Huebener, Mathias; Kuger, Susanne; Marcus, Jan

Article — Accepted Manuscript (Postprint)

Increased instruction hours and the widening gap in student performance

Labour Economics

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Huebener, Mathias; Kuger, Susanne; Marcus, Jan (2017) : Increased instruction hours and the widening gap in student performance, Labour Economics, ISSN 0927-5371, Elsevier, Amsterdam, Vol. 47, pp. 15-34, https://doi.org/10.1016/j.labeco.2017.04.007

This Version is available at:
https://hdl.handle.net/10419/214645

This is the postprint of an article published in Labour Economics 47 2017 15-34
available online at: https://doi.org/10.1016/j.labeco.2017.04.007

Increased instruction hours and the widening gap in student performance

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

[a] *DIW Berlin,* [b] *Freie Universität Berlin,* [c] *DIPF Frankfurt,* [d] *Universität Hamburg*

March 2017

Abstract

Do increased instruction hours improve the performance of all students? Using PISA scores of students in ninth grade, we analyse the effect of a German education reform that increased weekly instruction hours by two hours (6.5 percent) over almost five years. In the additional time, students are taught new learning content. On average, the reform improves student performance. However, treatment effects are small and differ across the student performance distribution. Low-performing students benefit less than high-performing students. We argue that the content of additional instruction time is an important determinant explaining 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, curriculum, quantile difference-in-differences, recentered influence function
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). Policymakers 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 high hopes of policymakers 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 spend 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 so-called G8-reform reduced the length of academic track schooling by one year, while increasing instruction hours in the remaining school years such that students will have covered a similar curriculum when they graduate from school in one year less. We focus on the performance of students in ninth grade, when they are typically 15 years old. These students are only affected by the additional instruction hours, but not yet by the reduced length of schooling. An important feature of the increased instruction time is that it covered more learning content. The reform serves as a natural experiment to estimate the effect of spending two additional instruction hours per school week (+6.5 percent) in the classroom from grade 5 to grade 9, i.e. between the ages of 11 and 15. The additional instruction time totals to about 350 hours. Our analyses rely on data from the Programme for International Student Assessment (PISA), pooled across five waves from 2000 through 2012. The reform was implemented with regional and temporal variations in only one school track, which we exploit in linear difference-in-differences models to estimate average treatment effects, as well as in non-linear difference-in-differences

models to estimate quantile treatment effects.

Estimates of the average treatment effects suggest that the reform increased PISA test scores of ninth graders in reading, mathematics, and science by 5 to 6 percent of an international standard deviation. The estimated quantile treatment effects reveal that the bottom of the student performance distribution shows almost no effects, while treatment effects increase further up in the performance distribution. This widening gap between low- and high-performing students is most pronounced in mathematics and science. Our findings are robust to various model specifications. Different placebo regressions support the main identification assumption.

This study contributes to the previous literature in three important aspects: First, we study a policy experiment in which additional classroom time was devoted to additional learning content rather than the same content. This is a highly relevant policy experiment, as policymakers are typically referring to more instruction hours covering more learning content 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 (e.g. Marcotte, 2007; Sims, 2008; Fitzpatrick et al., 2011; Herrmann & Rockoff, 2012; Goodman, 2014; Carlsson et al., 2015; Aucejo & Romano, 2016). Only a 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; Cortes & Goodman, 2014; Taylor, 2014; Cortes et al., 2015). Our study exploits a policy reform within the same school environment 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 relevant from a policy perspective. Increases in instruction time with additional learning content may have 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 in the literature.

We conclude that (i) additional instruction time improves average student performance; (ii) the effect sizes are rather small given the substantial increase in instruction time; and (iii) the student performance distribution widens, especially in mathematics and science. That the increased instruction time is spent on new learning content seems to be crucial for explaining why effect sizes are small on average, and why they increase as one moves up the performance distribution. Students' existing set of skills 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 content. When policymakers decide about additional classroom time, they should be aware of the potential to widen gaps in student performance when new learning content is added to the curriculum.

Previous studies on the G8-reform mainly analyse the joint effect of fewer years of schooling and additional weekly instruction hours (see Huebener & Marcus, 2015 and Thomsen, 2015 for overviews of these studies). Dahmann (2015) is an exception: She analyses the G8-reform effect on fluid and crystallised intelligence. Comparing students at age 17 (with different levels of instruction time) in survey data, she finds positive reform effects on crystallised intelligence of boys, but not for girls. At the end of academic track schooling, after treated students attended one year less of schooling, she finds no reform effects. In our study, we focus on a different set of outcomes and look at the effects of additional instruction time on student performance at age 15 in the three PISA domains of reading, mathematics, and science. The domain-specific effects are important because policymakers have an interest in learning about effective ways to improve student competencies in certain domains. Another distinct feature of our study is that we show differential effects across the performance distribution. Two further working papers also examine the effects of the G8-reform in PISA data (Andrietti, 2016; Andrietti & Su, 2016). The work has been developed independently and at the same time. The combined statistical findings of both these working papers are similar to our core findings. Additionally, we conducted extensive archival research on official timetable regulations, as decreed by the education ministries of each federal state, allowing us to determine the exact, subject-specific instruction hour increase induced by the G8-reform, which is not provided in previous work on the reform. Furthermore, we examine numerous other

channels, in addition to instruction time, through which the reform may impact student performance. In addition, we draw on another, large data set of teachers and study reform adjustments in the teacher body.

The remainder of this study is organised as follows. Section II reviews the related literature. Section III describes the institutional setting and the 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 reform channels in Section VI. Section VII concludes.

II Related literature

Understanding the effectiveness of school input factors in increasing student performance is important for policymakers allocating 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 correlating student performance with instruction time in cross-sectional data find at most small positive, but not robust, relationships (Card & Krueger, 1992; Grogger, 1996; Lee & Barro, 2001; Woessmann, 2003). Yet, observed cross-country correlations might be confounded by other features of education systems. In individual-level data, it is students' endogenous selection into more or less instruction time that poses challenges for the identification of causal effects. Lower-performing students might receive additional instruction hours in order to revise and understand the classroom content. Better-performing students might select additional courses in subjects they like the most. With the availability of better data sources in education research (Machin, 2014), new approaches can be applied to address this challenge.

To address endogeneity problems, two approaches dominate the 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 assumes 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 both the effect of instruction hours in earlier grade levels on current performance and about the learning content of additional 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 school instruction time prior to centralised 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 (2016) identify the effects with random variations in student and teacher absence days. These quasi-experimental studies find mostly beneficial impacts of more instruction time. Although the content of the additional classroom time is not explicitly stated, one can think of these studies as identifying the effects of spending varying amounts of time on a fixed curriculum. The variation in instruction time is typically small and not induced by specific policies.

Only a 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 significant institutional changes and large investments into the school infrastructure. Lavy (2012) studies a school funding 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 a selected group of students receive additional

instruction time. Battistin & Meroni (2016) 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 of more learning time 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.

Only limited research examines 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. Bellei (2009) finds that the introduction of all-day schooling in Chile had larger 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 learning content 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 learning content.

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. After joint primary

schooling for typically four years, students in Germany are tracked into different school types according to their ability. Academic track schools (*Gymnasium*) constitute the high-ability school track and are designed to prepare students for university education. Upon the successful completion of *Gymnasium*, students earn the university entrance qualification (*Abitur*) that is required for admission to university.[1] In general, the quality of the teachers and the peer environment is considered high, with about one-third of each cohort enrolled in this track.[2] A noteworthy feature of the German education system is that each federal state enacts school track-specific timetable regulations. These regulations are binding for schools and contain the distribution of weekly instruction hours across the different school subjects.

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 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 such that previously nine school years of learning content and time are now covered in eight school years.[3] Generally, the reform affected cohorts newly entering academic track schools after primary school.[4] Overall, one can think of the reform as consisting of two core elements. First, it eliminates 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 increase in instruction time as we look at students in grade 9. Thereby we inform the literature on the effects of additional instruction time covering more learning content.

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

[2] For more details on the education system in Germany, see, e.g. Dustmann et al. (2016).

[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] Exceptions are the states of Mecklenburg-Vorpommern (MV) and Saxony-Anhalt (ST), where the first affected cohorts were already in grade 9 when they were subject to the G8-reform. Cohorts of these states captured in PISA 2006 (the first treatment cohorts) were in grade 7 (ST) and 8 (MV) when the reform was implemented. Our results are robust to excluding these surprised cohorts (see Section VI).

