught new learning content. On average, the reform improves student performance. However, treatment effects are small and differ across the student performance distribution. While low-performing students do not benefit, high-performing students benefit the most. The findings suggest that increases in instruction hours can widen the gap between low- and high-performing students.

Keywords: Instruction time, student achievement, PISA, G8-high school reform, quantile regressions, curriculum, difference-in-differences
JEL: I21, I24, I28, D04, J24

Acknowledgements: This paper benefited from comments and suggestions by Steven Barnett, Mandy Baumann, Stefan Bauernschuster, Bernd Fitzenberger, Adam Lederer, Brian McCall, Tim Phillips, Friedhelm Pfeiffer, Ronny Scherer, Thomas Siedler, Katharina Spiess, Rainer Winkelmann, and seminar participants in Berlin, Hamburg, Heidelberg, and Nuremberg. Special thanks go to Ute Figgel-Dietrich and Geraldine Frantz for excellent research assistance. We thank the IQB Berlin for providing the data and Georgios Tassoukis from IZA for technical support with the remote access to the PISA data. We are grateful for funding of the German National Academic Foundation and the College for Interdisciplinary Education Research.
* Corresponding author: Mathias Huebener, Department Education and Family, DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany. Email: mhuebener@diw.de.

I Introduction

Substantial gaps in student achievements across countries are often attributed to large differences in school instruction time. Increasing the time that students spend in the classroom has moved into the policy focus in OECD countries. In the UK and the US, it is a central element of education policy agendas (OECD, 2016a). Policy-makers make two main arguments for increasing school instruction time. First, more instruction time could improve overall student performance by providing more learning opportunities. Second, more instruction time could help narrow performance gaps between low- and high-performing students by compensating for lacking resources or supervision outside school (OECD, 2016b). Despite the large hopes of policy-makers and the high costs of instruction time as a school input factor, the question of whether spending more time in the classroom can effectively improve student performance has received surprisingly little research attention (Patall et al., 2010; Lavy, 2015; OECD, 2016b). Even less is known about the effects of more instruction time on the performance gap between low- and high-performing students.

In this article, we study the impact of an increase in weekly instruction time on student performance induced by a large education reform in German academic track schools. The reform reduced the length of academic track schooling by one year, with the instruction hours and the curriculum gradually shifted forward in time. We focus on the performance of students in ninth grade, when they are typically 15 years old. For these students, the reform serves as a natural experiment to estimate the effect of spending 6.5 percent more time in the classroom through grades 5 to 9, i.e. between the ages of 11 and 15. This is equivalent to two additional instruction hours per school week, or about 350 hours overall. Our analyses rely on data from the Programme for International Student Assessment (PISA), pooled across the five waves from 2000 through 2012. The reform was implemented with regional and temporal variations, which we exploit in difference-in-differences models to estimate average and quantile treatment effects of the reform on student performance.

Estimates of the average treatment effect suggest that the reform increased PISA test scores of ninth graders in reading, mathematics, and science by about 6 percent of an international standard deviation. Quantile regressions reveal that students at

the bottom of the distribution show almost no effects, while students further up in the performance distribution benefit more from additional instruction time. The widening of the gap between low- and high-performing students is consistent across the three PISA domains of reading, mathematics, and science. Our findings are robust to various model specifications, and different placebo regressions support the common trend assumption.

This study helps address three limitations in the previous literature. First, many previous studies rely on small and short-lived exogenous changes in instruction time to estimate the effects on student performance. Such studies exploit variations in classroom time due to adverse weather conditions and unscheduled school closures (e.g. Marcotte, 2007; Goodman, 2014), quasi-random assignments of school start dates and assessment dates (e.g. Sims, 2008; Fitzpatrick et al., 2011), as well as student and teacher absences (e.g. Herrmann & Rockoff, 2012; Aucejo & Romano, 2014). Second, the studies generating insights from considerable, policy-induced increases in instruction time are often accompanied by changes in other school input factors or changes in the peer environment (Bellei, 2009; Lavy, 2012; Taylor, 2014; Cortes & Goodman, 2014; Cortes et al., 2015). We exploit a policy reform that led to a substantial and lasting increase in instruction hours, without affecting other school input factors or the peer environment. Third, the previous literature focuses predominantly on average treatment effects of instruction time. As increases in instruction time are regularly proposed in the debate on narrowing student performance gaps (OECD, 2016b), it is also important to determine which students benefit most from additional classroom time. Differential effects across the performance distribution are not previously discussed in the literature. We address this gap and contribute estimates of the reform effect on the distribution of student performance.

We conclude that (i) additional instruction time does improve average student performance; (ii) the effect sizes appear rather small given the substantial increase in instruction time; and (iii) the gap in the performance of low-performing and high-performing students widens. The additional learning content covered in the additional instruction time may be important in explaining why effect sizes are small

on average, and why effects increase as one moves up the performance distribution. The existing skill set of students may be important in transforming instructional input into student performance: Lower-performing students might need more time than better-performing students to process new learning inputs. Therefore, policy-makers increasing instruction time should be aware of the potential conflict between improvements in student performance and widening gaps in student performance. The learning content of additional time in school should be carefully considered.

The paper is structured as follows. Section II reviews the related literature. Section III describes the institutional setting and the German school reform from which we derive our findings. Section IV introduces the data, and section V outlines the econometric approach. We report the main findings in section VI, and check for the sensitivity of the findings and potential channels of the reform effect in section VII. Section VIII concludes.

II Related literature

Understanding the effectiveness of school input factors in increasing student performance is important for policy-makers assigning resources. The effectiveness of instruction time in increasing student performance has received surprisingly little attention, even though it is an omnipresent, easy-to-manage, but also costly input factor in education systems (Patall et al., 2010; Lavy, 2015; OECD, 2016b).

One reason may be the challenges involved in identifying the causal effects of instruction time on student performance. Some studies that correlate student performance in cross-sectional assessment data with instruction time find at most small positive, but not robust, relationships (Card & Krueger, 1992; Grogger, 1996; Lee & Barro, 2001; Woessmann, 2003). Yet, the observed cross-country correlations might be confounded by other features of education systems. In individual-level data, students' endogenous selection into more or less instruction time poses a challenge for the identification of causal effects. Lower-performing students might attend additional instruction hours to provide them with additional time to revise and understand the classroom content. Better-performing students might select courses that they

like most and that require more instruction hours. Two approaches are predominant in the microeconomics literature to address this challenge. The first looks at within-student variation in subject-specific instruction time. For instance, Lavy (2015) and Rivkin & Schiman (2015) use cross-subject variations in instruction time and controls for time-invariant, student-specific characteristics in student-fixed effects models. In contrast to previous correlation studies, they find a strong positive effect of instruction hours on student achievements. The other approach exploits quasi-experimental settings to analyse causal effects of instruction time on student performance. Marcotte (2007), Marcotte & Hemelt (2008) and Goodman (2014) use variation in winter weather that affected instruction time prior to centralised state school exams. Sims (2008) and Fitzpatrick et al. (2011) use school day variations induced by quasi-random assignments of school start dates and assessment dates. Herrmann & Rockoff (2012) and Aucejo & Romano (2014) identify the effects with random variations in student and teacher absence days. All these quasi-experimental studies identify the average effects of variations in actual instruction time, and find mostly beneficial effects of more instruction time.

The variation in instruction time that is used in these quasi-experimental studies is rather small and short-lived. Very few studies identify the effects of policy-induced increases in instruction hours. One exception is the introduction of full day schooling in Chile, evaluated by Bellei (2009). This reform increased instruction time, but also required large investments into the school infrastructure and significant institutional changes. Another exception is a school funding policy reform in Israel that altered weekly instruction hours, teaching budgets and the classroom time spent on core subjects. Lavy (2012) finds that both increases in funding and instruction hours improve student performance. Finally, there are studies on a policy program in the US that doubles mathematics instruction hours for low-performing students to provide extra time for remediation. Taylor (2014), Cortes & Goodman (2014) and Cortes et al. (2015) find positive effects of such reforms on student performance. Affected low-performing students are taught the additional instruction hours separately, which mixes the effects of additional instruction time with student composition effects. Also, it is not clear in how far the findings can be transferred to settings in which policy-makers also raise the number of instruction hours for

better-performing students. Two working papers exploit the same German reform to investigate the effect of additional instruction hours. Dahmann (2015) looks at survey data on fluid and crystallised intelligence and finds no overall reform effect. Andrietti (2015) finds positive average treatment effects on test scores using PISA 2000-2009 data. We go beyond this working paper in several dimensions. We additionally use PISA 2012 data, which allows us to include more treatment states. Further, we address a major shortcoming in the PISA data and merge detailed historical timetable information of students throughout secondary school from binding federal state regulations. We also provide heterogenous treatment effect estimates from more flexible model specifications. Most importantly, we go beyond the estimation of average treatment effects and analyse quantile treatment effects.

Overall, we add to the literature on the effects of instruction hours by looking at a substantial, lasting and exogenous increase in weekly instruction hours. This increase affects the allocated instruction time, a quantity that is of high policy-relevance. We extend the prevailing analytical framework of estimating average effects and provide important estimates for the effect on the distribution of student performance.

