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*

November 1, 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 has 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. We argue that the content of additional instruction time is an important determinant to explain this pattern. The findings demonstrate 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

I Introduction

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 raise two main arguments for increasing school instruction time: First, more instruction time could improve overall student performance by providing more learning opportunities. Second, it 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 how additional classroom time should be spent and how the effects differ between low- and high-performing students.

In this paper, 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, and increased instruction hours in the remaining school years such that students will have covered a similar curriculum when they graduate from school. These students are only affected by the additional instruction hours, and not yet by the reduced length of schooling. An important feature of the increased instruction time is that it covered more learning content. We focus on the performance of students in ninth grade, when they are typically 15 years old. 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 in only one school track, 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 effects 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. This widening 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 main identification assumptions.

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. That the increased instruction time is spent on new material seems to be crucial for 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 differential effects on student performance and the potential to widen gaps in student performance when they have to decide about how the additional classroom time is spent.

This study contributes to the previous literature in three important aspects. First, we study a policy experiment in which additional classroom time was not devoted to the same content, but used for additional classroom material. This is a highly relevant policy experiment, as policy-makers are typically referring to more instruction hours covering more material when they discuss increases in instruction time to improve student performance. Second, 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 or assessment dates (e.g. Sims, 2008; Fitzpatrick et al., 2011; Carlsson et al., 2015), as well as student and teacher absences (e.g. Herrmann & Rockoff, 2012; Aucejo & Romano, 2014). Only few studies generate insights from considerable, policy-induced increases in instruction time, and they are often accompanied by changes in other school input factors or the peer environment (Bellei, 2009; Lavy, 2012; Taylor, 2014; Cortes & Goodman, 2014; Cortes et al., 2015). Our study exploits a policy reform within the same school and peer environment that led to a substantial and lasting increase in instruction hours from a level close to the OECD average (OECD, 2015). Third, the previous literature mostly focuses on average treatment effects of instruction time. Differential effects by student ability received less attention (exceptions are Bellei, 2009; Carlsson et al., 2015; Cattaneo et al., 2016), but they are very interesting from a policy perspective. Increases in instruction time with additional learning content may have very different effects on students depending on their capabilities of understanding and processing new learning content. We estimate such effects across the performance distribution and address this gap.

The remainder of this study is organised 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 outlines the empirical approach. We report the main findings in Section V, and check the sensitivity of the findings and potential channels of the reform effect in Section VI. Section VII 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 little attention, even though classroom time is an omnipresent, easy-to-manage, but also costly input factor in education systems (Patall et al., 2010; Lavy, 2015; OECD, 2016b).

The challenges involved in identifying the causal effects of instruction time on student performance may be one reason. 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. With the availability of better data sources in education research (Machin, 2014), new approaches can be applied to address this challenge.

Two approaches are predominant in the microeconomics literature on this topic. The first looks at within-student variation in subject-specific instruction time. For instance, Lavy (2015), Rivkin & Schiman (2015), and Cattaneo et al. (2016) use cross-subject variations in instruction time and control for time-invariant, student-specific characteristics in student-fixed effects models. In contrast to previous correlation analyses, these studies find a strong positive effect of instruction hours on student achievements. Despite the advantages of this econometric approach, it needs to assume that only classroom time in a certain subject affects the performance of students in the respective subject, i.e. spillovers between subjects do not exist. As these studies typically relate the current level of instruction hours to student performance, little is known about the effect of instruction hours in earlier grade levels on current performance, and about the learning content of the time in school.

The second approach exploits quasi-experimental settings to learn about 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), Fitzpatrick et al. (2011) and Carlsson et al. (2015) use school day variations induced by quasi-random assignments of school start dates or assessment dates. Herrmann & Rockoff (2012) and Aucejo & Romano (2014) identify the effects with random variations in student and teacher absence days. These quasi-experimental studies find mostly beneficial impacts of more instruction time. Even though the content of the additional classroom time is not stated explicitly, one can think of these quasi-experimental studies as identifying the effects of spending different amounts of time on a fixed curriculum. The variation in instruction time that is used in these quasi-experimental studies is comparably small and not induced by specific policies.

Only few studies identify the effects of policy-induced increases in instruction hours. Bellei (2009) evaluates the introduction of all-day schooling in Chile, which increased instruction time, but was accompanied by large investments into the school infrastructure and significant institutional changes. Lavy (2012) studies a school funding policy reform in Israel that altered weekly instruction hours, teaching budgets and the classroom time spent on core subjects. Jensen (2013) analyses a national harmonisation of school timetables in Denmark involving increases in the number of classroom hours, but also in the number of school days per year. Finally, some studies evaluate programmes in which low-performing students receive additional instruction time. Battistin & Meroni (2013) and Meroni & Abbiati (2016) evaluate an EU school funding programme directed towards low-performing schools in Italy that provided afternoon programmes for low-performing students, and specialised classes for relatively higher-performing students. Taylor (2014), Cortes & Goodman (2014) and Cortes et al. (2015) examine programmes in the US that double mathematics instruction hours for low-performing students. All examined programmes generally find positive effects on student performance. However, the increases in instruction time are often accompanied by changes in other school input factors or changes in the peer environment.

There is only little research on effect heterogeneities of increased instruction time by student ability. Based on a student-fixed effects approach and the PISA assessments for Switzerland, Cattaneo et al. (2016) find important effect differences across school ability tracks, and increases in within-school variance of student performance. In a policy experiment, Bellei (2009) finds that the introduction of all-day schooling in Chile had larger treatment effects in higher quantiles of the student performance distribution. Kawaguchi (2016) examines the effects of abandoning compulsory Saturday schooling in Japan. He finds that the socio-economic gap in student performance increases. In contrast, Carlsson et al. (2015) do not find differences in treatment effects of more school days. Banerjee et al. (2007) analyse an intervention in India providing remediation classes. In contrast to the above-mentioned studies, the additional classroom time was most beneficial for students at the bottom of the performance distribution. The different findings in the literature are indicative that the content of additional classroom time (i.e. whether the time is spent on new material or on remediation) might be important to determine which students benefit the most. Overall, we add to this literature by looking at the average and quantile treatment effects of a substantial and lasting increase in weekly instruction hours that covered additional classroom material.[1]

III The G8 academic track school reform

This study 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 for four years (for more details on the education system in Germany, see e.g. Dustmann et al., 2016). Academic track schools (*Gymnasium*) constitute the high-ability school track, and intend to prepare students for university education.[2] The quality of the teachers and the peer environment is considered high. This track 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 regulations contain the distribution of weekly instruction hours across the different school subjects and they are binding for schools.