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. In all reform states, weekly instruction hours increase sharply with the reform implementation. The exact changes of the timetables were determined by the education ministries of the federal states after consulting education researchers and practitioners, with the objective to best map the previous curriculum to the new timetables. 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 of about 6.5 percent (see Table 2, where we report estimates of the G8-reform effects on instruction time). Across the different grades, the increase varies between 1.62 hours (+5.4 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, on average 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 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 in sum by 1.25 hours per week.[5]

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;

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

Klieme, 2013; Prenzel et al., 2015).[6] The data contain internationally standardised measures of student performance (PISA scores) in the three domains of reading, mathematics, and science. An important feature of PISA is that it goes beyond curriculum-based assessments and examines if students can make effective use of their knowledge and skills in situations likely to be encountered outside of school. Therefore, instruction hours in certain subjects cannot be directly mapped into the different PISA domains. 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.[7] We pool information over five PISA waves, obtaining a sample of 33,217 academic track students in ninth grade.[8] The German school year usually starts in August or September, with the German PISA assessments taking place in April and May. Therefore, we capture the effects of additional instruction time over a period of 4.7 school years.

In addition to the PISA assessment, students answer a separate questionnaire in which they are asked about their instruction hours in their current grade only. We complement the PISA data with information from official timetable regulations that each federal states enacts. 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 she or he lives in. The official timetable regulations match students' reported instruction hours for grade 9 in the PISA data quite well (Appendix Table A.1). This confirms the binding nature of the regulations, and provides confidence that the information for earlier grades is also reliable. We use this timetable

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

[7]In most German states, the university entrance qualification can also be earned at alternative, job-oriented academic track schools (*Berufliches Gymnasium*) that were not affected by the G8-reform. These schools typically start after grade 10 and are not contained in our analysis of students in grade 9. In the state of Baden-Württemberg, a small number of job-oriented academic track schools start in grade 7. The PISA data does not allow distinguishing between job-oriented and general academic track schools, such that we may accidentally assign the G8-treatment to untreated students. However, the fraction of these job-oriented academic track schools is small (less than 5% in Baden-Württemberg), and so is the chance that they are captured in the PISA data. Our results are robust to excluding the state of Baden-Württemberg.

[8]While the international PISA data sample 15-year old students, we focus on students in the modal grade 9 as the international PISA 2009 data for Germany includes only ninth graders.

information to estimate the effect of the G8-reform on instruction time (Table 2).

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 a 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 prior to the survey. Further, 64 percent of students have a 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

To estimate 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 average treatment effects of the reform on students' PISA performance in reading, mathematics, and science with the following difference-in-differences (DiD) model:

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

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 provides 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 applying to all federal states. The set of individual control variables, X_{ist}, contains a quadratic term for students' age, a gender dummy, a migration background dummy (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 G8-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.[9]

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).[10] Standard errors and coefficient estimates also take into account that each student has five plausible values for their PISA scores.[11]

The causal interpretation of the resulting estimates rests on three major assump-

[9]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 a treatment state.

[10]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). Inference based on wild cluster bootstrap p-values yields the same conclusions as inference based on p-values from clustered standard errors in OLS regressions (see Appendix Table A.2).

[11]Students answer only a subset of the total pool of PISA assessment questions and the subset differs between students ("multi-matrix design"). 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).

tions: 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 in detail.

While the OLS approach asks how the conditional mean of student performance is affected by the reform, the resulting average treatment effect estimates 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 estimate quantile treatment effects to learn how the G8-reform affected the distribution of student performance. Identifying quantile treatment effects in our setting requires assumptions about the counterfactual distribution of PISA scores in the absence of the G8-reform. Similar to the (mean) DiD procedure outlined above, we use changes in the distribution of PISA scores in the control group to construct the counterfactual PISA score distribution in the treatment group. Several non-linear DiD models are suggested for this purpose (see Athey & Imbens, 2006). In our main specification for quantile treatment effects, we look at unconditional quantiles by applying the method of recentered influence function (RIF) regression developed by Firpo et al. (2009), which Havnes & Mogstad (2015) applied in a difference-in-differences context (RIF-DiD). The RIF-DiD procedure looks at the fraction of students below the PISA score at a specific quantile and compares how this fraction changes in treatment and control states with the introduction of the G8 reform. RIF-DiD can be seen as a two-step procedure. In the first step, the outcome variable is dichotomized at each quantile (τ) according to the following transformation:

$$RIF_{\tau}(Y_{ist}) = Q_{\tau} + \frac{\tau - \mathbb{1}(Y_{ist} \leq Q_{\tau})}{f_Y(Q_{\tau})}, \tag{2}$$

where $RIF_{\tau}(Y_{ist})$ is the transformed PISA score of individual i in state s at time t for quantile τ. Q_{τ} denotes the unconditional PISA score at quantile τ and $\mathbb{1}(\cdot)$ is an indicator function indicating whether an individual's PISA score is below the PISA

score at quantile τ. $f_Y(Q_\tau)$ denotes the density of Y around Q_τ. In the second step, we run a DiD-style regression on these transformed (quantile-specific) outcomes to obtain quantile treatment effects. As before, we apply student sampling weights. Bootstrapped standard errors allow for clustering at the federal state level. The RIF-DiD estimator assumes that in the absence of the G8-reform, the change in the fraction of students who score below a specific PISA value Q_τ is the same in treatment and control group. As with all identification assumptions, this assumption cannot be directly tested. However, in Section VI we run placebo regressions that provide evidence for the plausibility of this identification assumption.

As a robustness check, we apply another non-linear DiD estimator for the estimation of quantile treatment effects, quantile difference-in-differences (QDiD, Athey & Imbens, 2006). QDiD builds on a different identification assumption than RIF-DiD. Essentially, it assumes that changes in PISA scores at a particular quantile would have been the same in treatment and control states, in the absence of the G8-reform. We estimate the reform effect at quantile τ of the conditional distribution with the following quantile regression model:

$$Q_{Y_{ist}}(\tau|G8_{st}, \mu_s, \kappa_t, X_{ist}) = \beta(\tau) \cdot G8_{st} + \mu_s(\tau) + \kappa_t(\tau) + X'_{ist} \cdot \lambda(\tau). \qquad (3)$$

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. We again apply student sampling weights and bootstrapped standard errors that allow for clustering at the federal state level. Havnes & Mogstad (2015) argue that the identification assumption for QDiD is more restrictive than the RIF-DiD identification assumption, particularly if pre-treatment outcome levels differ between the treatment and control group and changes in the outcome depend on the outcome level. Another advantage of the RIF-DiD estimator is that it is invariant to monotonic transformations of the outcome variable. For these reasons, we use the RIF-DiD estimator in our main specification and present the QDiD estimator for robustness purposes.[12]

[12]The estimation of the RIF-DiD model is performed with the user-written Stata command

V Results

A. Average treatment effects

We start our analysis by graphically inspecting the development of the raw PISA scores in treatment and control groups (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 for ninth graders before PISA 2006, between PISA 2006 and PISA 2009, as well as 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 control group. The pictures for mathematics and science are similar: Parallel trends before the reform implementation, with relative improvements in the treated groups following the implementation, thus indicating a positive reform effect.[13]

Note that the graphical comparisons of the average test scores do not account for other changes in the school system or for 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.[14] Ta-

rifreg, the QDiD model is estimated with Stata's *qreg* command. For the main results, we report bootstrapped standard errors that allow for heteroscedasticity and clustering at the federal state level. For the more than 500 quantile treatment effect models (for each RIF-DiD and QDiD) estimated in heterogeneity analyses and sensitivity checks, we report default standard errors to save computational resources of the remote access used for the analysis.

[13]For the states that had implemented the G8-reform prior to 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. Also note that the increase in student performance lags behind for this group. One explanation is that students captured in PISA 2006 in Saxony-Anhalt and Mecklenburg-Vorpommern were in grade 8 (ST) and 7 (MV) when the reform was implemented, i.e., they did not experience the increase in instruction hours from grade 5 onwards as other cohorts. Our results are robust to excluding these surprised cohorts (see Section VI).