III The G8-academic track school reform

This paper derives the effects of increased instruction time on student performance from an education reform in German academic track schools. Students in Germany are tracked into different school types according to their ability, after joint primary schooling usually through the fourth grade. Academic track schools (*Gymnasiums*) constitute the high-ability school track, and intend to prepare students for university education.[1] Only this track was affected by the reform. It is attended by about one-third of each cohort. A noteworthy feature of the German education system is that each federal state enacts school track specific timetable regulations. These

[1]In some federal states, the university entrance qualification can also be earned in alternative school tracks that were not affected by the G8-reform. Reform effects on the choice of the school track are not an important concern for our identification strategy. A detailed discussion on this issue is provided in section VII.A.

regulations contain the distribution of weekly instruction hours across the different school subjects and they are binding for schools.[2]

Between 2001 and 2007, 13 out of 16 German federal states reduced the length of academic track schooling from nine to eight years. The so-called G8-reform aimed at bringing students to the labour market earlier without significant changes to the school curriculum. The number of total instruction hours required for academic track school graduation and the school curriculum were redistributed over the remaining school years (KMK, 2013), consequently increasing the number of weekly instruction hours in the remaining school years starting from grade 5.[3] On average, the increase amounts to about 2 additional hours per week in grades 5 to 9, which corresponds to an increase in weekly instruction hours by about 6.5 percent (details are provided in section VI). Each additional instruction hour was intended to cover new content, gradually shifting the curriculum forward from previously higher grades. The 13 federal states implemented the reform at different points in time. Table A.1 in the appendix provides an overview of the timing.

IV Data

We use data from the German extension of the Programme for International Student Assessment (PISA) for 2000, 2003 and 2006, as well as international PISA data for 2009 and 2012 on students in ninth grade (Baumert, 2009; Prenzel, 2007, 2010; Klieme, 2013).[4] The data contain internationally standardised measures of student performance (PISA scores) in the three domains of reading, mathematics, and science. Each domain is standardised to have an international mean of 500 and a standard deviation of 100.[5] The PISA assessments go beyond curriculum-based assessments and examine if students can make effective use of their knowledge and

[2]For more details on the education system in Germany, see e.g. Dustmann et al. (2014).

[3]In some states tracking takes place after grade 6 (details are provided in table A.1). In these states the additional instruction hours are distributed over fewer years.

[4]For 2009 and 2012, the German extensions of PISA lack information on student performance in mathematics and reading; they focused on language skills.

[5]For each PISA domain, students are assigned five plausible values randomly drawn from a likely test score distribution. We deal with this multiply imputed data set following the recommended standard procedure outlined in Rubin (1987). Further details on the PISA assessment procedure are contained in the PISA technical reports.

skills in reading, mathematics, and science in situations likely to be encountered outside of school. In addition to the achievement data, the PISA data provide information from separate questionnaires for students and school principals.

In our main analyses, we focus on students in academic track schools as only this track was affected by the G8-reform. We pool information over five PISA waves, obtaining a sample of 33,217 *academic track* students in ninth grade.[6] The German school year usually starts in August or September, while the German PISA assessments take place in April and May. We therefore capture the effect of additional instruction time over a period of 4.7 school years, beginning in fifth grade.

In the student background questionnaire, students are asked about their instruction hours. However, this information focuses only on the grade students are currently in. The number of instruction hours in a specific grade may not be informative for the overall level of instruction hours students are exposed to throughout schooling; fewer instruction hours in one subject in one grade might be compensated with more instruction hours in other grades. A further complication is that the questions on instruction hours is asked differently across PISA waves, sometimes targeting only certain subjects. We overcome these shortcomings with information from official timetable regulations that the federal states enact. We carried out extensive archive research on historical timetable regulations and assign each student his effective timetable throughout academic track school, depending on the grade at the time of the PISA survey, and the federal state he lives in. The official timetable regulations match students reported instruction hours for grade 9 in the PISA data very well (table A.2 in the appendix). This confirms the binding nature of the regulations, and provides confidence that the information for earlier grades is also reliable. Figure 1 plots the average number of weekly instruction hours in grade 5 through 9 for the school entry cohorts 1991 to 2003 for each federal state. One can see a sharp increase in weekly instruction hours after the reform implementation.

Descriptive statistics of our pooled sample are provided in table 1. The mean PISA test scores are above the international mean of 500 because we focus on students

[6]While the international PISA data sample 15-year old students, we focus on students in the modal grade nine as the international PISA 2009 data for Germany includes only ninth-graders.

in the high-ability track. In grades 5 to 9, students have on average 31 instruction hours per week, with on average 4.2 instruction hours in German, 4 instruction hours in mathematics, and 3.6 instruction hours in science. Females constitute 54 percent of our sample and 13 percent of students have at least one parent who was not born in Germany. The students are 15.4 years old, on average. Approximately 7 percent of the students repeated a grade throughout their educational career. Further, 64 percent of students have at least one parent with a tertiary education degree. At the school level, the average school size is 850 students. Public schools make up 91 percent of the sample, with 36 percent of teachers working part-time. The average student-computer-ratio is 31.7 and the student-teacher-ratio is 16.7. Students affected by the G8-reform constitute 38 percent of our sample.

V Methodology

In order to obtain causal effect of the G8-reform, we exploit the fact that the reform was implemented at different points in time across the federal states. We estimate the average treatment effect of the reform on student performance with separate difference-in-differences (DiD) models for PISA scores in reading, mathematics, and science. The model we estimate is

$$y_{ist} = \beta \cdot G8_{st} + \mu_s + \kappa_t + X'_{ist} \cdot \lambda + \varepsilon_{ist} \tag{1}$$

where y_{ist} is the performance of student i in federal state s at time t in one PISA domain. $G8_{st}$ is a binary variable that identifies whether the student was affected by the G8-reform. β is the coefficient of core interest and identifies the reform effect on student performance. With the standardised PISA scores as outcome, β can be immediately interpreted as the effect in percent of an international standard deviation. State-fixed effects (μ_s) account for cohort-invariant differences in the outcome variables between different federal states, i.e. general state differences in terms of school funding, teacher quality, school quality, or student ability will not confound our findings. κ_t captures general differences between cohorts over time as well as student performance shocks common to all federal states, e.g. resulting from methodological changes across PISA waves. The set of individual control

variables, X_{ist}, contains a quadratic term for students' age, a gender dummy, a migration background dummy, as measured by whether at least one parent was born abroad, as well as a set of five indicator variables for parents' highest education level, as measured by the international standard classification of education, ISCED. In section VII we confirm that these control variables are orthogonal to our reform indicator. Their inclusion can increase the precision of our estimates. Given the state- and cohort-fixed effects, the variation in the G8-treatment indicator stems from the differential timing of the reform across the federal states (see also table A.1 in the appendix). By the time the PISA 2006 assessment was conducted, three federal states had changed to the G8-regime. By PISA 2009, seven more states had followed, and by PISA 2012 two more states had implemented the reform.[7]

We estimate equation 1 with ordinary least squares (OLS), using student sampling weights provided in the PISA data. Standard errors account for heteroskedasticity in the error term, ε_{ist}, and are clustered at the federal state level.[8] Standard errors and coefficient estimates also take into account that each student has five plausible values of PISA test scores, randomly drawn from the likely test score distribution. As recommended by the PISA technical reports, we run our regressions on each of the five plausible values and combine the estimated standard errors and point estimates according to the procedure outlined in Rubin (1987).

The causal interpretation of the resulting estimates rests on two major assumptions: We have to assume that there are no compositional changes in the student body due to the reform and that the PISA scores would have followed the same trend in the treatment and control group in the absence of the reform (common trend assumption). In section VII we provide evidence for the plausibility of these

[7]In the federal state of Schleswig-Holstein, cohorts affected by the G8-reform are outside the period of our analysis. The federal state of Hesse – accounting for about 8 percent of academic track students in Germany – implemented the G8-reform over a period of three years. While in the first year, only 10 percent of academic track schools implemented the reform, two years later all academic track schools had implemented the reform. For our analyses, we use Hesse as a control state in the first year of the implementation. In the next PISA wave, three years later, Hesse is treated as a treatment state.

[8]Our estimation results are based on 16 clusters. We also perform wild cluster bootstrap methods to account for the comparably small number of clusters (Cameron et al., 2008). The p-values are of similar magnitude as the p-values based on clustered standard errors from OLS regressions.

assumptions.

While OLS asks how the conditional mean of student performance is affected by the reform, this focus on average treatment effects might hide important differences across the performance distribution. In particular, it is crucial to understand whether additional instruction time could help narrow performance gaps between low- and high-performing students. We perform quantile regressions to obtain a more complete description of how the conditional distribution of student performance is affected by the reform. We estimate the reform effect at quantile τ of the conditional distribution with the following model:

$$Q_{Y_{ist}}(\tau|G8_{st}, \mu_s, \kappa_t, X_{ist}) = \beta(\tau) \cdot G8_{st} + \mu_s(\tau) + \kappa_t(\tau) + X'_{ist} \cdot \lambda(\tau). \qquad (2)$$

As before, $G8_{st}$ is a binary treatment indicator, μ_s denotes state-fixed effects, κ_t captures cohort-fixed effects and X_{ist} is the set of student characteristics. The quantile treatment effect at quantile τ is estimated by solving a linear programming algorithm. As before, we apply student sampling weights. Bootstrapped standard errors of the main results account for clustering at the federal state level.[9]

VI Results

A. Main results

In table 2, we first present results for the reform effect on instruction hours across grades and subjects for the students in our sample. On average, the weekly instruction time in grades 5 to 9 increased by 1.99 hours, or 6.5 percent. Across the different grades, the increase varies between 1.62 (5.3 percent) and 2.65 (8.4 percent) hours, with the largest absolute increases in grades 8 and 9. Across the different subjects,

[9]Using the German PISA data in combination with highly confidential federal state identifiers requires carrying out the analyses with Stata via a remote access. Standard Stata quantile regression commands allow for either weighting of the regressions (*qreg*) or clustering of the standard errors (*qreg2*). As it is common practice in applied work to report bootstrap standard errors for quantile regressions, we circumvented this limitation by bootstrapping the weighted quantile regressions for the main results in table 3. For the 378 quantile regression models estimated in heterogeneity analyses and sensitivity checks, we report conventional standard errors, as each regression with 200 bootstrap replications takes about two hours and occupies computer resources of the remote access.

the average increase in German (language arts) is 0.02 hours, in mathematics 0.1 hours and in science (biology, physics, and chemistry) 0.62 hours. The instruction hours of all other subjects, including foreign languages, geography, social sciences, sports, and arts increased on average by 1.25 hours per week.