In the last years, 13 out of 16 German federal states reduced the length of academic track schooling from nine to eight years. Table 1 provides an overview of the differential timing of the reform across states. The so-called G8-reform aimed at bringing students to the labour market earlier without significant changes to the core school

[1] Two other working papers examine the effects of the same German reform in PISA data (Andrietti, 2016; Andrietti & Su, 2016). This work has been developed independently and at the same time. The combined statistical findings of both these working papers are similar to ours.

[2] In some federal states, the university entrance qualification can also be earned in alternative school tracks that were not affected by the G8-reform. We discuss potential reform effects on the choice of the school track in Section VI.A.

curriculum. The minimum number of total instruction hours required for academic track school graduation has been kept constant (KMK, 2013). Consequently, the number of weekly instruction hours increased in the remaining school years, starting from grade 5.[3] Overall, one can think of the reform as consisting of two core elements. First, it removes the final school year. Second, it increases instruction time for each year of (academic track) schooling to cover a very similar curriculum. In this study, we focus on the second element as we look at students in grade 9, and thereby inform the literature on the effects of additional instruction time covering more learning content.[4] Therefore, our study differs in a very important aspect from other evaluations of the reform, which analyse the joint effect of fewer years of schooling and additional weekly instruction hours (see Huebener & Marcus (2015a) and Thomsen (2015) for overviews of these studies).[5]

Figure 1 plots the average number of weekly instruction hours in grades 5 through 9 for students in the school entry cohorts 1991 to 2003 for each federal state. One can see a sharp increase in weekly instruction hours following the reform implementation. The exact changes of the timetables have been determined by the education ministries of the federal states after consulting education researchers and practitioners, with the objective to best cover the previous curriculum. The average increase across federal states 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 (see Table 2). Across the different grades, the increase varies between 1.62 hours (+5.3 percent) and 2.65 hours (+8.4 percent), with the largest absolute increases in grades 8 and 9. Across grades 5 through 9, German language arts hours, which account for 13.6 percent of overall weekly instruction time, received almost no increase in instruction time under the reform. Most likely, education researchers and practitioners perceived that the required curriculum can also be covered in the given

[3] In some states tracking takes place after grade 6 (details are provided in Table 1). In these states the additional instruction hours increased from grade 7 onwards.

[4] The reform does not provide more instruction time across the life-cycle of a student.

[5] A study by Dahmann (2015) is an exception. She exploits the G8-reform to investigate the effect of instruction time on fluid and crystallised intelligence. Comparing students at age 17 in survey data, she finds positive, statistically significant effects on crystallised intelligence of boys, but not of girls. At the end of academic track schooling, i.e. when treated students experienced a similar level of instruction time in one year less of schooling, she finds no reform effects.

number of instruction hours. Mathematics hours, accounting for about 13 percent of weekly instruction time, increased by 0.1 hours per week. The subjects biology, physics, and chemistry cover 11.5 percent of the school week and increased by 0.62 hours per week. Instruction hours in other subjects, including foreign languages, history, geography, social sciences, arts, and sports, account for 62 percent of weekly instruction hours and increased by 1.25 hours per week.[6]

IV Data and empirical strategy

A. The Programme for International Student Assessment

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; Prenzel et al., 2015).[7] The data contain internationally standardised measures of student performance (PISA scores) in the three domains of reading, mathematics, and science. The PISA assessments go beyond curriculum-based assessments and examine if students can make effective use of their knowledge and skills in situations likely to be encountered outside of school.[8] Therefore, instruction hours in certain subjects cannot be mapped into the different PISA domains.[9] Each PISA domain is standardised to have an international mean of 500 and a standard deviation of 100.

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.[10] The German

[6] We have not further disentangled the timetable changes for the category *other subjects*, as there exist differences across federal states in the availability, combinations and names of other subjects.

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

[8] This is an important characteristic of the PISA assessment for this study as effects on curriculum-based assessments may be purely mechanical because students covered more of the school curriculum by the time of testing.

[9] Several studies demonstrate spillover effects between subjects (Machin & McNally, 2008; Battistin & Meroni, 2013; Rivkin & Schiman, 2015).

[10] 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.

school year usually starts in August or September, and 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.

In addition to the PISA assessment, students and school principals provide additional information in separate questionnaires. In the student questionnaire, students are asked about their instruction hours in their current grade only. We complement the PISA data with information from official timetable regulations that the federal states enact. We 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. Thereby, we obtain estimates for the reform-induced changes in instruction time. The official timetable regulations match students' reported instruction hours for grade 9 in the PISA data very well (Table A.1 in the appendix). This confirms the binding nature of the regulations, and provides confidence that the information for earlier grades is also reliable.

Descriptive statistics of our pooled sample of students are provided in Table 3. The mean PISA test scores are above the international mean of 500 because we focus on students 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 language arts, 4 instruction hours in mathematics, 3.6 instruction hours in biology, physics and chemistry, and 19.1 instruction hours in other subjects including history, geography, foreign languages, arts, music and sports. 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, and 36 percent of teachers work part-time. The average student-computer-ratio is 31.7 and the student-teacher-ratio is 16.7. Students affected by G8 constitute 38 percent of our sample.

B. Empirical strategy

In order to obtain estimates for the causal effects 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 students' PISA performance in reading, mathematics, and science with separate difference-in-differences (DiD) models. 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 or policy changes at the federal level. 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 indicators for parents' highest education level, as measured by the international standard classification of education (ISCED). In Section VI.A, 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-reform indicator stems from the differential timing of the reform across the federal states (see Table 1). 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.[11]

We estimate equation 1 with ordinary least squares (OLS), using student sampling weights provided in the PISA data. Standard errors are clustered at the federal state level, and thereby account for heteroskedasticity and correlations of the error term ε_{ist} at the federal state level (Bertrand et al., 2004).[12] Standard errors and coefficient estimates also take into account that each student has five plausible values for their PISA scores.[13]

The causal interpretation of the resulting estimates rests on three major assumptions: We have to assume that there are no compositional changes in the student body due to the reform, that the PISA scores would have followed the same trend in the treatment and control group in the absence of the reform, and that no other treatment coincides with the timing of the G8-reform across states. In Section VI we provide evidence for the plausibility of these assumptions, and discuss possible threats (e.g. through other policy changes) in detail.

While the OLS approach 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 affects low- and high-performing students differently. We perform quantile regressions to obtain a more complete description of how the conditional distribution of student performance is affected by the reform.