[14]As part of the sensitivity checks in Section VI.A, we also control for other education reforms that were introduced in certain states.

ble 4 shows our main regression results. Column 1 reports the results for the average treatment effects 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 scores of about 5.3 to 5.8 percent of an international standard deviation. While instruction time in language arts did not experience increases, students now spend more time in several different subjects, such as history, social sciences, geography, or biology, where reading texts and writing essays are common classroom activities. At the same time, better reading skills can help students understand mathematical problems (Machin & McNally, 2008). Our findings can be rationalised with spillover effects between subjects that are observed in the literature (Machin & McNally, 2008; Rivkin & Schiman, 2015; Battistin & Meroni, 2016).[15]

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, previous studies on instruction hour effects using PISA data, the gender differences in student performance, and the contribution to Germany's position in international PISA-ranking tables. On average, one more year of schooling in Germany raises PISA 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 about 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 stan-

[15]We provide some direct evidence for the effects of subject-specific instruction hours on PISA scores in Appendix Table A.3. We replace the G8-reform dummy in equation 1 with four continuous variables: the total number of instruction hours in (i) language arts; (ii) mathematics; (iii) science; and (iv) all other subjects for grades 5 through 9. Instruction hours in language arts and other subjects increase PISA scores in reading; instruction hours in mathematics and other subjects increase PISA scores in mathematics. The picture for PISA scores in science is less clear, as science performance is mostly related to instruction hours in mathematics. Note that the coefficients on subject-specific instruction hours are only identified by timetable changes in twelve reform states and that the changes across subjects may be correlated, i.e. the number of degrees of freedom is relatively small compared to the number of coefficients to be estimated. 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.

dard deviation for one additional instruction hour per week in subjects most closely related to the PISA domains.[16] Relating the findings to the gender gap in student performance, our point estimates for the average treatment effects also appear rather small. Girls outperform boys on average by 15 percent of a standard deviation in reading, but are worse off by 26 percent in mathematics and 30 percent in science.[17] 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.[18] Back-of-the-envelope calculations suggest that the reform contributed an increase in Germany's average PISA performance of less than 2 points.[19] The average rank of Germany in PISA 2012 would have been the same. Overall, although the average reform effects are statistically significant, the economic significance appears rather small.

Why are the reform effects comparably small? Rivkin & Schiman (2015) suggest that marginal benefits of additional instruction hours may diminish as 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 such that the level of instruction time may not be the most important explanation for small effects in our setting.

Another explanation may be the content of additional instruction time. Whereas in the Israeli case examined in Lavy (2012), the additional classroom time was also

[16] Comparing these findings to our results is somewhat complicated: Both studies proxy general differences in instruction time with the number of instruction hours in the grade 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.

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

[18] 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.

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

intended to cover the current curriculum in more depth, in our setting the additional instruction time covers new learning content. The relevance of this explanation is corroborated by findings from a high school programme in the US. As a consequence of teaching algebra courses from higher grades in earlier grades, 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 were not sufficiently prepared and that maturity effects of when students face certain learning content can play a role. In sum, the content of additional instruction time seems to be an important determinant explaining the small average effects in our policy experiment.

B. Quantile treatment effects

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 estimated G8-coefficients of the RIF-DiD regressions. For mathematics and science, effect sizes are positive, but small and insignificant in the lower deciles. The treatment effects increase as one moves up the performance distribution, becoming statistically significant. If we make the common assumption that the G8-reform preserves students' rank in the performance distribution, we can also interpret the effects on the student performance distribution at the student level: The reform appears more effective for students further up in the performance distribution, especially in mathematics and science. The differences across the performance distribution are less pronounced for reading. One reason for this pattern might be that learning content in science and mathematics builds strongly on previous learning content (Schmidt et al., 2001). Students who did not understand the previous content in mathematics and science will benefit much less from instruction time covering new content. Overall, the results suggest that the distribution of student performance widens because of the reform. This findings is robust to using quantile difference-in-differences (QDiD) instead of RIF-DiD (see Table 5). In this specification, the widening of the performance gap appears to be even larger than in the RIF-DiD specifications and it covers all three PISA domains.

Why do the results differ across the performance distribution? The content of ad-

ditional instruction time seems again important to explain our findings. Students further up in the performance distribution might cope with the additional content more easily, while other students might be overburdened by new learning content. This argument 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 learning content. 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 our results is consistent with skills and instruction hours (with new content) being complements in the educational production process. The pre-existing skill set may be important for processing 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 (see, e.g. Rangvid, 2007; Bellei, 2009; Mueller, 2013; Nicoletti & Rabe, 2014).

C. Further heterogeneities

Next to the effect differences across the performance distribution, the effects may also vary between girls and boys (e.g. Dee, 2007), as well as between students from low- and high-socio-economic backgrounds (e.g. Agasisti & Longobardi, 2014). In Table 6, we report the results for subsamples stratified by gender and parental education.[20]

Across the three domains of reading, mathematics, and science, the effects are very similar for girls and boys. Children from parents without a higher education de-

[20] We also estimated the effects separately for students with and without migration background. However, the share of students with migration background in academic track schools is small, and the students are a highly selective group of migrants. There were no significant effect differences. These results are reported in the discussion paper (Huebener et al., 2016).

gree exhibit slightly larger point estimates in mathematics and science, but slightly smaller estimates in reading. However, the small differences in the treatment effects between subgroups cannot be established with statistical significance. Overall, the findings in our setting suggest that there are no large differences between girls and boys or between children from lower and higher socio-economic status families.[21]

VI Sensitivity checks

A. Threats to the identification strategy

In this section, we discuss several threats to our identification strategy that 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 the reform. The choice for a lower quality school track has lasting consequences as the academic track school is the usual way to earn the general university entrance qualification. Commuting or moving to another federal state involves high costs to both the child and its family, and became increasingly difficult as more federal states implemented the reform. A general escaping behaviour should be evident from enrolment rates in academic track schools. However, Huebener & Marcus (2017) find no evidence of reform-induced lower enrolment rates at academic track schools using administrative data on all students in Germany.[22] We confirm this finding with the PISA data by running difference-in-differences regressions as outlined in equation 1 on observable student characteristics (without individual control variables). In column 1 of Table 7, we consider students across all school tracks in the PISA data, taking an indicator for attending the academic school track as the dependent variable. The probability

[21] The quantile treatment effect estimates for the subsamples are reported in Appendix Table A.4. Overall, a similar picture emerges.

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

of attending the academic track is not affected by the reform. In columns 2 to 6, we directly check for changes in observable characteristics of academic track students taking students' gender, parental education, migration background, grade repetition, and age as dependent variables. All coefficient estimates are close to zero and insignificant, providing no evidence for compositional changes in the student body. The small and insignificant coefficient on grade repetitions (and on students' age) is in line with Huebener & Marcus (2017) who show in administrative data that the G8-reform did not impact grade repetitions until grade 9.

The second main assumption of our identification strategy is the common trend in student performance between treatment and control states in the absence of the G8-reform. The way the reform was implemented across federal states and in one specific school track only enables us to simulate two placebo treatments that can add plausibility to the common trend assumption.[23] 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 (columns 2 and 3 of Table 8), adding plausibility to the common trend assumption.[24]

As a further examination of the plausibility of the common trend assumption, we include linear time trends in our main specification. The results are similar when

[23] All robustness tests of the mean DiD estimates are also conducted for RIF-DiD. The results are reported in Appendix Tables A.5 and A.6. Sensitivity checks for the QDiD regressions are reported in our discussion paper (Huebener et al., 2016).

[24] We can also use the result of the latter placebo regression for a difference-in-differences-in-differences (DiDiD or triple-difference) approach. DiDiD estimates result from the differences between the DiD estimates at academic track schools (column 1) and the placebo DiD estimates at alternative school tracks (column 3). The resulting estimates are 6.08 for reading, 6.1 for mathematics and 4.36 for science. While the placebo effect estimates are small and insignificant, they are estimated with uncertainty. Consequently, DiDiD estimates are less precisely estimated, although they give essentially the same results.

we control for separate time trends for East and West Germany (column 4), and for separate time trends according to the performance in PISA 2000 (column 5).[25]

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 (i.e. the introduction of a two-tier system). It is important to note that our difference-in-differences identification strategy does not need to rely on the absence of other reforms, but it requires that these reforms do not coincide with the introduction of the G8-reform. In columns 6 to 8 of Table 8, we add dummy variables to equation 1 for each of the reforms reported in Table 1. The robustness of our estimates suggests that the G8-indicator (varying across states and time) is sufficiently orthogonal to each of the other reforms.

Another concern may be a federal school investment programme that aimed at promoting the introduction of all-day schooling. All-day schools in Germany are usually attended voluntarily and incorporate leisure time activities, homework supervision and study time in the regular school day. The programme was passed in 2003, it addressed all school types (primary schools as well as all secondary school tracks) in all federal states, and it was rolled out slowly. 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 voluntary offers. We perform three tests to check whether the expansion of all-day schooling might confound our results. First, we show that there is no evidence that the G8-reform had an impact on PISA scores in other school tracks (see column 3 of Table 8). 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 turns significant in alternative school tracks. Second, we take the share of all-day students (KMK, 2016) in academic track schools in the federal state at the time of the PISA assessment as a dependent

[25]This specification includes four time trend variables based on the state's quartile in PISA 2000 (academic track students only), the first PISA assessment.

variable in our difference-in-differences model and estimate how the G8-reform impacts the share of all-day students in the federal state. The estimated coefficient is small (0.01) and insignificant. Third, we control for this share of all-day students in our main specification, but this does not impact our findings (see column 9 of Table 8).