Table 3 shows our main results. Column 1 reports the results for the average treatment effect of the G8-reform. The coefficients suggest a statistically significant increase in reading, mathematics, and science test scores of about 5.3 to 5.8 percent of an international standard deviation. To illustrate the magnitude, we compare the reform effects with three different quantities: the increase in PISA scores of a typical school year, the gender differences in test scores, and findings on the effects of instruction time in other PISA-based studies. On average, one year of schooling in Germany is estimated to raise test scores by 33 percent of a standard deviation (Prenzel et al., 2006). Students affected by the G8-reform received on average two additional instruction hours per school week for 4.7 school years, which amounts to one-third of an additional school year. The reform effects correspond to about one-fifth of the annual increase. This suggests that the increase in performance lags behind the increase in instruction hours. Also relating the findings to the gender reading gap, our point estimates for the average treatment effect seem to be rather small. Girls outperform boys on average by 15 percent of an international standard deviation in reading, but are worse off by 26 percent in mathematics and 30 percent in science.[10] Relating to findings in the economics literature using PISA data, Rivkin & Schiman (2015) and Lavy (2015) find effect sizes between 3 and 6 percent of a standard deviation for one additional subject-specific instruction hour per week. Comparing these findings to our results is somewhat complicated. Both studies proxy general differences in instruction time with a contemporaneous level of instruction hours reported at the time of the PISA test. The increase in instruction hours in the setting we analyse occurred across several grades and subjects, with some evidence for spill-over effects between subjects (Rivkin & Schiman, 2015, see also section VI.C). Also, increases in instruction time in earlier grades may matter for future learning (Rothstein, 2010).

[10]Estimates for the gender gaps are based on the estimate for the gender dummy in equation 1.

The quantile regression results are reported in columns 2 to 10 of table 3. Across all PISA domains, effect sizes are positive, but small and mostly insignificant until the third decile. The treatment effects increase almost monotonically as one moves up the performance distribution, and become statistically significant. Under the common assumption of student rank stability, the reform appears more effective for students further up in the performance distribution. The results suggest that the distribution of student performance widens because of the reform.

Our findings somewhat contrast with the small existing literature on instruction time and student performance. While we find that the lower part of the performance distribution does not benefit from additional instruction hours, policies doubling mathematics-instructions to support low-performing students show positive effects on student performance (Taylor, 2014; Cortes & Goodman, 2014; Cortes et al., 2015). Note that these policies devote extra learning time to remediation. But also within the group of low-performing students, Allensworth et al. (2009) find that better-performing students benefit more. Banerjee et al. (2007) show for an education intervention in India that remediation classes are most beneficial for students at the bottom of the performance distribution. An important difference compared to these settings is the curricular content of instruction time. In our setting, additional hours cover new content.

That the content may be an important determinant of the benefits of learning time is supported by findings from a high school programme in the US that teaches algebra courses from higher grades in earlier grades. As a consequence of teaching more difficult courses earlier, Allensworth et al. (2009) and Clotfelter et al. (2015) find negative effects on mathematics test scores, suggesting that the benefits from instruction time declined. The authors argue that students have not been sufficiently prepared.

The modest increase in student performance is also consistent with diminishing marginal returns to additional hours if student concentration and the capability to process new inputs declines with additional time (Rivkin & Schiman, 2015). Another explanation for our findings may lie in the opportunity costs of time (Rivkin & Schiman, 2015). Students' time spent outside school is substituted by classroom

time spent on new learning content. Leisure time could have been invested to revise and understand the content covered in the classroom. Also, sleep is important for processing new inputs (Eide & Showalter, 2012). This time substitution may be most problematic for low-performing students lacking time outside school to process the additional inputs.

The pattern in the results hints at skills and instruction hours being complements in the educational production process. The pre-existing skill set may be important for digesting new learning content and transforming it into student performance. Studies on other school input factors also reveal in quantile regressions that treatment effects increase with students' position in the performance distribution (Rangvid, 2007; Bellei, 2009; Mueller, 2013).[11]

B. Heterogeneity analyses

In this section, we estimate treatment effects for different student subgroups. Group differences in the effects can carry important implications for student performance gaps, well-documented between boys and girls (e.g. Dee, 2007), between natives and migrants (e.g. Lüdemann & Schwerdt, 2013), as well as between students from low and high socio-economic backgrounds (e.g. Agasisti & Longobardi, 2014).

In table 4, we report the results for subsamples stratified by certain socio-economic characteristics: gender, migration status, and parental education.[12] Across the three domains of reading, mathematics, and science, the effects are almost identical for girls and boys. In reading, there are somewhat larger point estimates for migrants. However, the low share of migrants at academic track schools reduces the sample size, and the coefficient is insignificant. The estimates on mathematics and science performance are clearly higher for natives and close to zero for migrants. Children from parents without a degree in higher education exhibit larger point estimates in

[11] Note that establishing the causal relationship between student performance and the complementarity of instruction hours and skills requires also exogeneity in students' skills as they may correlate with unobserved family investments or other child characteristics (Todd & Wolpin, 2003, 2007).

[12] Analogously, table A.3 in the appendix presents the results for the quantile regressions. In general, the quantile regression results confirm the picture of the average treatment effects for the heterogeneity analyses.

mathematics and science, but smaller estimates in reading.

Lavy (2015) suggests that the effects of instructional time are stronger for girls, migrants, and students from low socio-economic backgrounds. The findings in our setting suggest that the gender difference in treatment effects is negligible. The increase in instruction hours had smaller effects for migrants. Even though we find some differences in the treatment effects between the subgroups, we cannot establish the statistical significance of these differences.

C. Subject-spill-over-effects of instruction hours

The reform effects on reading, mathematics, and science scores are similar, despite differences in the subject-specific increase in instruction hours. German and mathematics displayed smaller increases in instruction time than science, but still the estimated effects are similar to the reform effect on science scores. Next to subject-specific heterogeneities in the benefits of instruction time, this pattern may stem from spill-over effects between subjects. PISA tests are not curriculum based and any increase in instruction hours involves interacting with classroom material, solving problems, and reading school material, which may improve student performance in all domains. Rivkin & Schiman (2015) provide indirect evidence for subject-spill-over effects. In this section, we present some direct evidence that is consistent with subject-spill-over effects of instruction time on student performance.

The major variation in instruction hours in our setting is caused by the G8-reform (see figure 1). This allows identifying the effect of instruction time on performance by within-subject variations. We replace the G8-reform dummy in equation 1 with four continuous variables, namely the total number of instruction hours in German, mathematics, science, and all other subjects for grades 5 through 9. The estimation results are presented in table 5. Generally, the findings are in line with our expectations. Student performance in reading is positively affected by German classes, and the performance in mathematics is positively affected by mathematics classes. Instruction hours in other subjects also have a positive and significant impact on reading and mathematics performance, suggesting spill-over effects between subjects. The category includes reading-intensive subjects like history, social

sciences, and foreign languages, which might explain the larger effect on reading scores. The findings on student performance in science are less straightforward. Additional science instruction hours seem to have no effect. Instead, the coefficient of mathematics hours is significant. Still, this is in line with implicit evidence provided by Rivkin & Schiman (2015), who also suggest that mathematics instruction hours affect performance in science.

However, the coefficients on subject-specific instruction hours should not be overemphasised for several reasons. First, only variation in the subject-specific changes in instruction hours in twelve reform states identify the coefficients, and the changes across subjects may be correlated. Second, the model assumes that instruction hours in grade 5 have the same effect as instruction hours in grade 9. It is not clear whether this assumption holds: One could argue that instruction hours in higher grades should receive a higher weight as the learning content covered in class is more readily available for the students. On the other hand, instruction hours in earlier grades might be more relevant because of skill complementarities. Therefore, the effects of subject-specific instruction hours should be interpreted with caution.

VII Sensitivity checks

In this section, we present sensitivity checks for the robustness of our findings to different model specifications. We begin by investigating two main threats to our identification strategy: reform-induced compositional changes in the student body and generally differing time trends between treatment and control states. We then discuss the sensitivity of our results to changes in the set of control variables and the sample definition, before we discuss whether the reform might have worked through other channels than instruction hours. Finally, we discuss the external validity of our findings.

A. Threats to the identification strategy

The consistency of our reform effect estimates rests on two main assumptions. The first assumption is that the G8-reform must not have affected the composition of students attending academic track schools, the only school track affected by the

reform. As all academic track schools within a federal state were required by law to implement the reform in one specific year, students may opt for a lower quality school track, or move to another federal state that has not (yet) implemented the reform. The choice for a lower quality school track has long-lasting consequences as the academic track school is the usual way to earn the general university entrance qualification. Commuting or moving to another federal state involves high costs to both the child and its family, and became increasingly difficult as more federal states implemented the reform. Any kind of avoidance behaviour should be evident from enrolment rates in academic track schools. Huebener & Marcus (2015) find no evidence for reform-induced lower enrolment rates at academic track schools using administrative data on all students in Germany.

The PISA data allow to directly check for compositional changes in the student body based on observable student characteristics (gender, parental education, migration background, age). We run baseline difference-in-differences regressions as outlined in equation 1 without individual control variables. Dependent variables are the stated student characteristics. The results are reported in columns 1 to 4 of table 6. All coefficients are close to zero and insignificant. Hence, there is no evidence for compositional changes in the student body at academic track schools following the G8-reform. Another reason for compositional changes could be increases in grade repetitions due to the reform. However, Huebener & Marcus (2015) show that the reform did not affect grade repetitions until grade 9. We can confirm this finding in the PISA data as well (column 5 of table 6). This notion is also supported by the absence of a reform effect on the students' age in ninth grade (column 4 of table 6).