[11] 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.

[12] 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.

[13] In the PISA assessments, students answer only a subset of the total pool of questions. This subset differs between students. In order to deal with the missing information on questions outside the student's subset, each student is assigned five so-called plausible values for each PISA domain, which are random draws from a likely test score distribution. We deal with this multiply imputed data set 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).

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 β 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.[14]

V Results

A. OLS regressions results

Before we turn to the regression results for the effects of the G8-reform on student performance, we first inspect the development of the raw PISA scores in treatment and control groups graphically (see Figure 2). Due to the staggered implementation of the reform, the graphs compare the (control) group of states that did not change their treatment status during our period of analysis (Rhineland-Palatinate, Saxony, Schleswig-Holstein, Thuringia) with three different groups of treatment states that had implemented the reform by PISA 2006, between PISA 2006 and PISA 2009, and between PISA 2009 and PISA 2012, respectively. Before the implementation of the reform, the trends in reading (Panel A) appear similar between the control group and each of the three treatment groups. After the reform, the reading scores of all three groups of treated states improved compared to the treatment group. The pictures for mathematics and science are similar: Parallel trends before the

[14] 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 4. For the 351 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.

reform implementation, with relative improvements in the treated groups following the implementation, thus indicating a positive reform effect.[15]

Note that the graphical comparisons of the average test scores do not take into account other changes in the school system or changes in socio-economic characteristics of the student body over time that are unrelated to the reform. The regression framework outlined in equation 1 uses the full variation across cohorts and federal states, and can also control for socio-economic characteristics of the students.[16] Table 4 shows our main regression results. Column 1 reports the results for the average treatment effect of the G8-reform and it generally confirms the picture from the graphical inspection. The coefficient estimates 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. While language arts did not experience increases in instruction time, students now spent more time in several different subjects, such as history, social sciences, geography, or biology, where reading texts and writing essays is a common classroom activity. As PISA tests transferable skills, it is no surprise that the PISA reading score increases on average. Furthermore, better reading skills can help students understand mathematical problems (Machin & McNally, 2008). Our findings are therefore in line with spillover effects between subjects that have previously been observed in the literature (Machin & McNally, 2008; Battistin & Meroni, 2013; Rivkin & Schiman, 2015).[17]

To illustrate the magnitude of the reform effects, we relate them to four different quantities: the increase in PISA scores of a typical school year, the gender differences in student performance, previous studies on instruction hour effects using PISA data, and the contribution to Germany's position in international PISA-ranking tables.

[15] For the group of states that had implemented the G8-reform by PISA 2006 (Mecklenburg-Vorpommern, Saarland, Saxony-Anhalt), the pre-treatment trend in mathematics does not look very similar to the control group. These three states are rather small in terms of their population and our results are robust to excluding them.

[16] As part of the sensitivity checks in Section VI.A, we also include control variables for other education reforms that have been introduced in certain states to our main model outlined in equation 1. Our results remain robust to controlling for them.

[17] We provide some direct evidence for subject spillover effects in our setting in Table A.2 in the appendix. Note that the coefficients for the subject-specific changes in this model are only identified by the twelve reform states, and the changes across subjects may be correlated. Furthermore, the model assumes that instruction hours in grade 5 have the same effect as instruction hours in grade 9. Therefore, the effects of subject-specific instruction hours should not be over-interpreted.

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 overall increase in instruction hours. Relating our findings to other studies on instruction time using PISA data, Rivkin & Schiman (2015) and Lavy (2015) find effect sizes between 3 and 6 percent of a standard deviation for one additional instruction hour per week in subjects most closely related to the PISA domains.[18] Relating the findings to the gender gap in student performance, our point estimates for the average treatment effects also 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.[19] Next, we consider the reform impact on Germany's ranking in cross-country PISA comparisons. In PISA 2012, Germany reached on average 514 points, and was ranked below Finland (519), Canada (518), Poland (518), and Belgium (515). It was ranked above Vietnam (511), Austria (506) and Australia (504). By 2012, the reform affected about 29.7 percent of all students in Germany enrolled in grade 9.[20] Back-of-the-envelop calculations suggest that the reform contributed an increase in Germany's average PISA performance of less than 2 points.[21] The average rank of Germany in PISA 2012 would have been the same. Overall, even though the average reform effects are statistically significant, the economic significance of the average reform effects appear rather small.

Why are the obtained reform effects comparably small? Rivkin & Schiman (2015) discuss classroom quality as an important determinant for the effectiveness of addi-

[18] 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 increases in instruction time in earlier grades may matter for future learning (Rothstein, 2010). Furthermore, the identification strategy of Rivkin & Schiman (2015) and Lavy (2015) relies on the assumption of no spillover effects between subjects.

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

[20] The academic school track accommodated 34.9 percent of ninth-graders, with 85.2 percent from federal states that have introduced the reform between 2000 and 2012.

[21] Average change $= 0.297 * (5.76 + 5.26 + 5.71)/3 = 1.65$

tional instruction time. However, our setting considers the high-ability school track in which teacher qualifications and the quality of the peer environment is considered high. Low classroom quality therefore seems an unlikely explanation. Diminishing marginal returns to additional instruction hours, as suggested by Rivkin & Schiman (2015), might be another explanation for the small effect sizes if students' concentration and the capability to process new inputs declines with additional time. To see whether this is an important explanation in our setting, we compare our findings to Lavy (2012). He analyses the effect of additional instruction time in Israel where the baseline level of weekly instruction hours is higher than in our setting. Still, he finds sizeable effects.

Another explanation for the small reform impact may be the content of additional instruction time. Whereas in the Israeli case examined in Lavy (2012) the additional classroom time was also intended to cover the current curriculum in more depth, in our setting the additional instruction time covers new material. The relevance of this explanation is corroborated 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 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, and that maturity effects of when students face certain material can play a role. This might explain why the benefits of additional instruction time in our policy experiment are comparably small. In sum, the content of additional instruction time seems to be an important determinant to explain the small average effect sizes.

B. Quantile regression results

Next, we examine whether the rather small average effects mask important heterogeneities across the performance distribution. Columns 2 to 10 of Table 4 report the quantile regression results. Across all PISA domains, effect sizes are positive, but small and mostly insignificant until the third decile. The treatment effects increase 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.