B. Specification issues

In this section, we show that our results are not sensitive to the choice of control variables and to accounting for exceptional cohorts. In column 2 of Table 9, we estimate the main model without the set of student characteristics, X_{ist}. In column 3, we add a set of school characteristics (teacher-student-ratio, student-computer-ratio, public school dummy) to the main model.[26] As certain individual control variables are missing for approximately 6 percent of the sample, in column 4 we include these observations in our sample and re-estimate the main model accounting for missing socio-economic control variables with dummy variables. Our findings are robust to these alternative specifications.

In the next set of robustness checks, we examine how sensitive our findings are to controlling for exceptional circumstances of certain cohorts. In column 5 of Table 9, we include a dummy variable controlling for students of the first G8-cohort and the last G9-cohort that are part of the so-called double graduation cohort. Due to the nature of the G8-reform, these cohorts graduate at the same time, which may create exceptional performance incentives. In column 6, we control for cohorts that were tested in PISA in the year when the double graduation cohort was in their last year, i.e. the year when the school's overall teaching load was highest. Finally, in column 7 we control for students in Saxony-Anhalt and Mecklenburg-Vorpommern tested in PISA 2006. These students were "surprised" by the reform as they were already in higher grades when the reform was introduced (see Section III). Our results are not sensitive to accounting for these exceptional cohorts.

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

C. Other channels

In the following, we examine whether the G8-reform might affect student performance through other channels besides the increase in weekly instruction hours.

Given that students have a restricted time budget set, the reform could affect the time they spend 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 learning content in private remediation classes. In 2003 and 2012, the student questionnaires contain questions on homework, out-of-school classes and tutoring. We use this information to estimate difference-in-differences models as in equation 1 (this time based on only two time periods). Table 10 reports the results. The average number of hours per week spent on homework did not change significantly with the reform (column 1). The estimated G8-effect on out-of-school classes and private tutoring suggests a reduction of 5 percentage points that is borderline significant, indicating a small substitution of out-of-school classes with classroom time in school. However, note that this comparison is only based on the 2003 and 2012 PISA waves and we cannot check for the common trend in the pre-treatment period. But in which activities do students reduce time if they spend more time in school? Meyer & Thomsen (2015) and Hübner et al. (2017) investigate this question for single federal states at the end of academic track schooling, when students are about three years older than students in our sample. Meyer & Thomsen (2015) also cannot find effects on homework. Further, there are no effects on sports and music activities. Their results rather suggest that students spend less time reading, watching TV, surfing the internet, listening to music, or doing nothing. Likewise, Hübner et al. (2017) find small reductions in leisure time spent on meeting friends and watching TV.

Related to the time use of students is the question of whether students actually 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 questionnaires ask students how often they missed school, skipped classes, or arrived late for school during the previous two weeks. We again use difference-in-differences models to estimate reform effects on these outcomes. Columns 3-5 of Table 10 show that the propensity of students to miss or skip classes, or to arrive late for school did not change significantly with the G8-reform. There is no evidence that increases in actual instruction time lag behind increases in allocated instruction time.[27]

Next, we examine whether the G8-reform affected the composition of the teacher body at academic track schools, and thereby potentially the returns to instruction time. If policymakers want to increase instruction hours, the demand for teaching hours increases. In our setting, the potential impact of changes in the teacher body is likely to be very 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) because schooling was shortened by one year as well. The updated G8-timetable regulations applied to cohorts newly entering academic track schools. Schools and teachers could gradually adapt to the increased demand for teaching during the transition period. Furthermore, academic track teachers need to satisfy high qualification standards. They are trained in at least two subjects and teach in parallel in several grade levels (about 50 percent of teachers teach in five or more grade levels, according to the PISA 2006 teacher questionnaire), allowing them in general to flexibly react to changing timetable requirements.

To empirically analyse G8-related changes in the teacher body, we gathered information from the German Microcensus for the years 2001-2012 (FDZ, 2016). The Microcensus randomly samples one percent of households in Germany each year. We rely on the scientific use file, which covers 70 percent of the original sample.

[27]Besides 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. However, there is no evidence that the G8-reform impacts the term length (see Appendix Table A.7).

The contained information on individuals' profession allow identifying teachers at academic track schools. Based on this sample of about 14,500 academic track school teachers, we investigate changes in the teaching staff related to the G8-reform. As schools encounter a stepwise increase in instruction hours with every new G8-cohort entering academic track school during the transition period, we substitute the G8-dummy in our DiD model with a measure for this additional teaching load. This measure equals the number of grades already affected by the G8-reform during the 8-year transition period (i.e. it ranges from 1 to 8), and zero otherwise.

To develop a general understanding about changes in the teacher body with the introduction of G8, we estimate the G8-effect on the probabilities of being female, of being of foreign nationality, of cohabiting, of working on a fixed-term contract, and of holding a university degree. The large sample size allows estimating the reform effects very precisely, but we cannot find evidence for any change in these teacher characteristics (see Panel A of Table 11). However, we find evidence that teachers extend their working hours slightly in response to each new G8-cohort entering academic track schools by, on average, 0.6 percent (0.23 hours per week). We also find that teachers are slightly younger (-0.23 years). The share of teachers below the age of 30 increases by 0.23 percentage points (significant at the 10 percent level), while the share of teachers above the age of 60 increases by 0.18 percentage points (insignificant). As a placebo test, we repeat the analysis for teachers in other school tracks (Panel B of Table 11), where all coefficients on socio-demographic characteristics, working hours and age are close to zero and statistically insignificant. This provides additional support for our identification strategy for the teacher effects. School principals seem to react to the increased demand for instruction hours with working hours extensions of teachers, keeping older teachers a little longer (though this is not statistically significant), and hiring new teachers.[28] Overall, the changes in the academic track teacher body are small (particularly when compared to the overall means), occur gradually, and also affect G9 students during the transition phase.

[28]The adjustments in the teaching staff body can explain why there is no G8-reform effect on the student-teacher ratio in the PISA data (reported in our discussion paper Huebener et al., 2016).

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. At the end of the transition phase, when the double cohort is in the final year, the additional workload of teachers is largest (and presumably potential changes in teachers' motivation). Excluding cohorts from the sample that were tested in PISA at this time hardly changes the point estimates (see column 6 of Table 9), suggesting that changes in teacher motivation and effort do not play a major role in explaining the G8-reform effects on student performance.

Finally, classroom quality can be an important determinant of the returns to instruction time (Rivkin & Schiman, 2015). Note that the G8-reform was implemented within a given school infrastructure and school environment, and we cannot find evidence for G8-effects on the composition of the student body. Together with the small effects on the teacher body, the classroom quality is also unlikely to have changed substantially with the G8-reform.

In sum, the assembled arguments suggest that the main G8-effect is induced by increased instruction hours.[29] In the following section we discuss to what extend the effects may be specific to the German context and the fact that G8-students can graduate from school in one year less.

[29]One may want to use the G8-reform as an instrumental variable (IV) in the identification of the causal effects of instruction time. IV estimates are re-scaled reduced-form estimates and depend on the choice of the first stage dependent variable. It is debatable what the appropriate first stage is in our case and whether the required exclusion restriction is satisfied. One could use the total cumulative instruction time between grade 5 and the time of the PISA test, assuming that an additional instruction hour in grade 5 has the same effect on student performance as an additional hour in grade 9. It is not clear whether instruction time in higher grades is more relevant for PISA scores (due to being more recent at the time of the test), or instruction time in lower grades (due to dynamic complementarities, see, e.g. Cunha & Heckman, 2007). If we still draw the first stage from the cumulative instruction time from Table 2 (1.99 hours increase), the resulting IV-estimates are 2.9 for reading, 2.6 for mathematics and 2.9 for science. Further note that IV-estimations require that the G8-reform affected student performance exclusively through increased school instruction time. While we argue that channels other than school instruction time play only a minor role, we cannot entirely rule out that the reform operates through other channels. Therefore, the IV-approach is not our preferred choice.

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, policymakers 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), the German experience is informative for policymakers in other countries considering increases in classroom time and needing to decide how the additional time is spent.