The second main assumption of our identification strategy is the common trend in student performance between treatment and control states if the reform was not implemented. The way the reform was implemented across the federal states and in one specific school track only, enables us to simulate two different placebo treatments that can add plausibility to the common trend assumption. First, we add a placebo reform dummy to equation 1 that assumes the reform would have taken place one PISA-wave (three years) earlier. A significant coefficient for this placebo policy would indicate that the treatment and control group followed different

trends in the outcome variables before the onset of the G8-reform. Second, we investigate the reform effect on alternative school tracks that were not affected by the reform. Significant results in this placebo specification would indicate that other factors unrelated to the G8-reform changed simultaneously in the treatment states also affecting other school types. The results are reported in column 2 and 3 of table 7.[13] Both placebo-reforms produce coefficients that are small and statistically insignificant, adding plausibility to the common trend assumption.

Another violation of the common trend assumption could stem from confounding effects of other education reforms implemented over the same time period. Major reforms affecting academic track schools include the introduction of central exit exams, changes in the grade in which students are tracked, and changes in the number of alternative school tracks next to the academic school track. Table A.1 in the appendix reports the timing of these reforms. In columns 4 to 6 of table 7, we add dummy variables to equation 1 for each of these reforms. Our findings remain robust.

B. Specification issues

In this section, we show that our results do not depend on the choice of control variables and the restriction of our sample. In column 7 of table 7, we estimate the model without the set of student characteristics, X_{ist}. As certain individual control variables are missing for approximately 6 percent of the sample, in column 8 we include these observations in our sample and re-estimate the model without socio-economic control variables. In column 9, we add a set of school characteristics (teacher-student-ratio, student-computer-ratio, public or private school dummy) to the model in equation 1 in order to show the robustness of our results to this additional set of control variables.[14] The stable estimated reform effects suggest that changes with respect to the set of control variables or sample restrictions do not threate our findings.

[13]The pattern for the quantile treatment effects are very similar to the main effects. The results are reported in table A.4 in the appendix.

[14]This is not our main specification as several schools completely lack these information. In order to maintain the sample size, we set missing values to zero, and include dummy variables indicating the missing values on each of the school characteristics.

C. Other channels

In the following, we examine whether the G8-reform might affect student performance through other channels besides the increase in weekly instruction hours. For example, the reform could affect the time spent on out-of-school learning activities, such as homework, attending out-of-school classes, or receiving private tutoring. *A priori*, the direction of such an effect is ambiguous. Teachers could assign more homework proportional to the increase in instruction hours, or reduce it in order to provide more time for recreation. Attending out-of-school classes or private tutoring may decrease if these activities are substituted with classroom time. Or, the demand increases in order to better understand the classroom material in private remediation classes. In 2003 and 2012, the student questionnaire contains similar questions on homework, out-of-school classes and tutoring. This allows for the development of a general idea on the importance of these channels to determine the estimated effects on student performance outside the classroom. Table A.5 in the appendix compares the means of students in all states that introduced the G8-reform between 2003 and 2012 to states that did not. The average number of hours per week spent on homework is very similar between both groups in 2003 and 2012. The share of students attending out-of-school classes and private tutoring increased more strongly in control states than in treatment states between 2003 and 2012. This suggests some small substitution effects of out-of-school classes with classroom time in school. We interpret the baseline difference-in-differences estimates as a sign that changes in the amount of homework and in the use of out-of-school classes play a minor role in explaining the effects.

The reform enacted increases in the *allocated* instruction time, but increases in students' *actual* instruction time could be different if the reform affected students' behaviour to skip or miss classes. In PISA 2000 and 2012, the student questionnaire asked students how often they missed school, skipped classes or arrived late for school during the previous two weeks. We again calculate baseline difference-in-differences estimates, reported in table A.6 in the appendix. The propensity of students to miss class, skip class, or to arrive late for school was very similar prior to the reform and did not develop differently over time between treatment and control states. Increases

in actual instruction time do not lag behind increases in allocated instruction time.

May variations in the term length confound the findings? G8-treatment effects on school and bank holidays also show that these outcomes are not affected by the G8-reform. The estimation results are reported in table A.7 in the appendix.

While the classroom quality is shown to be a potentially important determinant of the returns to instruction time (Rivkin & Schiman, 2015), reform effects on the classroom quality are not significant drivers of the observed reform effects in our setting. Our effects are derived from variations within a given school infrastructure and school environment. As the composition of the student body at academic track schools was not affected by the reform, students' peer environment is unlikely to have changed. The slow-moving labour markets for teachers and high certification standards for teachers also do not point to relevant changes in the quality of teachers.

Did the reform change the composition of the teacher body at academic track schools? In general, if instruction hours are increased, schools would need to proportionally increase the teaching load of the present teachers or hire new teachers. Hence, any increase in the demand for teachers would be part of overall effects of increasing instruction hours. Note that in our setting, the potential impact of changes in the teacher body is exceptionally small. The total number of instruction hours taught at a given school increased in the transition period only, i.e. the period in which students in the 8-year academic track and older students still in the 9-year academic track run parallel. While the G8-reform increased instruction hours, it also reduced the length of the academic track by one school year. Rather than hiring new teachers, anecdotal evidence suggests that schools expanded the teaching load of existing teachers during the transition period, for instance through increases in working hours of part-time teachers, postponed retirements, and returns of recently retired teachers. In columns 6 and 7 of table 6, we report the reform effect on the share of full time teachers in the total teacher pool and the effect on the student-teacher-ratio measured at the school level. A small positive, but insignificant point estimate suggests that the share of full-time teachers slightly increased, which is consistent with the anecdotal evidence. At the same time it shows that changes in the composition of the teacher-body play a negligible role in explaining the effect

patterns of increased instruction time. In addition, the student-teacher-ratio was not affected by the reform.

The reform may also have changed teacher motivation and effort. On the one hand, teachers could have become more motivated and exert more effort if they see students struggling. On the other hand, prolonged working days of teachers could lead to decreasing motivation and lower effort. If the reform affected teacher motivation, it would be part of the reform effect as the reform was passed with the idea of a permanent change. Similarly, parental investments in family education inputs may respond to the increase in school instruction hours, and explain portions of the observed effects. But also with parental investments, any change would be part of the reform mechanism, which is not specific to the institutional context.

Summing up, the assembled arguments suggest that the major effect is indeed induced by increased instruction hours that can also be realised in other education systems.[15] Any adjustments in the behaviour of students, parents, and teachers are likely to be part of the effect of increases in instruction hours. They seem not specific to the German context.

D. External validity

The implementation of the reform facilitates contrasting developments across states, cohorts and school tracks, so the findings should have good internal validity. But are the findings also informative beyond the German experience, and have external validity to other contexts? Due to potentially diminishing benefits of additional classroom time, policy-makers have a natural interest in knowing whether student

[15] One may want to use the G8-reform as an instrument in the identification of the causal effects of instruction time. However, using an instrumental variable approach in this setting is not our preferred choice. First, we demonstrate that the reform changed instruction hours across several grades and subjects, so that the results in table 2 can all be seen as first-stage effects in instrumental variable estimations. They could be used to re-scale the reduced-form effects of the reform, reported in table 3. However, it is not clear which of the increases in instruction time constitute the relevant first-stage. Second, instrumenting total instruction hours with the G8-reform may violate the exclusion restriction, especially if other channels than school instruction time play a role. While we argue in the previous section that channels other than school instruction time play only a minor role, we cannot entirely rule out that the reform operates through other channels. If spill-over effects across subjects exist, instrumenting subject-specific instruction hours with the G8-reform may also violate the exclusion restriction. Our regression results in table 5, and also findings by Rivkin & Schiman (2015), are consistent with spill-over effects between subjects.

performance can still be improved at the given level. As the level of instruction hours in Germany before the reform is very similar to many other OECD countries (OECD, 2015), the German experience is informative for other countries. However, our estimated treatment effects may be too optimistic for school systems without tracking. Compared to other countries, the German school system tracks students relatively early into different school types according to their ability. Lavy (2015) finds that effects of instruction time are smaller in school systems without tracking. In addition, in systems without tracking, classroom heterogeneity in student ability is larger, thus the variation of treatment effects across the student performance distribution may even be wider if additional classroom time is spent on new content. Furthermore, the benefits of more instruction time may also be smaller in less favourable classroom environments (Rivkin & Schiman, 2015). The G8-reform affected the high-ability school track, in which the quality of teachers and the peer environment is considered to be better than in alternative school tracks.

VIII Conclusion

Even though instruction time is a key lever in education systems, little research examines its causal effects on student performance. Most quasi-experimental settings identify effects of small and short-lived variations in instruction time. The existing literature concentrates on average treatment effects of more instruction time, but no study looks at the effects across the distribution of student performance. We address these issues and examine the impact of a substantial and lasting increase in instruction hours across the performance distribution. We derive the effects of more instruction time from the German G8-reform, and estimate reform effects on students' PISA test scores in reading, mathematics, and science of students in ninth grade.

The reform significantly increased average test scores in reading, mathematics, and science. However, the increase in student performance appears rather small in relation to the substantive increase in instruction time. Quantile regressions reveal that treatment effects increase almost monotonically across the performance distribution. While the effects are very small and insignificant in the lower part of the distribu-

tion, students further up benefit the most from additional instruction time. This pattern persists across the three PISA domains and various model specifications. As a consequence, the reform widens the gap between low- and high-performing students.

The effect pattern across the performance distribution may be related to the content covered during additional instruction time. Our study estimates the effects of increased instruction time devoted to additional learning content - a setting in which the effect variations across the performance distribution may be particularly pronounced. We encourage future research to further examine the role of the content in additional instruction time, and to re-examine the effects on the student performance distribution in other institutional contexts. One important question is whether the effect pattern across the performance distribution is less pronounced, or even reversed, if additional time is spent on revising and deepening the curriculum.