Why do the results differ across the performance distribution? The content of additional instruction time seems important to explain our findings again. Students further up in the performance distribution might cope with the additional content more easily, while other students might be overburdened by new material. This argumentation is in line with findings from other studies. Kawaguchi (2016) suggest that abandoned Saturday schooling with the same national curriculum (i.e. learning the same content in fewer school days) increased the socio-economic gap in student performance. In contrast, experimental evidence by Banerjee et al. (2007) from an education intervention in India shows that remediation classes are most beneficial for students at the bottom of the performance distribution. In this setting, students spend more classroom time on the same material. This variation in findings across studies, combined with our quasi-experimental evidence on increased instruction time covering more learning content, suggests that the content of learning time is an important determinant of the benefits of additional classroom time.

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 that treatment effects increase with students' position in the performance distribution (Rangvid, 2007; Bellei, 2009; Mueller, 2013; Nicoletti & Rabe, 2014).[22]

C. Further heterogeneities

Next to the effect differences in student ability, the effects may also vary between boys and girls (e.g. Dee, 2007), and between students from low and high socio-economic backgrounds (e.g. Agasisti & Longobardi, 2014). In Table 5, we report

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

the results for subsamples stratified by gender and parental education.[23] Across the three domains of reading, mathematics, and science, the effects are very similar for girls and boys. Children from parents without a degree in higher education exhibit larger point estimates in mathematics and science, but smaller estimates in reading. This may relate to the selectivity of students from lower socio-economic status families to academic track schools. However, the small differences in the treatment effects between subgroups cannot be established with statistical significance. Overall, the findings in our setting suggest no large differences between boys and girls and children from lower and higher socio-economic status families.[24]

VI Sensitivity checks

In this section, we present a broad set of sensitivity checks. First, we concentrate on potential threats to the main identification assumptions that underlie the causal interpretation of our estimates. Second, we discuss the sensitivity of our results to changes in the model specification and sample definitions. Third, we discuss whether the reform might have worked through other channels than increased weekly 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 three main assumptions. The first assumption is that the G8-reform has not affected the composition of students attending academic track schools. As all academic track schools within a federal state were required by law to implement the reform starting with one specific cohort, students can only escape the treatment by opting for a different school track, or by moving to another federal state that has not (yet) implemented

[23] We also estimated the effects separately for students with and without migration background. However, the share of students with migration background on academic track schools is small, and the students are a highly selective group of migrants. There were no significant effect differences. The results are available on request.

[24] The quantile regression results from the subsample analyses are reported in Table A.3 in the appendix. Overall, a very similar picture emerges. For reading competencies of boys and children from families with high socio-economic status, the treatment effects are more similar across the distribution.

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 has become increasingly difficult as more federal states have implemented the reform. A general escaping behaviour should be evident from enrolment rates in academic track schools. However, Huebener & Marcus (2015b) find no evidence of reform-induced lower enrolment rates at academic track schools using administrative data on all students in Germany.[25] We confirm this finding with the PISA data. In Table 6, we run difference-in-differences regressions as outlined in equation 1 without individual control variables. In column 1, we consider students across all school tracks in the PISA data, and take an indicator for attending the academic school track as the dependent variable. The probability of attending the academic track is not affected by the reform. In columns 2 to 5, we directly check for compositional changes in the student body at academic track schools. We take observable student characteristics (gender, parental education, migration background, and age) as the dependent variable, and estimate the reform effect on these characteristics at academic track schools. All coefficient estimates 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 (2015b) 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 students' age in ninth grade (column 6 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. The results

[25] Dahmann & Anger (2014) do not find any evidence for moving between states induced by the G8-reform.

are reported in column 2 and 3 of Table 7.[26] First, we assume that the reform would have taken place one PISA-wave (three years) earlier, and add a placebo reform dummy to equation 1. A significant coefficient estimate 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. Both placebo-reforms produce coefficient estimates that are small and statistically insignificant, adding plausibility to the common trend assumption.[27]

The third main assumption is that the timing of the G8-reform does not coincide with other significant reforms that affect student performance. 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. It is important to note that our difference-in-differences identification strategy does not need to rely on the absence of other reforms, but requires that these reforms do not correlate with the introduction of the G8-reform, such that they can be accounted for by the comparison to control states. Table 1 reports the timing of other reforms across the different federal states. No other reform perfectly coincides with the introduction of G8. In columns 4 to 6 of Table 7, we add dummy variables to equation 1 for each of the reforms. Even though the inclusion of dummy variables alone may not entirely rule out their confounding influences, the robustness of our estimates suggest that the G8-indicator (varying across states and time) is sufficiently orthogonal to each of the other reforms. We further test explicitly how the G8 indicator and other policy reforms correlate, i.e. whether students affected by the G8-reform are also more

[26] The pattern for the quantile treatment effects are very similar to the main effects. The results for these and all other robustness tests are reported in Table A.4 in the appendix.

[27] We can also use the results from column 3 of Table 7 for a difference-in-differences-in-differences approach, where the triple-difference estimator is the difference between columns 1 and 3. Such an approach could account for state-time specific shocks common to different school tracks. However, the placebo check on alternative school tracks provides no evidence for the existence of such effects correlated with the G8-reform.

likely than students in the control group to be affected by other school reforms. The correlation is small in magnitude and insignificant (see Table A.5 in the appendix).

Another concern may be a federal investment programme that aimed at promoting the introduction of all-day schooling in Germany, which offers mostly voluntary afternoon school activities to children. The programme was passed in 2003, and addressed all federal states and all school types (primary schools as well as all tracks of secondary schooling) and was slowly rolled out. In PISA 2009, 20.5 percent of students in the academic track attended an all-day school, compared to 33.7 percent of students in alternative school tracks. Less than one third of affected students report using the mostly voluntary offers. We perform three tests to check whether the expansion of all-day schooling might confound our results. First, we check whether the G8-reform had an impact on PISA scores in other school tracks (see column 3 of Table 7). If the federal investment programme coincides with the G8-reform and if voluntary all-day schooling has an impact on student performance, we would expect that the G8-indicator also has an effect in alternative school tracks. However, there is no evidence for that. Second, we control for the share of all-day students in academic track schools in the federal state at the time of the PISA assessment, but it does not impact on our findings (column 6 of Table 7, information obtained from KMK, 2016). Third, we take this share of all-day students as a dependent variable in our difference-in-differences model and estimate how the G8-reform effect impacted on the share of all-day students in the federal state. The estimated coefficient is close to zero and insignificant (see column 4 of Table A.5 in the appendix).

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. Our findings are also robust to this additional set of control variables.[28] The stable estimated reform effects suggest that changes with respect to the set of control variables or sample restrictions do not threaten our findings.