However, for other school systems our small estimated treatment effects may even be too optimistic for three reasons: First, 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, classroom heterogeneity in student ability is larger in systems without tracking, thus the variation of effects across the student performance distribution may even be wider if additional time is spent on new content. Second, the G8-reform affected the high-ability school track, in which the quality of teachers and the peer environment is considered high. Rivkin & Schiman (2015) suggest that more instruction time is more beneficial in more favourable classroom environments. Third, the gradual introduction and the nature of the G8-reform make it less likely that the estimated effects are importantly impacted by changes in the teacher body. Other school systems increasing instruction time would, *ceteris paribus*, have to increase the number of teachers (or their working hours) to provide extra time as well. The increase in working hours in the G8-context may partly contain (positive) self-selection of teachers for longer working hours. Consequently, our small treatment effects could be even smaller in other education systems.

Another concern for the external validity of our results would be behavioural changes of students that are not related to increased instruction time, but mainly to the shorter time to graduation also implied by the G8-reform. Our identification strategy

would capture both of these effects. An important question in this context is whether students react to their earlier graduation with a changed study behaviour in grade 9, i.e. more than three years before graduation. For their post-secondary educational career, the final grade point average (GPA) matters. However, only grades earned in the final two years and in the final exams count towards the final GPA. Exerting more effort already in grade 9 would require a great amount of discipline, forward-looking behaviour, and the awareness of students about dynamic complementarities of their acquired skills. However, there is evidence that school children are far less forward-looking and patient than adults (see, e.g. Bettinger & Slonim, 2007; Lavecchia et al., 2016). If G8-students still exerted exceptional effort because of the shortened school duration, we would overestimate the effects of additional instruction time and our conclusion of small overall effects would remain.

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 contribute 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 ninth graders PISA scores in reading, mathematics, and science. 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 student performance distribution widens, especially in mathematics and science. These findings 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 has important implications for policymakers. They often associate increased instruction time as a means to improve student performance and to narrow performance gaps between low- and high-performing students. We demonstrate that student performance can indeed be improved. However, the magnitude of effects seems small, while increases in instruction time may also widen the gap between low- and high-performing students. The content of additional classroom time should be considered carefully.

Acknowledgements

This paper benefited from comments and suggestions by two anonymous referees, the guest editor Edwin Leuven, 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*.

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

Athey, S. & Imbens, G. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431–497.

Aucejo, E. M. & Romano, T. F. (2016). Assessing the effect of school days and absences on test score performance. *Economics of Education Review*, *55*, 70–87.

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. (2016). Should we increase instruction time in low achieving schools? Evidence from Southern Italy. *Economics of Education Review*, *55*, 39–56.

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.

Bettinger, E. & Slonim, R. (2007). Patience among children. *Journal of Public Economics*, *91*(1-2), 343–363.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The 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. *The 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.

Cunha, F. & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

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.

FDZ (2016). Mikrozensus, Erhebungsjahre 2001-2012. *FDZ der Statistischen Ämter des Bundes und der Länder*.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, *77*(3), 953–973.

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.

Havnes, T. & Mogstad, M. (2015). Is universal child care leveling the playing field? *Journal of Public Economics*, *127*, 100–114.

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

Hübner, N., Wagner, W., Kramer, J., Nagengast, B., & Trautwein, U. (2017). Die G8-Reform in Baden-Württemberg: Kompetenzen, Wohlbefinden und Freizeitverhalten vor und nach der Reform. *Zeitschrift für Erziehungswissenschaft*, forthcoming.

Huebener, M., Kuger, S., & Marcus, J. (2016). Increased instruction hours and the widening gap in student performance. *DIW Discussion Paper*, *1561*.

Huebener, M. & Marcus, J. (2015). Empirische Befunde zu Auswirkungen der G8-Schulzeitverkürzung. *DIW Roundup Politik im Fokus*, *57*.

Huebener, M. & Marcus, J. (2017). Compressing instruction time into fewer years of schooling and the impact on student performance. *Economics of Education Review*, *58*(June), 1–14.

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. *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*.

Lavecchia, A., Liu, H., & Oreopoulos, P. (2016). Behavioral economics of education. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 5). Chapter 1, (pp. 1–74). Amsterdam.

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

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

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

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.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Schulferien.org (2016). Kalender mit Schulferien und Feiertagen. *Retrieved from http://www.schulferien.org/Kalender_mit_Ferien/*, downloaded in January, 2016.

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.

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

Figures

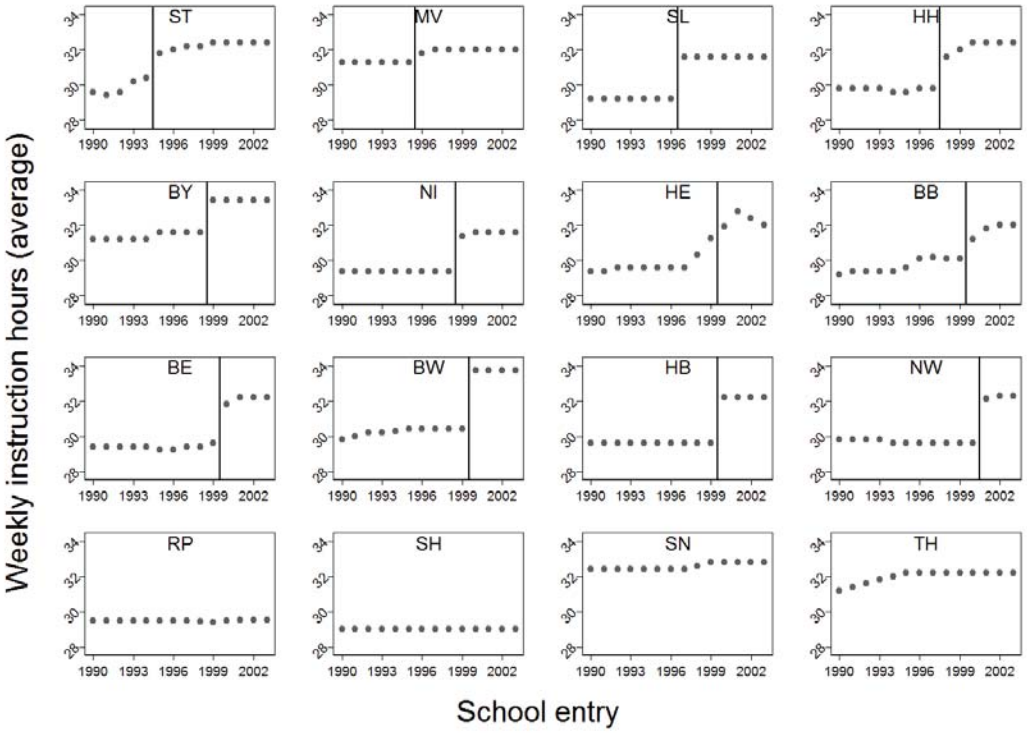

Figure 1: Number of weekly instruction hours by school entry cohort (averaged over grades 5 to 9). In the order of reform introduction: ST: Saxony-Anhalt, MV: Mecklenburg-Vorpommern, SL: Saarland, HH: Hamburg, BY: Bavaria, NI: Lower-Saxony, HE: Hesse, BB: Brandenburg, BE: Berlin, BW: Baden-Württemberg, 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.