This study carries important implications for policy-makers. Our findings can be used to relate the effects of more instruction time to the effects of changes in other school input factors, which may ultimately allow to carry out cost-effectiveness analyses. Regarding the hopes of policy-makers associated with increases in instruction time, this study demonstrates that student performance can indeed be improved. However, increases in instruction time may also widen the gap between low- and high-performing students. The content of the additional time that students spend in the classroom should be carefully considered.

References

Agasisti, T. & Longobardi, S. (2014). Inequality in education: Can Italian disadvantaged students close the gap? *Journal of Behavioral and Experimental Economics*, *52*, 8–20.

Allensworth, E., Nomi, T., Montgomery, N., & Lee, V. E. (2009). College preparatory curriculum for all: Academic consequences of requiring Algebra and English I for ninth graders in Chicago. *Educational Evaluation and Policy Analysis*, *31*(4), 367–391.

Andrietti, V. (2015). The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment. *Universidad Carlos III de Madrid Working Papers*, *2015*(September), 1–41.

Aucejo, E. M. & Romano, T. F. (2014). Assessing the effect of school days and absences on test score performance. *CEP Discussion Paper*, *1302*.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235–1264.

Baumert, J. (2009). Programme for International Student Assessment 2000 (PISA 2000). Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. *Max-Planck-Institut für Bildungsforschung (MPIB)*, http://doi.org/10.5159/IQB_PISA_2000_v1.

Bellei, C. (2009). Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile. *Economics of Education Review*, *28*(5), 629–640.

Cameron, C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, *90*(3), 414–427.

Card, D. & Krueger, A. B. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy*, *100*(1), 1–40.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2015). The aftermath of accelerating Algebra: Evidence from district policy initiatives. *Journal of Human Resources*, *50*(1), 159–188.

Cortes, K. E. & Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of double-dose algebra on student achievement. *American Economic Review: Papers & Proceedings*, *104*(5), 400–405.

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

Dahmann, S. (2015). How does education improve cognitive skills? Instructional time versus timing of instruction. *SOEPpapers on Multidisciplinary Panel Data Research*, *769*.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, *42*(3), 528 – 554.

Dustmann, C., Puhani, P. A., & Schönberg, U. (2014). The long-term effects of early track choice. *The Economic Journal*, (forthcoming).

Eide, E. R. & Showalter, M. H. (2012). Sleep and student achievement. *Eastern Economic Journal*, *38*(4), 512–524.

Fitzpatrick, M. D., Grissmer, D., & Hastedt, S. (2011). What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review*, *30*(2), 269–279.

Goodman, J. S. (2014). Flaking out: Student absences and snow days as disruptions of instruction time. *NBER Working Paper*, *20221*.

Grogger, J. (1996). Does school quality explain the recent black/white wage trend? *Journal of Labor Economics*, *14*(2), 231–53.

Herrmann, M. A. & Rockoff, J. E. (2012). Worker absence and productivity: Evidence from teaching. *Journal of Labor Economics*, *30*(4), 749–782.

Huebener, M. & Marcus, J. (2015). Moving up a gear: The impact of compressing instructional time into fewer years of schooling. *DIW Discussion Paper*, *1450*.

Klieme, E. (2013). Programme for International Student Assessment 2009 (PISA 2009). Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. *Deutsches Institut für Internationale Pädagogische Forschung*, http://doi.org/10.5159/IQB_PISA_2009_v1.

KMK (2013). Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II. Beschluss der Kultusministerkonferenz vom 07.07.1972 i.d.F. vom 06.06.2013. Technical report, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany), Bonn/Berlin.

Lavy, V. (2012). Expanding school resources and increasing time on task: Effects of a policy experiment in Israel on student academic achievement and behavior. *NBER Working Paper*, *18369*.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, *125*(588), F397–F424.

Lee, J.-W. & Barro, R. J. (2001). Schooling quality in a cross-section of countries. *Economica*, *68*(272), 465–488.

Lüdemann, E. & Schwerdt, G. (2013). Migration background and educational tracking: Is there a double disadvantage for second-generation immigrants? *Journal of Population Economics*, *26*(2), 455–481.

Marcotte, D. E. (2007). Schooling and test scores: A mother-natural experiment. *Economics of Education Review*, *26*(5), 629–640.

Marcotte, D. E. & Hemelt, S. (2008). Unscheduled closings and student performance. *Education Finance and Policy*, *3*(3), 316–338.

Mueller, S. (2013). Teacher experience and the class size effect – Experimental evidence. *Journal of Public Economics*, *98*, 44–52.

OECD (2015). Education at a glance 2015: OECD indicators. *OECD Publishing*.

OECD (2016a). How is learning time organised in primary and secondary education? *Education Indicators in Focus*, *38*, OECD Publishing, Paris.

OECD (2016b). Student learning time: A literature review. *OECD Education Working Papers*, *127*, OECD Publishing, Paris.

Patall, E. A., Cooper, H., & Allen, A. B. (2010). Extending the school day or school year: A systematic review of research (1985-2009). *Review of Educational Research*, *80*, 401–436.

Prenzel, M. (2007). Programme for International Student Assessment 2003 (PISA 2003). Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. *Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik an der Universität Kiel*, http://doi.org/10.5159/IQB_PISA_2003_v1.

Prenzel, M. (2010). Programme for International Student Assessment 2006 (PISA 2006). Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. *Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik an der Universität Kiel*, http://doi.org/10.5159/IQB_PISA_2006_v1.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., & Al., E. (2006). *PISA 2003 - Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Münster: Waxmann.

Rangvid, B. S. (2007). School composition effects in Denmark: Quantile regression evidence from PISA 2000. *Empirical Economics*, *33*(2), 359–388.

Rivkin, S. G. & Schiman, J. C. (2015). Instruction time, classroom quality, and academic achievement. *The Economic Journal*, *125*(588), F425–F448.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Sims, D. P. (2008). Strategic responses to school accountability measures: It's all in the timing. *Economics of Education Review*, *27*(1), 58–68.

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162–181.

Todd, P. E. & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33.

Todd, P. E. & Wolpin, K. I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital*, *1*(1), 91–136.

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

Figures

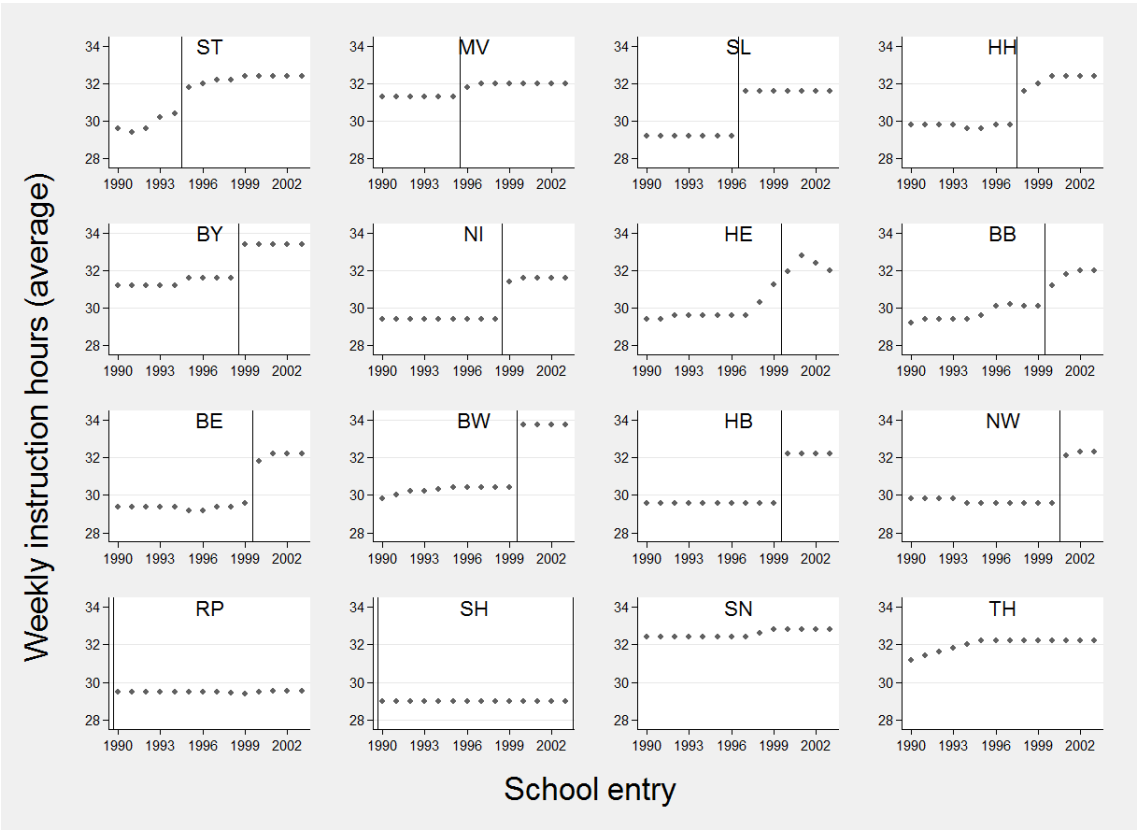

Figure 1: Number of weekly instruction hours by school entry cohort (averaged over grades 5 to 9). In the order of reform introduction: ST: Saxony-Anhalt, MV: Mecklenburg-Vorpommern, SL: Saarland, HH: Hamburg, BY: Bavaria, NI: Lower-Saxony, HE: Hesse, BB: Brandenburg, BE: Berlin, BW: Baden-Württemberg, HB: Bremen, NW: North Rhine-Westphalia, RP: Rhineland-Palatinate, SH: Schleswig-Holstein, SN: Saxony, TH: Thuringia.
Source: Official historical time-table regulations, own calculations.