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. Given that students have a restricted time budget set, the reform could affect the time they 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 provides some indication on the importance of these channels to determine the estimated effects on student performance outside the classroom. Table A.6 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 only 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. But in which other activities do students cut time if they spent more hours in school? Meyer & Thomsen (2015) investigate this in one federal state at the end of academic track schooling. These

[28] 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.

students are about three years older than students in our sample. As in the PISA data, the authors cannot find effects on homework. Further, there are no effects on sports and music activities. Some evidence suggests that students spent less time on reading, watching TV and surfing the internet, and on listening to music or doing nothing. They also spent less time on volunteering activities.[29]

Related to the time use of students is the question of whether students take up additional instruction time. 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.7 in the appendix. The propensity of students to miss class, skip class, or arrive late for school was very similar prior to the reform and did not develop differently over time between treatment and control states. There is no evidence that increases in actual instruction time lag behind increases in allocated instruction time.

Next to the observed changes in instruction time, it could also be that additional differences in the length of the school year confound the analysis. To largely rule out this possibility, we collected information on the official school holiday calendars and bank holidays of the federal states.[30] Based on the school year when students are in grade 9, we assign students the official number of school holidays, bank holidays and total holidays they were exposed to between grades 5 through 9. We take this number as the dependent variable in our difference-in-differences model and estimate the treatment effect on it.[31] The estimates of the G8-reform effects on school and bank holidays are reported in Table A.8 in the appendix. They show very small point estimates and no significant impacts on the term length.

[29] Meyer & Thomsen (2015) also find a reduction in students' probability of having a side-job. This, however, seems less important in the sample of 15 year old students in our sample.

[30] The school calendar and bank holidays vary at the federal state level. They are identical across school types. The data has been collected from Schulferien.org (2016).

[31] If the student is in grade 9 in the school year of 1999/2000, we assign the summed information for the school years 1995/1996 through 1999/2000. The assignment assumes no grade repetition and no movements between federal states.

Did the reform change the composition of the teacher body at academic track schools? If policy makers want to increase instruction hours, schools will 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 7 and 8 of Table 6, we report the reform effects on the share of full time teachers in the total teacher pool and 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.

While the classroom quality is shown to be a potentially important determinant of the returns to instruction time (Rivkin & Schiman, 2015), we cannot find evidence that the school environment and the peer groups have changed differentially between treatment and control states. The reform effects are derived from changes within a given school infrastructure and school environment, and the composition of the student body at academic track schools did not change differently between treatment and control states with the introduction of the reform. Therefore, students' peer environment is unlikely to have changed substantially. With respect to changes in the teacher quality, slow-moving labour markets for teachers and high certification standards also do not point to relevant changes. This may suggest that changes in the classroom quality do not impact our findings meaningfully.[32]

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 constituted a permanent change. Similarly, parental investments in 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.[33] Adjustments in the behaviour of students, parents, and teachers are likely to be part of the effect of increases in instruction hours.

D. External validity

The implementation of the reform facilitates contrasting developments across states, cohorts and school tracks, so that 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 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), and as the reform covers new content in the additional

[32] One may also be concerned with changes in instruction material and text books affecting the treatment effects. Textbooks vary across states, and are updated on a regular basis in treatment and control states to best accompany the curriculum of the federal states. Sometimes, only the layout changes, or the presentation of learning content is updated based on new insights from educational science. While we cannot entirely rule-out that updated classroom material has an impact on our findings, we would expect an upward bias in the already small effects as the material is meant to ease student learning.

[33] 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 is not our preferred choice. The reform changed instruction hours across several grades and subjects. Therefore, it is unclear which of the increases in instruction time would constitute the relevant first stage.

time, the German experience is informative for policy-makers in other countries that consider increases in classroom time and need to decide how additional time is spent. 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 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 high. Overall, we believe that the G8-reform generates insights that are relevant beyond the German experience.

VII Conclusion

Even though instruction time is a key lever in education systems, its causal effects on student performance are not well understood. We make three contributions to the research on this topic by examining the impact of a substantial and lasting increase in instruction hours, by highlighting the importance of the content of additional instruction time, and by providing new insights on the effects of increased instruction time on the distribution of student performance.

We derive our findings from the German G8-reform, and estimate reform effects on PISA scores in reading, mathematics, and science of students in ninth grade. The reform substantially increased instruction hours that covered new learning content. We find that the reform (i) improves average student performance; (ii) the effect sizes appear rather small; and (iii) the gap in the performance of low-performing and high-performing students widens. The small average effect sizes and the pattern across the performance distribution suggest that students need different amounts of time to learn, and that the content of instruction time may be an important determinant of its benefits for different students. Lower-performing students might need more time than better-performing students to process new learning inputs. We encourage future research to further examine the role of the content of additional instruction time, and to re-examine the effects on the student performance distribution in other institutional contexts.

This study carries important implications for policy-makers. Our findings can be used to compare 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, the magnitude of effects seems small, and increases in instruction time may also widen the gap between low- and high-performing students. Therefore, the content of additional classroom time should be carefully considered.

Acknowledgements

This paper benefited from comments and suggestions by Steven Barnett, Stefan Bauernschuster, Bernd Fitzenberger, Ludovica Gambaro, Mandy Huebener, Victor Lavy, Adam Lederer, Brian McCall, Sandra McNally, Friedhelm Pfeiffer, Ronny Scherer, Thomas Siedler, C. Katharina Spiess, Rainer Winkelmann, participants of the EALE conference 2016 in Ghent, and seminar participants in Berlin, Hamburg, Hanover, Heidelberg, Nuremberg, and London. 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.

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. (2016). The causal effects of an intensified curriculum on cognitive skills: Evidence from a natural experiment. *Universidad Carlos III de Madrid, Working Paper Economic Series*, *16-06*(April).

Andrietti, V. & Su, X. (2016). Education curriculum and student achievement: Theory and evidence. *Universidad Carlos III de Madrid, Working Paper Economic Series*, *16-07*(April).

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.

Battistin, E. & Meroni, E. C. (2013). Should we increase instruction time in low achieving schools? Evidence from Southern Italy. *IZA Discussion Paper*, *7437*.

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.

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

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.

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

Cattaneo, M. A., Oggenfuss, C., & Wolter, S. C. (2016). The more, the better? The impact of instructional time on student performance. *Leading House Working Paper Series*, (115).

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*.

Dahmann, S. & Anger, S. (2014). The impact of education on personality: Evidence from a German high school reform. *IZA Discussion Paper*, *8139*.