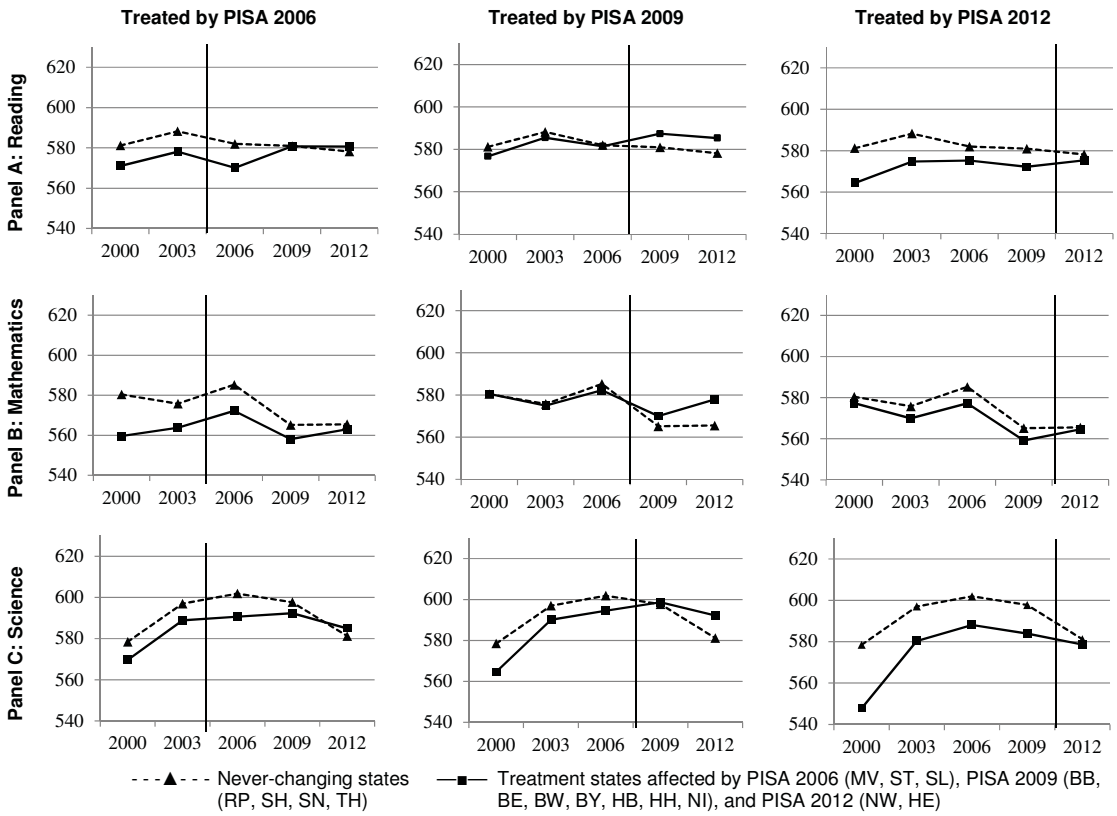

Figure 2: Development of PISA scores in the control group of states that did not change their treatment status in the period of analysis, and in treated states that implemented the reform before PISA 2006 (first column), between PISA 2006 and PISA 2009 (second column), and between PISA 2009 and PISA 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. *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	1.99***	1.94***	1.62***	1.66**	2.09***	2.65***
weekly instruction hours	(0.44)	(0.46)	(0.41)	(0.69)	(0.54)	(0.46)
$\% - change$	6.53	6.79	5.44	5.32	6.66	8.37
		By subject				
		All	Language arts	Mathematics	Biology, physics, chemistry	Others (e.g. history, geography, foreign languages)
Average change in		1.99***	0.02	0.10*	0.62***	1.25**
weekly instruction hours		(0.44)	(0.06)	(0.06)	(0.16)	(0.52)
$\% - change$		6.53	0.51	2.48	18.41	6.61
N				33,217		

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	1,322	
Number of students	33,217	

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 treatment effects								
		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***	5.42*	6.00**	5.69**	6.52**	6.83**	6.99**	7.23***	7.89***	6.91**
	(1.91)	(2.82)	(2.42)	(2.76)	(2.79)	(2.67)	(2.98)	(2.80)	(2.93)	(2.74)
Mathematics										
G8-reform	5.26**	1.57	2.41	2.75	4.96*	6.88**	6.08**	7.26**	6.95**	7.15**
	(2.55)	(3.46)	(2.86)	(3.23)	(3.00)	(3.23)	(2.99)	(3.06)	(3.02)	(3.46)
Science										
G8-reform	5.71*	0.17	1.59	4.12	5.64	6.48*	8.80**	8.84***	8.53***	8.36***
	(2.99)	(4.33)	(4.38)	(4.61)	(3.73)	(3.52)	(3.79)	(3.39)	(3.31)	(2.67)
N	33,217	33,217	33,217	33,217	33,217	33217	33,217	33,217	33,217	33,217

Notes: The table reports OLS and RIF-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 RIF-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: Main results: QDiD estimates on student performance distribution

Dependent variable: Domain specific PISA score									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Quantile treatment effects								
	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	2.92	4.59**	4.15*	5.77***	6.01***	6.65**	7.56**	8.30***	7.93***
	(2.25)	(1.79)	(2.42)	(2.13)	(2.15)	(2.94)	(3.03)	(3.19)	(2.98)
Mathematics									
G8-reform	1.95	0.62	3.18	4.96	5.56*	6.72**	7.87**	8.49***	8.34**
	(3.32)	(2.95)	(2.61)	(3.19)	(2.87)	(3.00)	(3.06)	(2.87)	(3.32)
Science									
G8-reform	1.95	3.24	4.28	5.20	6.63*	7.31**	7.79**	7.85**	7.58**
	(3.59)	(3.42)	(3.77)	(3.48)	(3.72)	(3.55)	(3.54)	(3.45)	(3.05)
N	33,217	33,217	33,217	33,217	33,217	33,217	33,217	33,217	33,217

Notes: The table reports estimates of the G8-reform effect from quantile difference-in-differences (QDiD) models. 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). Clustered standard errors (federal state level) for QDiD-regressions are reported in parentheses and are bootstrapped (200 replications). 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: 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	17,990	15,227	12,301	20,916

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 7: OLS estimates of the G8-reform effect on student composition

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable:					
	At academ. track	Girls	Parents with ISCED≥5	Migrants	Grade repeated	Age in years
G8-reform	-0.0100	-0.0021	-0.0143	-0.0100	0.0046	0.0153
	(0.0330)	(0.0186)	(0.0185)	(0.0214)	(0.0124)	(0.0299)
Sample mean	0.34	0.54	0.64	0.13	0.07	15.38
N	100,972	33,217	33,217	33,217	32,990	33,217

Notes: The table reports OLS regression estimates of the G8-reform effect on student characteristics. 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. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

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

Table 8: Threats to identification: OLS results for average treatment effects

Dependent variable: Domain specific PISA score									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Placebo treatments		Linear time trends		Controlling for other reforms			
	Main	Treatment one period earlier	Treatment in other school tracks	East-West trends	PISA 2000 performance (4 groups)	Central exit exams	Tracking after grade 6	Reduced no. of tracks	Expansion of all-day schooling
Reading									
G8-reform	5.76***	-0.31	-0.32	6.31***	4.63**	6.03***	5.81***	5.09**	5.79***
	(1.91)	(2.45)	(2.56)	(2.03)	(1.94)	(2.32)	(1.98)	(2.04)	(1.98)
Mathematics									
G8-reform	5.26**	-1.56	-0.85	4.58*	4.81*	4.88**	4.10*	5.18*	5.53**
	(2.55)	(2.56)	(2.79)	(2.59)	(2.60)	(2.43)	(2.37)	(2.77)	(2.50)
Science									
G8-reform	5.71*	-1.00	1.36	4.82*	5.78**	5.35**	5.05*	6.04**	6.06**
	(2.99)	(3.40)	(3.35)	(2.70)	(2.84)	(2.72)	(2.99)	(2.92)	(2.55)
N	33,217	33,217	67,755	33,217	33,217	33,217	33,217	33,217	33,217

Notes: The table reports OLS regression estimates for the effect of the G8-reform, relating to various threats to the identification strategy. 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). Column 9 allows for a linear trend of four state-groups based on their PISA 2000 performance (BY, SH, NI, RP; SN, BW, NW, TH; HE, SL, BE, MV; HH, ST, HB, BB). 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 9: Sensitivity checks: OLS estimates for alternative model specifications

Dependent variable: Domain specific PISA score							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
					Controlling for exceptional cohorts		
		Control variables			Double cohort		
	Main	No controls	Individual & school level	Full sample[a]	In grade 9	In final year	Surprised cohorts
Reading							
G8-reform	5.76***	5.73***	6.03***	5.86***	5.93***	5.65***	5.67**
	(1.91)	(2.00)	(2.04)	(1.90)	(1.66)	(1.84)	(2.35)
Mathematics							
G8-reform	5.26**	5.19*	5.15**	5.31**	5.95**	5.27**	6.73**
	(2.55)	(2.98)	(2.60)	(2.49)	(2.46)	(2.60)	(2.66)
Science							
G8-reform	5.71*	5.75*	5.72*	5.63**	6.07**	5.63*	6.74**
	(2.99)	(3.10)	(3.01)	(2.83)	(2.76)	(2.89)	(3.17)
N	33,217	33,217	33,217	35,429	33,217	33,217	33,217

Notes: The table reports OLS regression estimates of the G8-reform effect for varying model specifications. 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) unless stated differently. "Surprised cohorts" refers to students captured in PISA 2006 in ST and MV, as they were already in grade 7 (ST) and 8 (MV) when the reform was implemented. 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 35,429, for mathematics 34,619 and for science 34,280.