Tables

Table 1: Descriptive statistics of the main sample

Variable	Mean	SD
PISA test scores		
Reading	573.75	(60.42)
Mathematics	579.33	(61.43)
Science	585.29	(65.08)
Average weekly instruction hours, grade 5-9		
Total	30.96	(1.48)
German	4.22	(0.13)
Mathematics	4.04	(0.20)
Science	3.55	(0.61)
Other subjects	19.14	(1.39)
Socio-economic characteristics		
Female, dummy	0.54	(0.50)
Migrant, dummy	0.13	(0.34)
Age in years	15.38	(0.46)
Grade repeated, dummy	0.07	(0.26)
High parental education (ISCED $\geq$ 5)	0.64	(0.48)
School characteristics		
School size	850.44	(309.82)
Public school, dummy	0.91	(0.29)
Share of part-time teachers	0.36	(0.18)
Student-computer-ratio	31.68	(67.91)
Student-teacher-ration	16.69	(4.28)
G8-reform, dummy	0.38	(0.49)
Number of federal states	16	
Number of schools	1322	
Number of students	33217	

Note: The table reports descriptive statistics of the main sample, weighted by PISA sampling weights. Standard deviations are reported in parentheses. The sample includes all academic track schools in the German PISA data for 2000, 2003, 2006, 2009, and 2012.

Table 2: Estimates of the G8-reform effect on weekly instruction hours

	(1)	(2)	(3)	(4)	(5)	(6)
	Average change in weekly instruction	Change in weekly instruction hours by grade				
Subject	hours in grades 5 to 9	grade 5	grade 6	grade 7	grade 8	grade 9
All subjects	1.99***	1.94***	1.62***	1.66**	2.09***	2.65***
	(0.44)	(0.46)	(0.41)	(0.69)	(0.54)	(0.46)
$\% - change$	6.53	6.79	5.44	5.32	6.66	8.37
German	0.02	0.07	-0.08	0.04	-0.21**	0.29*
	(0.06)	(0.11)	(0.19)	(0.11)	(0.08)	(0.14)
Mathematics	0.10*	0.27	0.10	0.03	-0.11	0.21
	(0.06)	(0.16)	(0.15)	(0.09)	(0.19)	(0.21)
Science	0.62***	0.28	0.35	0.79*	1.10***	0.59
	(0.16)	(0.17)	(0.28)	(0.39)	(0.24)	(0.44)
Other subjects	1.25**	1.31*	1.25*	0.79	1.31**	1.56*
	(0.52)	(0.62)	(0.68)	(0.61)	(0.47)	(0.81)
N	33217					

Note: OLS regressions include federal state- and cohort-fixed effects. G8-reform effects estimated in separate regressions. Standard errors are reported in parentheses and allow for clustering at the federal state level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 3: Main results: OLS and quantile regression estimates of the G8-reform effect on student performance

Dependent variable: Domain specific PISA score										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	OLS	Quantile regressions								
		q=0.1	q=0.2	q=0.3	q=0.4	q=0.5	q=0.6	q=0.7	q=0.8	q=0.9
Reading										
G8-reform	5.76***	2.92	4.59***	4.15*	5.77**	6.01***	6.65**	7.56***	8.30***	7.93***
	(1.91)	(2.25)	(1.79)	(2.42)	(2.13)	(2.15)	(2.94)	(3.03)	(3.19)	(2.98)
Mathematics										
G8-reform	5.26**	1.95	0.62	3.18	4.96	5.56**	6.72**	7.87***	8.49***	8.34***
	(2.55)	(3.32)	(2.95)	(2.61)	(3.19)	(2.87)	(3.00)	(3.06)	(2.87)	(3.32)
Science										
G8-reform	5.71*	1.95	3.24	4.28	5.20	6.63*	7.31**	7.79**	7.85**	7.58***
	(2.99)	(3.59)	(3.42)	(3.77)	(3.48)	(3.72)	(3.55)	(3.54)	(3.45)	(3.05)
N	33217	33217	33217	33217	33217	33217	33217	33217	33217	33217

Note: OLS and quantile regressions include federal state-fixed effects, cohort-fixed effects, and socioeconomic controls (highest parental education, quadratic term for student age, migration background, gender). Standard errors are reported in parentheses and allow for clustering at the federal state level. Clustered standard errors for quantile regressions are bootstrapped (200 replications). Estimations apply PISA sampling weights and consider the five plausible values per domain for each student, as suggested in the PISA technical reports. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 4: Heterogeneity analyses: Subsample OLS estimates of the G8-reform effect on student performance

Dependent variable: Domain specific PISA score						
	(1)	(2)	(3)	(4)	(5)	(6)
	Gender		Migration background		Parental education	
Subsample:	Girls	Boys	Natives	Migrants	ISCED<5	ISCED≥5
Reading						
G8-reform	6.24*	5.10*	5.46***	6.58	4.84	6.22***
	(3.20)	(2.80)	(2.04)	(6.32)	(3.69)	(1.78)
Mathematics						
G8-reform	5.80	4.20	5.68**	2.35	6.86*	4.41*
	(3.81)	(3.22)	(2.50)	(6.93)	(3.79)	(2.56)
Science						
G8-reform	5.65	5.54*	6.51**	0.86	7.57*	4.80*
	(4.10)	(3.11)	(3.28)	(7.37)	(4.53)	(2.86)
N	17990	15227	27820	5397	12301	20916

Note: Subsample OLS regressions include federal state-fixed effects, cohort-fixed effects, and socioeconomic controls (highest parental education, quadratic term for student age, migration background, gender). Standard errors are reported in parentheses and allow for clustering at the federal state level. Estimations apply PISA sampling weights and consider the five plausible values per domain for each student. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 5: OLS estimates of the effect of instruction hours on student performance

Dependent variable: Domain specific PISA score			
	(1) Reading	(2) Mathematics	(3) Science
Total hours, grade 5-9			
German	3.73*	-1.38	-1.30
	(1.97)	(1.99)	(1.78)
Mathematics	1.03	4.69**	4.23**
	(1.56)	(2.09)	(1.78)
Science	0.33	-0.47	0.02
	(0.43)	(0.70)	(0.65)
Other subjects	0.54**	0.46*	0.38
	(0.24)	(0.24)	(0.29)
N	33217	33217	33217

Note: OLS regressions include federal state-fixed effects, cohort-fixed effects, and socioeconomic controls (highest parental education, quadratic term for student age, migration background, gender). Standard errors are reported in parentheses and allow for clustering at the federal state level. Estimations apply PISA sampling weights and consider the five plausible values per domain for each student. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 6: OLS estimates of the G8-reform effect on student composition, full-time teacher share and student-teacher-ratio.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent variable:						
	Girls	Parents with ISCED≥5	Migrants	Age	Grade repeated	Share of full time teachers	Student-teacher-ratio
G8-reform	-0.00	-0.01	-0.01	0.02	0.00	0.05	-0.34
	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)	(0.05)	(1.47)
N	33217	33217	33217	33217	32990	29475	28229

Note: OLS regressions include federal state-fixed effects and cohort-fixed effects. Standard errors are reported in parentheses and allow for clustering at the federal state level. Estimations apply PISA sampling weights and consider the five plausible values per domain for each student. 227 students in our sample do not provide information on their grade repetition history. For 3742 students, we lack information on the school share of full time teachers, and for 4988 students, we lack information on the student-teacher-ratio. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 7: Sensitivity checks: OLS estimates of the G8-reform effect for alternative model specifications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Placebo treatments		Other reforms			Control variables		
	Main	Treatment one period earlier	Treatment in other school tracks	Central exit exams	Tracking after grade 6	Reduced no. of tracks	No controls	Full sample[a]	Individual & school level controls
Reading									
G8-reform	5.76***	-0.23	-0.51	6.04**	5.82***	5.11**	5.73***	5.74***	6.43***
	(1.91)	(2.47)	(2.59)	(2.35)	(2.02)	(2.06)	(2.00)	(2.02)	(2.29)
Mathematics									
G8-reform	5.25**	-1.47	-0.96	4.87*	4.09*	5.17*	5.19*	5.40*	6.31**
	(2.62)	(2.57)	(2.77)	(2.51)	(2.43)	(2.85)	(2.98)	(3.18)	(2.50)
Science									
G8-reform	5.82*	-0.90	1.21	5.44**	5.16*	6.15**	5.75*	5.89*	5.94**
	(2.99)	(3.42)	(3.38)	(2.69)	(2.98)	(2.93)	(3.10)	(3.08)	(3.02)
N	33217	33217	67755	33217	33217	33217	33217	35557	33217

Note: OLS regressions include federal state-fixed effects and cohort-fixed effects, and socioeconomic controls (highest parental education, quadratic term for student age, migration background, gender) unless stated differently. School level controls include student-teacher-ratio, student-computer-ratio, school size, public school. Standard errors are reported in parentheses and allow for clustering at the federal state level. Estimations apply PISA sampling weights, and consider the five plausible values per domain for each student. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

[a] The sample size for reading is 36644, for mathematics 35894 and for science 35557.

A Appendix

Table A.1: Implementation of G8 and other education reforms in the federal states

	G8	Central exit exams	Tracking after grade 6	Two-tier system
Change from G9 to G8				
Saxony-Anhalt (ST)	from 1995	all	1993-1997	from 1993
Mecklenburg-Vorpommern (MV)	from 1996	all	from 1999	from 1998
Saarland (SL)	from 1997	all	none	from 1993
Hamburg (HH)	from 1998	from 1992	none	none
Bavaria (BY)	from 1999	all	none	none
Lower-Saxony (NI)	from 1999	from 1993	until 1997	none
Baden-Württemberg (BW)	from 2000	all	none	none
Bremen (HB)	from 2000	from 1994	until 1998	from 2000
Berlin (BE)	from 2000	from 1994	all	none
Brandenburg (BB)	from 2000	from 1992	all	from 2000
Hesse (HE)	from 2000	from 1994	none	none
North Rhine-Westphalia (NW)	from 2001	from 1994	none	none
Always G8				
Saxony (SN)	all	all	none	all
Thuringia (TH)	all	all	none	all
Always G9 (during the sample period)				
Rhineland-Palatinate (RP)	none	none	none	none
Schleswig-Holstein (SH)	from 2004	from 1995	none	none

Notes: The table reports how the school entry cohorts in our sample period, 1991-2003, are affected by different education reforms and institutional changes. The official abbreviations of the federal states are reported in parentheses for later reference. *Centralised school exit examinations* shift the design of exit exams from high schools to federal state institutions such that all students in the specific state sit the same exit exam. *Tracking after grade 6* indicates reforms that changed the age at which students are tracked. *Two-tier system* indicates reforms that combine the low and middle track in the traditional German three-tier school track system. Sources for the reform dates are available from the authors upon request.