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. (2016). The long-term effects of early track choice. *The Economic Journal*, forthcoming.

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. (2015a). Empirische Befunde zu Auswirkungen der G8-Schulzeitverkürzung. *DIW Roundup Politik im Fokus*, *57*.

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

Jensen, V. (2013). Working longer makes students stronger? The effects of ninth grade classroom hours on ninth grade student performance. *Educational Research*, *55*(2), 180–194.

Kawaguchi, D. (2016). Fewer school days, more inequality. *Journal of the Japanese and International Economies*, *39*, 35–52.

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, Bonn/Berlin.

KMK (2016). Allgemeinbildende Schulen in Ganztagsform in den Ländern in der Bundesrepublik Deutschland - Statistik 2010 bis 2014 -. *Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.*

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*(August).

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.

Machin, S. (2014). Developments in economics of education research. *Labour Economics*, *30*, 13–19.

Machin, S. & McNally, S. (2008). The literacy hour. *Journal of Public Economics*, *92*(5-6), 1441–1462.

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.

Meroni, E. C. & Abbiati, G. (2016). How do students react to longer instruction time? Evidence from Italy. *Education Economics*, *24*(6), 592–611.

Meyer, T. & Thomsen, S. L. (2015). Schneller fertig, aber weniger Freizeit? Eine Evaluation der Wirkungen der verkürzten Gymnasialschulzeit auf die außerschulischen Aktivitäten der Schülerinnen und Schüler. *Schmollers Jahrbuch*, *135*(2015), 249–278.

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

Nicoletti, C. & Rabe, B. (2014). School inputs and skills: Complementarity and self-productivity. *IZA Discussion Paper*, *8693*.

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.

Prenzel, M., Sälzer, C., Klieme, E., Köller, O., Mang, J., Heine, J.-H., Schiepe-Tiska, A., & Müller, K. (2015). Programme for International Student Assessment 2012 (PISA 2012). Version: 2. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. *Deutsches Institut für Internationale Pädagogische Forschung*, http://doi.org/10.5159/.

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.

Schulferien.org (2016). Kalender mit Schulferien und Feiertagen. *http://www.schulferien.org/Kalender_mit_Ferien/.*

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.

Thomsen, S. L. (2015). The impacts of shortening secondary school duration. *IZA World of Labor*, *166*(July), 1–10.

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

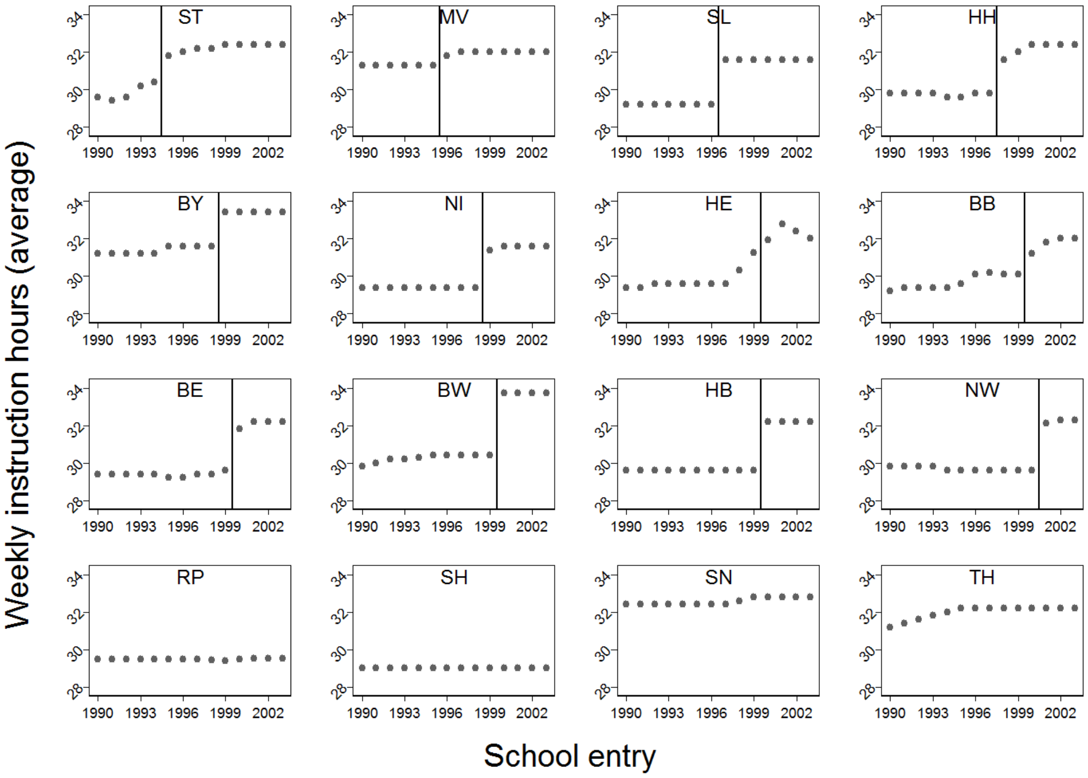

Figure 1: Number of weekly instruction hours by school entry cohort (averaged over grades 5 to 9). In the order of reform introduction: ST: Saxony-Anhalt, MV: Mecklenburg-Vorpommern, SL: Saarland, HH: Hamburg, BY: Bavaria, NI: Lower-Saxony, HE: Hesse, BB: Brandenburg, BE: Berlin, BW: Baden-Württemberg, HB: Bremen, NW: North Rhine-Westphalia. States that did not change their treatment status are: RP: Rhineland-Palatinate, SH: Schleswig-Holstein, SN: Saxony, TH: Thuringia.
Source: Official timetable regulations, own calculations.