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

Table 10: OLS estimates of the G8-reform effect on other channels

	(1)	(2)	(3)	(4)	(5)
	Dependent variable:				
	Homework	Tutoring	Missing school	Skipping classes	Arriving late
G8-reform	0.01	-0.05*	-0.02	-0.01	0.02
	(0.60)	(0.03)	(0.03)	(0.02)	(0.04)
Sample mean	6.40	0.33	0.13	0.09	0.23
N	9,426	8,190	11,493	11,479	11,492

Notes: The table reports OLS regression estimates of the G8-reform effect on homework (in hours per week) as well as attending private tutoring/out-of-school classes, missing school, skipping classes and arriving late for school in the previous two weeks prior to PISA (all binary). 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. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.
Source: PISA 2000, 2003, 2012 for Germany.

Table 11: G8-reform effects on teacher characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dependent variable:								
	Female	Foreign nationality	Cohabiting	Fixed-term contract	University degree	Working hours	Age	Age ≤ 30	Age ≥ 60
Panel A: Teachers in academic track schools									
G8-cohorts	-0.0004	0.0005	-0.0024	-0.0005	-0.0004	0.2301***	-0.2269**	0.0023*	0.0018
	(0.0023)	(0.0007)	(0.0030)	(0.0011)	(0.0016)	(0.0763)	(0.0885)	(0.0013)	(0.0016)
Mean	0.5400	0.0197	0.7091	0.0997	0.9449	37.9972	45.9690	0.1103	0.1042
%-change	-0.07	2.53	-0.35	-0.46	-0.04	0.61	-0.49	2.13	1.68
N	14,479	14,479	10,319	14,435	14,118	14,479	14,479	14,479	14,479
Panel B: Teachers in other school tracks (primary school, lower- and middle-secondary school)									
G8-cohorts	0.0028	-0.0000	-0.0003	-0.0007	0.0011	-0.0683	-0.1076	-0.0001	0.0004
	(0.0021)	(0.0004)	(0.0013)	(0.0011)	(0.0025)	(0.0507)	(0.0694)	(0.0015)	(0.0013)
Mean	0.7581	0.0172	0.7283	0.0706	0.7552	34.3623	46.0141	0.1058	0.0951
%-change	0.36	-0.13	-0.04	-0.97	0.15	-0.20	-0.23	-0.11	0.39
N	30,180	30,180	21,247	30,030	29,432	30,180	30,180	30,180	30,180

Notes: The table reports OLS regression estimates for the effect of the number of G8 cohorts during the transition period on teacher characteristics. All estimates are obtained from separate regressions for the sample indicated by the panel name and include federal state and year fixed effects. "Cohabiting" is only observed for the years 2005-2012. 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: German Microcensus 2001-2012.

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: Wild cluster bootstrap

	Coefficient (1)	p-value Clustering at state level (2)	p-value Wild cluster bootstrapping (3)
Reading			
G8-reform	5.76***	[0.003]	[0.000]
Mathematics			
G8-reform	5.26**	[0.045]	[0.016]
Science			
G8-reform	5.71*	[0.052]	[0.026]

Notes: The table presents OLS estimates for the average reform effect (column 1). Column 2 shows p-values based on conventional clustering at the state level and column (3) based on a wild cluster bootstrap procedure (999 replication, Mammen weights, testing under H_0). * $p<0.1$, ** $p<0.05$, *** $p<0.01$.
Source: PISA 2000, 2003, 2006, 2009, 2012 for Germany.

Table A.3: OLS estimates of the effect of subject specific instruction hours on student performance

	(1)	(2)	(3)
	Dependent variable: Domain specific PISA score in		
	Reading	Mathematics	Science
Language arts	3.73*	-1.38	-1.30
	(1.97)	(1.99)	(1.78)
Mathematics	1.03	4.69**	4.23**
	(1.56)	(2.09)	(1.78)
Biology, physics and chemistry	0.33	-0.47	0.02
	(0.43)	(0.70)	(0.65)
Other subjects	0.54**	0.46*	0.38
	(0.24)	(0.24)	(0.29)
N	33,217	33,217	33,217

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.4: Heterogeneity analysis: Subsample estimates of the G8-reform effect on the distribution of 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=17,990]									
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=15,227]									
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=12,301]									
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=20,916]									
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 RIF-DiD estimates of the G8-reform effect on student performance for various subsamples. 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.5: Threats to validity: RIF-DiD results for quantile treatment effects

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=33,217]									
Reading	-4.70	-1.93	-0.51	0.15	0.54	1.14	1.79	2.12	2.99
	(3.21)	(2.91)	(2.85)	(2.62)	(2.65)	(2.73)	(2.56)	(2.74)	(3.27)
Mathematics	1.70	-0.83	-1.47	-1.69	-1.80	-1.94	-2.81	-4.40	-2.70
	(4.09)	(3.29)	(2.75)	(2.58)	(2.98)	(3.01)	(3.01)	(2.89)	(4.39)
Science	-0.71	0.22	0.44	-0.81	-1.49	-2.07	-1.35	-2.44	-2.05
	(3.66)	(2.68)	(3.41)	(2.78)	(2.83)	(2.80)	(3.24)	(3.10)	(3.95)
Placebo treatment: Treatment in other school tracks [N=67,755]									
Reading	-3.57	1.45	3.54	3.50	2.61	3.01	0.74	-0.53	-2.01
	(4.44)	(3.12)	(3.01)	(2.90)	(2.85)	(2.53)	(3.05)	(2.45)	(3.60)
Mathematics	-1.63	-2.86	-1.02	-0.72	0.63	0.62	0.51	0.13	-1.78
	(3.44)	(3.01)	(2.82)	(2.68)	(2.43)	(2.67)	(2.48)	(2.75)	(2.95)
Science	-1.06	0.39	1.08	1.50	2.26	3.18	3.15	3.24	1.35
	(3.71)	(3.86)	(3.09)	(2.87)	(2.57)	(2.68)	(2.91)	(3.76)	(3.25)
Linear time trends: East-West trends [N=33,217]									
Reading	5.92	6.56*	6.30**	7.14***	7.39**	7.56***	7.78***	8.53***	7.52*
	(4.03)	(3.94)	(2.99)	(2.62)	(2.91)	(2.49)	(2.60)	(2.41)	(3.88)
Mathematics	0.92	1.81	2.06	4.15	6.13**	5.33**	6.44**	6.25**	6.40
	(4.44)	(3.19)	(2.44)	(2.69)	(2.43)	(2.47)	(2.58)	(2.97)	(4.21)
Science	-1.21	0.28	2.87	4.48	5.45**	8.03***	8.12**	8.13**	8.20*
	(4.39)	(3.88)	(2.85)	(2.84)	(2.45)	(2.53)	(3.28)	(3.27)	(4.53)
Linear time trends: PISA 2000 performance (4 groups) [N=33,217]									
Reading	4.19	4.96	4.33	5.26**	5.65**	5.77**	6.20**	6.89***	5.97
	(3.99)	(4.12)	(3.10)	(2.62)	(2.87)	(2.44)	(2.64)	(2.42)	(3.83)
Mathematics	0.37	1.34	2.19	4.52*	6.33**	5.75**	7.15***	7.14**	7.71*
	(4.68)	(3.36)	(2.45)	(2.71)	(2.47)	(2.49)	(2.61)	(3.07)	(4.14)
Science	-0.08	1.57	4.04	5.67**	6.40**	8.59***	8.86***	8.81**	9.17**
	(4.35)	(3.90)	(2.92)	(2.82)	(2.50)	(2.63)	(3.28)	(3.42)	(4.58)
Other reforms: Central exit exams [N=33,217]									
Reading	5.63	6.21	5.92**	6.79***	7.09**	7.29***	7.57***	8.26***	7.25*
	(3.95)	(3.91)	(2.98)	(2.60)	(2.96)	(2.49)	(2.60)	(2.38)	(3.89)
Mathematics	1.17	2.05	2.34	4.53*	6.51***	5.70**	6.85***	6.58**	6.83
	(4.52)	(3.20)	(2.41)	(2.68)	(2.36)	(2.48)	(2.54)	(2.94)	(4.26)
Science	-0.51	0.94	3.63	5.20*	6.12**	8.53***	8.63***	8.39**	8.40*
	(4.44)	(3.87)	(2.86)	(2.77)	(2.44)	(2.51)	(3.23)	(3.31)	(4.44)