Table A.2: Comparing instruction hour information provided in PISA data to official timetable regulations.

Survey year	PISA question	PISA data	Enacted regulations
2000	"In the last full week you were in school, how many instruction hours *(each 45 minutes)* did you spend in ...?"		
	German	3.28 (0.66)	3.36 (0.33)
	Mathematics	3.57 (0.71)	3.64 (0.36)
	Science	5.32 (1.49)	5.07 (0.73)
2003	"In the last full week you were in school, how many instruction hours *(each 45 minutes)* did you have in **total**?"	30.60 (3.28)	31.40 (1.06)
	"In the last full week you were in school, how many instruction hours *(each 45 minutes)* did you spend in **Mathematics**?"	3.68 (0.73)	3.60 (0.42)
2006	"How much time do you typically spend per week studying the following subjects in regular lessons?" (Categories: "No time", "<2 hours", "2 to <4 hours", "4 to <6 hours", "≥6 hours", one hour corresponds to 60 rather than 45 minutes, the length of a usual German instruction hour)		
	German (share with "2 to <4 hours")	0.62 (0.49)	1.00 (0.00)
	Mathematics (share with "2 to <4 hours")	0.55 (0.50)	1.00 (0.00)
	Science (share with "2 to <4 hours")	0.32 (0.47)	0.38 (0.49)
2009	"In a normal, full week at school, how many instruction hours *(each 45 minutes)* do you have in **total**?"	33.22 (2.49)	33.25 (1.81)
	"How many instruction hours *(each 45 minutes)* per week do you typically have for the following subjects?"		
	German	3.71 (0.58)	3.68 (0.37)
	Mathematics	3.73 (0.58)	3.79 (0.32)
	Science	5.52 (1.29)	5.57 (0.73)
2012	"In a normal, full week at school, how many instruction hours *(each 45 minutes)* do you have in **total**?"	33.91 (3.28)	33.91 (1.27)
	"How many instruction hours *(each 45 minutes)* per week do you typically have for the following subjects?"		
	German	3.75 (0.77)	3.59 (0.45)
	Mathematics	3.81 (0.77)	3.80 (0.30)
	Science	5.68 (1.30)	5.81 (0.57)

Note: The table reports the mean of information on instruction hours from PISA data and of official timetable regulations matched to the PISA data. Standard deviations are reported in parentheses. Prior to the comparison, the PISA data on subject-specific instruction hours is set to missing for implausible values as done by Rivkin & Schiman (2015). We remove observations that report numbers of weekly classes exceeding 10, or equalling zero, which is implausible given the binding timetable regulations. The official timetable regulations are very similar to information in the provided PISA data but for PISA 2006. Information in PISA 2006 raise concerns about substantial measurement error, as the instruction hour question related to hours corresponding to 60 minutes, rather than instruction hours that typically last 45 minutes in Germany. While in other PISA waves, about 95 percent of mathematics hours fall in the "2 to <4 hours" category, in 2006 the distribution is more evenly split across the different categories. This has also been noted by Rivkin & Schiman (2015) in international PISA data.

Table A.3: Heterogeneity analysis: Subsample quantile estimates of the G8-reform effect on student performance

Dependent variable: Domain specific PISA score									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5	q=0.6	q=0.7	q=0.8	q=0.9
Gender: Girls [N=17990]									
Reading	3.36	3.10	2.45	5.10	5.41*	6.90**	8.62***	10.97***	9.30*
	(4.35)	(4.46)	(3.79)	(3.67)	(2.87)	(3.05)	(2.98)	(3.79)	(5.40)
Mathematics	3.34	1.57	3.21	4.87	6.17	8.28**	8.98**	9.81***	8.20
	(5.64)	(4.07)	(3.43)	(4.07)	(3.89)	(3.66)	(3.92)	(3.47)	(5.55)
Science	0.33	2.90	3.94	6.23	6.94*	8.49**	8.93***	7.60**	5.40
	(6.49)	(4.56)	(3.96)	(4.59)	(4.05)	(4.00)	(3.34)	(3.41)	(6.62)
Gender: Boys [N=15227]									
Reading	2.66	5.83	6.45*	6.07	6.59*	7.08*	5.29	5.18	5.27
	(6.51)	(4.68)	(3.62)	(4.89)	(3.45)	(3.62)	(4.12)	(4.30)	(4.49)
Mathematics	-1.52	-0.46	2.80	4.68	3.70	4.94	6.69*	5.90	7.45
	(5.34)	(4.86)	(4.10)	(3.65)	(3.47)	(3.02)	(3.75)	(3.92)	(5.70)
Science	1.39	3.56	3.39	4.03	5.23	5.52	7.72*	7.57	9.23
	(5.37)	(4.97)	(3.99)	(4.11)	(3.68)	(4.06)	(4.29)	(5.56)	(6.36)
Migration background: Natives [N=27820]									
Reading	1.81	4.32	3.81	5.14**	5.69**	6.69**	7.25***	8.50**	8.45**
	(3.40)	(2.88)	(2.64)	(2.48)	(2.62)	(2.88)	(2.62)	(3.43)	(3.80)
Mathematics	2.04	0.72	3.18	5.11*	5.98**	7.04***	8.45***	9.24***	8.37**
	(4.66)	(3.33)	(3.22)	(2.64)	(2.40)	(2.57)	(2.94)	(3.20)	(4.17)
Science	2.59	4.25	5.43	6.62**	7.09***	7.76***	8.22***	7.54**	8.99**
	(4.19)	(3.71)	(3.39)	(3.18)	(2.54)	(2.54)	(2.61)	(3.10)	(3.81)
Migration background: Migrants [N=5397]									
Reading	9.47	7.14	6.45	4.33	4.46	5.34	7.71	8.82	5.76
	(9.63)	(7.27)	(7.02)	(8.70)	(8.61)	(6.59)	(6.46)	(9.13)	(6.68)
Mathematics	2.76	-0.61	0.28	3.17	1.06	2.75	5.87	4.18	6.28
	(12.48)	(10.15)	(7.91)	(5.63)	(8.04)	(7.54)	(7.67)	(7.09)	(9.26)
Science	-3.79	-4.10	-3.49	-1.76	0.11	3.81	6.64	4.51	0.98
	(11.05)	(7.97)	(8.47)	(8.45)	(8.49)	(7.26)	(6.81)	(6.98)	(10.45)

Table A.3 continued on the next page

Table A.3 – continued from the previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5	q=0.6	q=0.7	q=0.8	q=0.9
Parental education: ISCED<5 [N=12301]									
Reading	-0.33	1.35	1.74	4.08	6.33	6.56*	7.87**	9.61**	9.53
	(4.71)	(4.79)	(4.93)	(3.74)	(4.46)	(3.40)	(3.93)	(4.88)	(6.07)
Mathematics	2.28	2.71	4.30	5.59	5.37	8.26**	9.77**	10.21*	12.49*
	(6.15)	(4.51)	(4.01)	(3.96)	(3.47)	(3.38)	(4.02)	(5.55)	(6.44)
Science	2.07	5.88	6.89	7.92	8.55*	9.52***	11.52***	9.63	8.80
	(5.81)	(5.03)	(5.91)	(5.43)	(4.39)	(3.39)	(4.24)	(6.47)	(6.49)
Parental education: ISCED≥5 [N=20916]									
Reading	4.54	6.91*	5.84*	6.63*	6.28**	6.50*	6.68**	8.21**	6.92
	(4.29)	(3.71)	(3.43)	(3.81)	(2.99)	(3.55)	(3.35)	(3.42)	(4.51)
Mathematics	1.72	0.61	2.05	4.54	5.02*	6.40**	6.76*	7.10**	6.16
	(4.72)	(4.96)	(4.42)	(2.90)	(2.99)	(2.69)	(3.85)	(3.16)	(5.04)
Science	2.85	2.14	2.56	4.28	5.29	5.65	6.93*	6.76*	6.89*
	(5.33)	(3.79)	(3.18)	(3.77)	(3.47)	(4.61)	(3.72)	(3.67)	(3.76)

Note: Subsample quantile regressions include federal state-fixed effects, cohort-fixed effects, and socioeconomic controls (highest parental education, quadratic term for student age, migration background, gender). Conventional standard errors are reported in parentheses. Estimations apply PISA sampling weights, and consider the five plausible values per domain for each student. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A.4: Sensitivity checks: Quantile estimates of the G8-reform effect for alternative model specifications