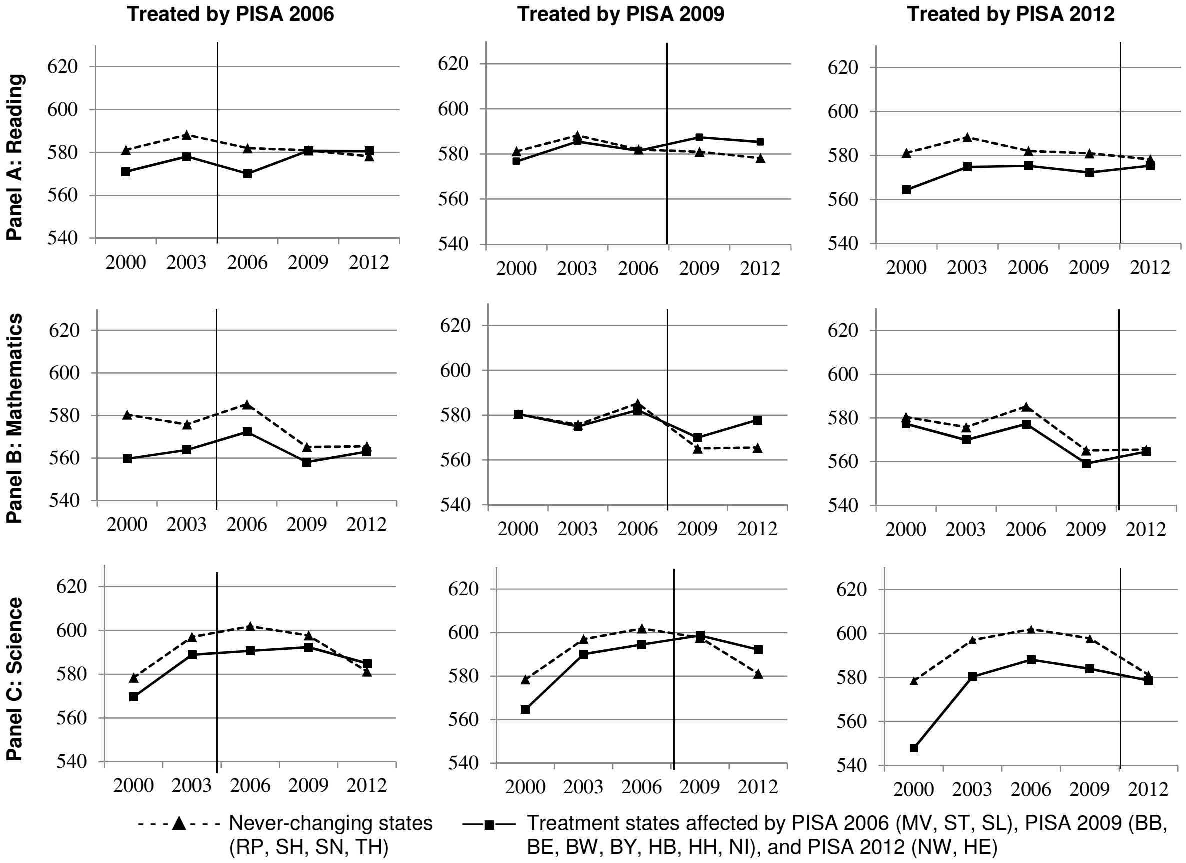

Figure 2: Development of PISA scores in the control group of states that did not change their treatment status in the period of analysis, and in treated states that implemented the reform before PISA 2006 (first column), between PISA 2006 and PISA 2009 (second column), and between PISA 2009 and PISA 2012 (third column).

Tables

Table 1: Implementation of G8 and other education reforms in the federal states by affected school entry cohort

	G8	First G8 in PISA ...	Central exit exams	Tracking after grade 6	Two-tier system
Change from G9 to G8					
Saxony-Anhalt (ST)	from 1995	2006	all	1993-1997	from 1993
Mecklenburg-Vorpommern (MV)	from 1996	2006	all	from 1999	from 1998
Saarland (SL)	from 1997	2006	all	none	from 1993
Hamburg (HH)	from 1998	2009	from 1992	none	none
Bavaria (BY)	from 1999	2009	all	none	none
Lower-Saxony (NI)	from 1999	2009	from 1993	until 1997	none
Baden-Württemberg (BW)	from 2000	2009	all	none	none
Bremen (HB)	from 2000	2009	from 1994	until 1998	from 2000
Berlin (BE)	from 2000	2009	from 1994	all	none
Brandenburg (BB)	from 2000	2009	from 1992	all	from 2000
Hesse (HE)	from 2000	2012	from 1994	none	none
North Rhine-Westphalia (NW)	from 2001	2012	from 1994	none	none
Always G8					
Saxony (SN)	all	all	all	none	all
Thuringia (TH)	all	all	all	none	all
Always G9 (during the sample period)					
Rhineland-Palatinate (RP)	none	none	none	none	none
Schleswig-Holstein (SH)	from 2004	none	from 1995	none	none

Notes: The table reports how the cohorts in our sample are affected by different education reforms and institutional changes. In order to have a common comparison base, the table refers to the year of (primary) school entry. 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.

Source: Numerous sources for the reform dates are available from the authors on request.

Table 2: G8-reform changes on weekly instruction hours

	(1)	(2)	(3)	(4)	(5)	(6)
	grades	By grade				
	5 to 9	grade 5	grade 6	grade 7	grade 8	grade 9
Average change in weekly instruction hours	1.99*** (0.44)	1.94*** (0.46)	1.62*** (0.41)	1.66** (0.69)	2.09*** (0.54)	2.65*** (0.46)
% − *change*	6.53	6.79	5.44	5.32	6.66	8.37
		By subject				
		All	Language arts	Mathematics	Biology, physics, chemistry	Others (e.g. history, geography, foreign languages)
Average change in weekly instruction hours		1.99*** (0.44)	0.02 (0.06)	0.10* (0.06)	0.62*** (0.16)	1.25** (0.52)
% − *change*		6.53	0.51	2.48	18.41	6.61
N				33217		

Notes: The table reports the G8-reform effect on weekly instruction hours for students in the main sample. Regressions include federal state- and cohort-fixed effects, and apply PISA sampling weights. Coefficient estimates obtained from separate OLS 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.

Source: Decreed timetable regulations of the federal states for academic track schools, and PISA 2000, 2003, 2006, 2009, 2012 for Germany.

Table 3: 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)
Language arts	4.22	(0.13)
Mathematics	4.04	(0.20)
Biology, physics, chemistry	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	

Notes: The table reports descriptive statistics of the main sample, weighted by PISA sampling weights. Standard deviations are reported in parentheses.
Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

Table 4: 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

Notes: The table reports OLS and quantile regression estimates of the G8-reform effect on student performance. All estimates are obtained from separate regressions including federal state-fixed effects, cohort-fixed effects, and socio-economic 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. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

Table 5: Heterogeneity analyses: Subsample OLS estimates of the G8-reform effect on student performance

Dependent variable: Domain specific PISA score				
	(1)	(2)	(3)	(4)
	Sample stratified by			
	Gender		Parental education	
	Girls	Boys	ISCED<5	ISCED≥5
Reading				
G8-reform	6.24*	5.10*	4.84	6.22***
	(3.20)	(2.80)	(3.69)	(1.78)
Mathematics				
G8-reform	5.80	4.20	6.86*	4.41*
	(3.81)	(3.22)	(3.79)	(2.56)
Science				
G8-reform	5.65	5.54*	7.57*	4.80*
	(4.10)	(3.11)	(4.53)	(2.86)
N	17990	15227	12301	20916

Notes: The table reports subsample OLS regression estimates of the G8-reform effect on student performance. All estimates are obtained from separate regressions including federal state-fixed effects, cohort-fixed effects, and socio-economic 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.

Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

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)	(8)
	Dependent variable:							
	At academ. track	Girls	Parents with ISCED≥5	Migrants	Grade repeated	Age in years	Share of full time teachers	Student-teacher-ratio
G8-reform	-0.01	-0.00	-0.01	-0.01	0.00	0.02	0.05	-0.34
	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)	(0.03)	(0.05)	(1.47)
N	100972	33217	33217	33217	32990	33217	29475	28229

Notes: The table reports OLS regression estimates of the G8-reform effect on student characteristics, the full-time teacher share and the student-teacher-ratio. All estimates are obtained from separate regressions including 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. 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.

Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
		Placebo treatments		Controlling for 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	Expansion of all-day schooling	No controls	Full sample^a	Individual & school level controls
Reading										
G8-reform	5.76***	-0.23	-0.51	6.04**	5.82***	5.11**	5.60***	5.73***	5.74***	6.43***
	(1.91)	(2.47)	(2.59)	(2.35)	(2.02)	(2.06)	(2.04)	(2.00)	(2.02)	(2.29)
Mathematics										
G8-reform	5.25**	-1.47	-0.96	4.87*	4.09*	5.17*	5.15*	5.19*	5.40*	6.31**
	(2.71)	(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**	5.75*	5.89*	5.94**
	(2.99)	(3.42)	(3.38)	(2.69)	(2.98)	(2.93)	(2.89)	(3.10)	(3.08)	(3.02)
N	33217	33217	67755	33217	33217	33217	33217	33217	35557	33217

Notes: The table reports OLS regression estimates of placebo treatments and of varying model specifications. All estimates are obtained from separate regressions including federal state-fixed effects and cohort-fixed effects, and socio-economic 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

Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

A Appendix

Table A.1: 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 ...?"		
	Language arts	3.28 (0.66)	3.36 (0.33)
	Mathematics	3.57 (0.71)	3.64 (0.36)
	Biology, physics, chemistry	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 Language arts instruction hour)		
	Language arts (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)
	Biology, physics, chemistry (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?"		
	Language arts	3.71 (0.58)	3.68 (0.37)
	Mathematics	3.73 (0.58)	3.79 (0.32)
	Biology, physics, chemistry	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?"		
	Language arts	3.75 (0.77)	3.59 (0.45)
	Mathematics	3.81 (0.77)	3.80 (0.30)
	Biology, physics, chemistry	5.68 (1.30)	5.81 (0.57)

Notes: 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.
Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

Table A.2: OLS estimates of the effect of subject specific instruction hours on student performance

Dependent variable: Domain specific PISA score			
	(1) Reading	(2) Mathematics	(3) Science
Other subjects	0.50***	0.29**	0.32
	(0.17)	(0.14)	(0.25)
Language arts	3.76**		
	(1.50)		
Mathematics		3.48	
		(2.29)	
Biology, physics, chemistry			3.32*
			(1.77)
N	33217	33217	33217

Notes: The table reports OLS regression results for average subject-specific instruction hours in grades 5 through 9. Results for each column are obtained from separate regressions including federal state-fixed effects, cohort-fixed effects, and socio-economic 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.

Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany and decreed timetable regulations.

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)
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)

Notes: The table reports subsample quantile regression estimates of the G8-reform effect on student performance. All estimates are obtained from separate regressions including 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.

Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

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
Other reforms: Expansion in all-day schooling programmes [N=33217]									
Reading	3.32	4.21	4.43*	5.85**	5.95***	6.45***	7.47***	8.29***	7.43**
	(3.71)	(2.85)	(2.66)	(2.63)	(2.04)	(2.44)	(2.23)	(2.88)	(3.25)
Mathematics	1.36	0.47	3.07	4.85**	5.39**	6.46**	7.82***	8.35***	8.17**
	(4.16)	(3.56)	(2.71)	(2.35)	(2.33)	(2.52)	(2.92)	(2.88)	(4.07)
Science	2.40	3.60	4.23	5.26*	6.61***	7.27***	7.77***	7.74**	7.55*
	(4.44)	(3.82)	(2.61)	(2.94)	(2.50)	(2.69)	(2.57)	(3.03)	(4.00)
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)

Notes: The table reports the sensitivity checks described in Section VI for the quantile estimations. All estimates are obtained from separate quantile regressions including federal state-fixed effects and cohort-fixed effects, and socio-economic controls (highest parental education, quadratic term for student age, migration background, gender) unless stated differently. School level controls include the student-teacher-ratio, the student-computer-ratio, the school size, and public school indicator. Conventional standard errors are reported in parentheses. Estimations apply PISA sampling weights and consider the five plausible values per domain for each student. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.
Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

Table A.5: OLS estimates of the G8-reform effect on other education reforms

Dependent variable in the column title				
	(1) Central exit exams	(2) Tracking after grade 6	(3) Reduced no. of tracks	(4) Share of acad. track students in all-day programme
G8-reform	0.05	-0.10	0.09	0.01
	(0.16)	(0.14)	(0.06)	(0.08)
N	33217	33217	33217	33217

Notes: This table reports estimates of the relationship between the G8-reform indicator and other education reforms as reported in Table 1 for students in the main sample. Whereas the outcome in the first three columns is binary (and given by the column header), the outcome in the fourth column consists of the share of academic track students in all-day programmes in the federal state in the school year of the PISA assessment. All 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. * p<0.1, ** p<0.05, *** p<0.01.

Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany, and KMK (2016). See Table 1 for information regarding the other education reforms.

Table A.6: 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]

Notes: 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.
Source: PISA 2003, 2012 for Germany.

Table A.7: 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]

Notes: 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.
Source: PISA 2000, 2012 for Germany.

Table A.8: 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

Notes: The table reports the estimated G8-reform effect on students' holidays. OLS estimations include federal state- and cohort-fixed effects. The outcome variables vary 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.
Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany and school holiday information provided by Schulferien.org (2016).