Table A.5 continued on the next page

Table A.5 – 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: Tracking after grade 6 [N=33,217]									
Reading	5.64	6.47*	5.91**	6.55**	6.73**	6.88***	7.17***	7.97***	6.84*
	(4.09)	(3.83)	(2.90)	(2.59)	(2.87)	(2.48)	(2.66)	(2.40)	(3.79)
Mathematics	0.21	1.05	1.52	3.69	5.49**	4.94**	6.19**	5.95*	6.09
	(4.44)	(3.23)	(2.42)	(2.70)	(2.39)	(2.44)	(2.55)	(3.06)	(4.13)
Science	-0.31	0.86	3.13	4.69	5.58**	7.92***	7.99**	7.90**	8.08*
	(4.48)	(3.98)	(2.80)	(2.86)	(2.46)	(2.51)	(3.23)	(3.40)	(4.48)
Other reforms: Reduced no. of tracks [N=33217]									
Reading	5.49	5.80	5.25*	5.99**	6.16**	6.10**	6.16**	6.71***	5.68
	(3.91)	(4.14)	(3.12)	(2.64)	(2.90)	(2.53)	(2.67)	(2.44)	(3.71)
Mathematics	2.23	2.62	2.97	5.00*	6.68***	5.80**	6.80**	6.52**	6.88
	(4.64)	(3.31)	(2.50)	(2.73)	(2.43)	(2.52)	(2.70)	(3.05)	(4.28)
Science	0.89	2.52	4.75	6.08**	6.79***	8.88***	8.97***	8.55**	8.39*
	(4.29)	(3.93)	(3.03)	(2.83)	(2.52)	(2.68)	(3.32)	(3.34)	(4.59)
Other reforms: Expansion in all-day schooling programmes [N=33,217]									
Reading	5.51	6.08	5.69*	6.54**	6.83**	6.99***	7.23***	7.89***	6.95*
	(4.04)	(3.93)	(2.99)	(2.61)	(2.97)	(2.48)	(2.63)	(2.38)	(3.88)
Mathematics	1.90	2.70	3.07	5.28*	7.14***	6.37**	7.56***	7.24**	7.41*
	(4.44)	(3.18)	(2.43)	(2.73)	(2.40)	(2.49)	(2.57)	(2.97)	(4.23)
Science	0.94	2.14	4.57	6.01**	6.75***	9.01***	9.03***	8.65**	8.40*
	(4.40)	(3.83)	(2.85)	(2.79)	(2.47)	(2.55)	(3.27)	(3.37)	(4.50)

Notes: The table reports the sensitivity checks described in Section VI for the quantile treatment effects. All estimates are obtained from separate RIF-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.6: Sensitivity checks: RIF-DiD estimates 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
Control variables: No controls [N=33,217]									
Reading	5.53	5.98	5.65*	6.46**	6.74**	6.90***	7.14***	7.81***	6.80*
	(4.00)	(3.96)	(3.02)	(2.67)	(2.99)	(2.53)	(2.63)	(2.40)	(3.85)
Mathematics	1.65	2.42	2.72	4.89*	6.76***	5.97**	7.10***	6.78**	6.97*
	(4.50)	(3.26)	(2.49)	(2.74)	(2.45)	(2.54)	(2.62)	(3.03)	(4.23)
Science	0.43	1.71	4.19	5.68**	6.46***	8.76***	8.77***	8.45**	8.25*
	(4.44)	(3.87)	(2.89)	(2.81)	(2.51)	(2.59)	(3.30)	(3.35)	(4.52)
Control variables: Individual & school level controls [N=33,217]									
Reading	5.93	6.49	6.13**	6.90**	7.06**	7.21***	7.42***	8.03***	7.32*
	(4.02)	(3.99)	(3.08)	(2.69)	(2.86)	(2.39)	(2.55)	(2.44)	(3.87)
Mathematics	1.30	2.03	2.33	4.84*	6.81***	6.15**	7.26***	6.97**	7.24*
	(4.46)	(3.33)	(2.43)	(2.77)	(2.46)	(2.58)	(2.59)	(2.98)	(4.18)
Science	-0.01	1.39	4.03	5.60**	6.53***	8.98***	9.12***	8.81***	8.57*
	(4.44)	(3.99)	(2.93)	(2.74)	(2.49)	(2.56)	(3.26)	(3.36)	(4.47)
Control variables: Full sample [N=35,429]									
Reading	5.06	5.99	5.82*	6.76***	7.23***	7.33***	7.71***	8.05***	7.33**
	(3.96)	(3.92)	(3.17)	(2.49)	(2.71)	(2.26)	(2.50)	(2.32)	(3.71)
Mathematics	2.72	2.66	3.05	4.80*	6.46***	5.91**	7.02***	6.63**	6.82
	(4.69)	(3.58)	(2.41)	(2.70)	(2.39)	(2.32)	(2.37)	(2.96)	(4.34)
Science	-0.74	1.49	4.08	5.72**	6.52***	8.86***	8.90***	8.74***	8.18**
	(4.09)	(4.12)	(2.90)	(2.70)	(2.43)	(2.43)	(3.14)	(3.18)	(4.17)
Exceptional cohorts: Double cohort in grade 9 [N=33,217]									
Reading	5.00	5.96	5.85**	6.76**	7.21**	7.55***	7.91***	8.60***	7.41*
	(4.00)	(3.87)	(2.85)	(2.64)	(2.85)	(2.43)	(2.55)	(2.47)	(3.82)
Mathematics	2.01	2.93	3.32	5.46**	7.47***	6.92***	8.48***	7.92***	8.18*
	(4.46)	(3.37)	(2.44)	(2.76)	(2.46)	(2.55)	(2.55)	(2.87)	(4.30)
Science	0.34	1.96	4.60	6.22**	6.99***	9.21***	9.25***	8.99***	8.74*
	(4.28)	(3.86)	(2.87)	(2.91)	(2.47)	(2.55)	(3.25)	(3.41)	(4.75)
Exceptional cohorts: Double cohort in final year [N=33,217]									
Reading	5.38	5.91	5.60*	6.40**	6.70**	6.86***	7.09***	7.75***	6.77*
	(3.97)	(3.94)	(2.97)	(2.62)	(2.95)	(2.48)	(2.59)	(2.38)	(3.83)
Mathematics	1.61	2.44	2.74	4.96*	6.86***	6.08**	7.28***	6.92**	7.12*
	(4.47)	(3.20)	(2.42)	(2.69)	(2.39)	(2.48)	(2.57)	(2.94)	(4.23)
Science	0.17	1.52	4.04	5.56**	6.41***	8.69***	8.74***	8.43**	8.21*
	(4.41)	(3.89)	(2.88)	(2.80)	(2.46)	(2.55)	(3.26)	(3.33)	(4.47)

Table A.6 continued on the next page

Table A.6 – 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
Exceptional cohorts: Surprised cohorts [N=33,217]									
Reading	5.91	5.62	5.10	6.08**	6.31*	6.70**	6.98**	7.58***	6.50
	(4.48)	(4.23)	(3.14)	(2.88)	(3.34)	(2.74)	(2.80)	(2.66)	(4.13)
Mathematics	3.52	3.92	4.03	6.47**	8.50***	7.44***	8.45***	8.08***	8.03*
	(4.45)	(3.48)	(2.72)	(3.03)	(2.58)	(2.80)	(2.85)	(3.11)	(4.59)
Science	1.63	2.25	5.09	6.62**	7.21***	9.92***	10.11***	9.56***	9.08*
	(4.77)	(4.29)	(3.21)	(2.96)	(2.74)	(2.81)	(3.50)	(3.49)	(4.70)
Alternative estimation method: Quantile difference-in-differences (QDiD) [N=33,217]									
Reading	2.92	4.59	4.15	5.77**	6.01***	6.65***	7.56***	8.30***	7.93**
	(3.41)	(2.80)	(2.72)	(2.69)	(2.23)	(2.49)	(2.18)	(2.79)	(3.52)
Mathematics	1.95	0.62	3.18	4.96**	5.56**	6.72***	7.87**	8.49***	8.34**
	(3.73)	(3.31)	(2.84)	(2.36)	(2.32)	(2.44)	(3.17)	(2.84)	(4.07)
Science	1.95	3.24	4.28*	5.20*	6.63***	7.31***	7.79***	7.85**	7.58*
	(4.12)	(4.00)	(2.59)	(2.91)	(2.50)	(2.63)	(2.65)	(3.05)	(3.95)

Notes: The table reports the sensitivity checks described in Section VI for the quantile treatment effects. All estimates are obtained from separate RIF-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.7: G8-reform effect on instruction hours and holidays

	(1)	(2)	(3)
	Dependent variable aggregated from grade 5-9		
	School holidays	Bank holidays	Total holidays
G8-reform	0.93	-2.00	-1.07
	(1.17)	(1.24)	(0.74)
Sample mean	326.04	35.94	361.98
N	33,217	33,217	33,217

Notes: The table reports the estimated G8-reform effect on the students' number of school holidays, bank holidays and total holidays they were exposed to between grades 5 through 9. 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).