Dependent variable: Domain specific PISA score									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5	q=0.6	q=0.7	q=0.8	q=0.9
Placebo treatment: Treatment one period earlier [N=33217]									
Reading	-2.32	-2.19	-0.09	0.78	0.16	1.21	1.77	2.05	2.24
	(3.21)	(3.17)	(3.04)	(2.64)	(2.68)	(2.83)	(2.28)	(2.18)	(3.01)
Mathematics	2.34	0.47	-0.54	-1.05	-1.77	-2.65	-3.68	-2.87	-2.99
	(3.20)	(3.01)	(2.86)	(2.99)	(3.06)	(3.42)	(2.68)	(2.72)	(3.98)
Science	-0.36	0.18	0.66	0.06	-0.80	-1.22	-2.28	-3.09	-3.29
	(3.39)	(2.97)	(2.79)	(2.36)	(2.56)	(2.28)	(3.33)	(2.59)	(3.82)
Placebo treatment: Treatment in other school tracks [N=67755]									
Reading	-6.79*	-2.47	-0.28	0.94	2.55	2.69	2.99	0.11	0.02
	(3.72)	(3.23)	(2.99)	(3.00)	(2.52)	(2.35)	(2.61)	(2.32)	(3.73)
Mathematics	-3.24	-1.80	-0.11	0.80	1.02	0.39	0.56	-0.81	-2.22
	(3.01)	(2.64)	(2.23)	(2.31)	(2.00)	(2.19)	(1.95)	(2.59)	(2.61)
Science	-1.73	1.74	3.11	2.09	1.85	2.06	2.01	2.41	0.14
	(3.09)	(3.34)	(2.72)	(3.19)	(2.72)	(2.29)	(2.71)	(2.49)	(2.89)
Other reforms: Central exit exams [N=33217]									
Reading	3.19	5.04*	4.67*	6.28**	6.28***	6.69***	7.59***	8.59***	7.86**
	(3.29)	(2.66)	(2.65)	(2.82)	(2.24)	(2.27)	(2.28)	(2.94)	(3.24)
Mathematics	1.63	0.14	2.88	4.83**	5.07**	6.48***	7.83**	8.12***	7.85*
	(3.83)	(3.07)	(3.13)	(2.33)	(2.25)	(2.27)	(3.21)	(2.82)	(4.05)
Science	1.87	3.05	3.62	4.89	5.91**	6.80**	7.60***	7.69***	7.46*
	(4.63)	(3.84)	(2.59)	(2.97)	(2.31)	(2.69)	(2.67)	(2.96)	(4.14)
Other reforms: Tracking after grade 6 [N=33217]									
Reading	3.02	4.75*	4.32*	5.72**	5.91***	6.53***	7.52***	8.69***	8.19**
	(3.48)	(2.74)	(2.60)	(2.70)	(2.16)	(2.38)	(2.21)	(2.72)	(3.53)
Mathematics	0.91	-0.85	2.11	3.60	4.26*	5.64**	6.81**	7.39***	7.27*
	(3.82)	(3.37)	(3.05)	(2.34)	(2.29)	(2.35)	(3.31)	(2.74)	(4.06)
Science	1.30	2.47	3.17	4.63	5.74**	6.57**	6.82***	6.91**	7.41*
	(4.25)	(3.95)	(2.72)	(3.05)	(2.38)	(2.92)	(2.64)	(3.14)	(4.21)
Other reforms: Reduced no. of tracks [N=33217]									
Reading	2.94	4.24	3.86	5.34**	5.55**	6.06**	6.58***	7.37**	6.71**
	(3.50)	(3.07)	(2.82)	(2.63)	(2.24)	(2.48)	(2.32)	(2.88)	(3.37)
Mathematics	2.65	1.16	3.47	5.12**	5.39**	6.45**	7.48**	7.80***	7.58*
	(3.98)	(3.46)	(3.03)	(2.45)	(2.45)	(2.59)	(3.22)	(2.75)	(3.87)
Science	2.98	3.97	5.13*	6.02*	7.20***	7.69***	7.93***	7.70***	7.44*
	(3.94)	(4.03)	(2.62)	(3.10)	(2.64)	(2.92)	(2.94)	(2.96)	(3.98)

Table A.4 continued on the next page

Table A.4 – continued from the previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	q=0.1	q=0.2	q=0.3	q=0.4	q=0.5	q=0.6	q=0.7	q=0.8	q=0.9
Control variables: No control variables [N=33217]									
Reading	3.16	4.64	5.30*	4.55	5.86**	6.17**	6.44**	7.72**	9.21***
	(3.72)	(3.09)	(2.95)	(2.94)	(2.37)	(2.87)	(2.71)	(3.03)	(3.34)
Mathematics	2.34	0.97	2.82	4.16*	6.26**	6.64**	6.91***	7.63***	7.19*
	(4.24)	(4.04)	(2.24)	(2.51)	(2.49)	(2.74)	(2.61)	(2.68)	(4.33)
Science	2.58	3.52	5.08	5.49*	5.92**	7.97***	6.69**	7.15***	6.60
	(5.62)	(3.46)	(3.34)	(2.99)	(2.81)	(2.80)	(3.29)	(2.61)	(4.71)
Control variables: Full sample									
Reading	3.03	4.67	5.86**	4.84*	5.89**	6.01**	6.29**	7.43***	8.75***
[N=36644]	(3.38)	(3.38)	(2.73)	(2.94)	(2.84)	(2.50)	(2.98)	(2.36)	(3.08)
Mathematics	3.81	2.38	3.13	4.21*	6.28**	6.81***	6.59**	7.49***	6.58*
[N=35894]	(4.42)	(3.34)	(2.56)	(2.26)	(2.55)	(2.57)	(2.59)	(2.66)	(3.50)
Science	3.25	3.78	5.03*	6.09**	5.83**	8.16***	6.76**	7.21***	7.33*
[N=35557]	(4.38)	(3.08)	(2.82)	(2.43)	(2.46)	(2.76)	(2.94)	(2.76)	(3.99)
Control variables: Individual and school level controls [N=33217]									
Reading	3.52	5.19*	4.96*	6.31**	6.87***	7.18***	7.72***	9.07***	8.60**
	(3.21)	(2.83)	(2.95)	(2.87)	(2.17)	(2.31)	(2.34)	(3.21)	(3.61)
Mathematics	2.84	1.48	4.31	5.94**	6.16***	8.05***	8.95***	9.15***	8.88**
	(3.63)	(3.55)	(2.81)	(2.53)	(2.12)	(2.20)	(2.82)	(3.29)	(4.14)
Science	2.79	3.52	4.34*	5.36*	6.61***	7.28***	7.62***	7.24**	6.86*
	(4.08)	(3.74)	(2.56)	(3.05)	(2.39)	(2.66)	(2.70)	(2.94)	(3.94)

Note: The table reports the sensitivity checks described in section VII for the quantile estimations. Quantile regressions include federal state-fixed effects and cohort-fixed effects, and socioeconomic controls (highest parental education, quadratic term for student age, migration background, gender) unless stated differently. School level controls include student-teacher-ratio, student-computer-ratio, school size, public school. Conventional standard errors are reported in parentheses. Estimations apply PISA sampling weights and consider the five plausible values per domain for each student. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A.5: Out-of-school learning activities over time in treatment and control states.

	2003	2012	Difference (2012-2003)
Homework, in hours per week			
Treatment states	7.41	5.57	-1.83***
	(4.59)	(4.08)	[0.23]
N	5885	1810	
Control states	7.12	5.20	-1.92***
	(4.66)	(4.11)	[0.58]
N	1825	287	
			DiD
Difference (treatment - control)	0.28	0.36	0.09
	[0.44]	[0.43]	[0.57]
Attending out-of-school classes or private tutoring, yes/no			
Treatment states	0.28	0.38	0.10***
	(0.45)	(0.49)	[0.02]
N	5013	1660	
Control states	0.21	0.36	0.15***
	(0.41)	(0.48)	[0.01]
N	1597	253	
			DiD
Difference (treatment - control)	0.07***	0.02	-0.05*
	[0.02]	[0.02]	[0.02]

Note: The table reports the weighted mean of out-of-school learning activities in treatment and control states. Treatment states: BB, BE, BY, BW, HB, HE, HH, MV, NW, NI, ST, SL. Control states: SH, RP, SN, TH. Standard deviations are reported in parentheses. Standard errors of the differences in means are reported in brackets and account for clustering at the federal state level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A.6: Missing class, skipping class, and arriving late for school in treatment and control states.

	2000	2012	Difference (2012-2000)
Missing school, yes/no			
Treatment states	0.24	0.03	-0.22***
	(0.43)	(0.17)	[0.01]
N	6334	2748	
Control states	0.21	0.02	-0.19***
	(0.41)	(0.14)	[0.02]
N	2254	444	
			DiD
Difference (treatment - control)	0.03	0.01	-0.03
	[0.02]	[0.01]	[0.02]
Skipping classes, yes/no			
Treatment states	0.10	0.08	-0.02
	(0.31)	(0.27)	[0.02]
N	6325	2749	
Control states	0.08	0.07	-0.01
	(0.28)	(0.26)	[0.02]
N	2248	444	
			DiD
Difference (treatment - control)	0.02*	0.01	-0.01
	[0.01]	[0.03]	[0.03]
Arriving late for school, yes/no			
Treatment states	0.25	0.23	-0.02
	(0.43)	(0.42)	[0.02]
N	6331	2753	
Control states	0.23	0.19	-0.04
	(0.42)	(0.39)	[0.04]
N	2252	444	
			DiD
Difference (treatment - control)	0.02	0.04	0.02
	[0.02]	[0.03]	[0.04]

Note: The table reports the weighted mean of students missing and skipping class, and of arriving late to school in treatment and control states in the previous two weeks prior to PISA (dummy variables yes/no). Treatment states: BB, BE, BY, BW, HB, HE, HH, MV, NW, NI, ST, SL. Control states: SH, RP, SN, TH. Standard deviations are reported in parentheses. Standard errors of the differences in means are reported in brackets and account for clustering at the federal state level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A.7: G8-reform effect on instruction hours and holidays

	(1)	(2)	(3)
	Dependent variable aggregated from grade 5-9		
	School holidays	Bank holidays	Total holidays
G8-reform	0.93	-2.00	-1.07
	(1.17)	(1.24)	(0.74)
N	33217	33217	33217

Note: OLS estimations with federal state- and cohort-fixed effects. The outcome variables vary only at the state and time level. Standard errors are reported in parentheses and allow for clustering at the federal state level. Estimations apply PISA sampling weights. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.