ifo WORKING PAPERS

353
2021

May 2021
(Revised January 2022)

School Track Decisions and Teacher Recommendations: Evidence from German State Reforms

Elisabeth Grewenig

ifo INSTITUTE
Leibniz Institute for Economic Research
at the University of Munich

Imprint:

ifo Working Papers
Publisher and distributor: ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49(0)89 9224 0, Telefax +49(0)89 985369, email ifo@ifo.de
www.ifo.de

An electronic version of the paper may be downloaded from the ifo website:
www.ifo.de

ifo Working Paper No. 353

School Track Decisions and Teacher Recommendations: Evidence from German State Reforms*

This version: January 2022
First version: May 2021**

Abstract

This paper explores the effects of selective admission policies in the context of school tracking. Depending on the federal state in Germany, either teachers or parents have the discretion to decide which secondary school track a child may attend after primary school. Applying a differences-in-differences approach, I exploit variation in the implementation and abolition of binding teacher recommendations across states and over time. Using data from large-scale assessments, I find that binding teacher recommendations significantly improve student achievement in fourth grade, right before track assignment. Effects persist into ninth grade. Further analyses show that effects are driven by increased time investments in students' skill development.

JEL Code: I21, I28, J24
Keywords: School tracking, admission policies, student performance

Elisabeth Grewenig
ZEW – Centre for
European Economic Research
68161 Mannheim, Germany
grewenig@zew.de

∗For helpful comments and discussion, I would like to thank participants at the ifo Center for the Economics of Education in Munich as well as the CESifo EffEE Workshop, virtual. I am grateful to Frauke Witthoeft for outstanding research assistance. The paper benefited from helpful discussions with Benjamin Arold, Peter Bergman, Marcel Helbig, Mikael Lindahl, Lukas Mergele, Mårten Palme, Alexander Patzina, Marc Piopiunik, Florian Schoner, Norbert Sendzik, Ludger Woessmann, and Larissa Zierow. Financial support by the Leibniz Competition (SAW 2019) is gratefully acknowledged. Among others, this paper uses data from the IGLU, PISA and IQB National Assessment Studies. The data were made available by the Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB) with project numbers 1803-05a/b, March 2018, 1910-08a/b, October 2019 and 2004-09a, April 2020. PISA-2000 was conceived in Germany as a national research program by the German PISA Consortium (Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, Manfred Weiß). The lead was Professor Dr. Jürgen Baumert, Max Planck Institute for Educational Research, Berlin. Results of the primary research are published in Baumert et al. (2001b, 2002, 2003). The survey instruments are documented in Kunter et al. (2002). I thank the German PISA Consortium and the Research Data Center (FDZ) in Berlin for permission and support of the secondary analysis.
** The first version of this paper was prepared while the author was working at the Center for Economics of Education at the ifo Institute.

1 Introduction

School tracking, i.e. the streaming of students into different school types according to their career goals or educational needs (e.g., Betts, 2011; OECD, 2013), is common in many educational systems around the world and has been proven to bear important consequences for students' educational attainment and wages later in life (e.g., Meghir and Palme, 2005; Hanushek and Woessmann, 2006; Pekkala Kerr et al., 2013; Canaan, 2020).

To govern the allocation of students into different schools, tracking systems need to rely on some sort of admission or placement policies which can vary in their degree of selectivity. While selective tracking systems base track admission on prior performance and allow students only to attend higher tracks if they show proof of earlier accomplishments, non-selective tracking systems offer free choice of school tracks by allowing students to sort themselves into the different streams according to their preferences. Both, selective and non-selective tracking policies, are widely used between and within countries. Figure 1 lists all OECD countries where 15 year old students are tracked into different school types (see OECD, 2013, for corresponding information on the categorization of countries). It reveals that 58 percent of those students attend selective schools where admission depends on previous academic performance or recommendations of feeder schools. The between-country difference in academic selectivity is also notable as the respective shares vary heavily, ranging from more than 80 percent in Hungary, Netherlands and Japan to less than 10 percent in Greece. Despite their common and diverse usage, evidence on the impact of selective vs. non-selective tracking systems is relatively scarce. This paper studies how changes in the selectivity of tracking affect students' educational outcomes in the short and medium run.

To that end, I exploit German state-level variation in the tracking procedure. In Germany, tracking takes place early, at the transition from primary to secondary schools (at age 10). While all students receive a recommendation by their primary school teacher on which school track she advises the child to pursue, the extent to which the content of this recommendation is used for admission to the academic schools (the highest track) and, therefore, the selectivity of the tracking system, varies by state.[1] States with binding teacher recommendations have a selective tracking system since children can only attend academic schools if they have a recommendation to do so. States with non-binding recommendations have a non-selective tracking system as each student still receives a recommendation but may attend any school type regardless of the recommendation outcome. Over time, several states reformed the binding nature of teacher recommendations, with some moving from non-binding to binding,

[1] The use of teacher recommendation is not Germany-specific. Other countries, such as Italy or the Netherlands, also employ teacher recommendations to facilitate school track assignment (e.g., Checchi and Flabbi, 2013; Timmermans et al., 2018).

some from binding to non-binding and others moving back and fourth. Thus, the reforms induce between-state variation over time in whether teacher recommendations are used as selection criterion for school track admission. As all students in all states always receive a recommendation by their primary school teacher, the reforms allow me to isolate the effect of their selective nature from any informational values that recommendations may also have.

There are several reasons why selective tracking systems, such as those with binding teacher recommendations, can affect students' outcomes in the short to medium run. First—while still attending primary school—binding recommendations can serve as incentive for children and parents to increase students' academic performance in order to be accepted to the academic schools (see e.g., Benabou and Tirole, 2000; Lindo et al., 2010, for a more general discussion on performance standards). Second, if teachers are better able to asses a child's academic potential than parents, binding recommendations may lead to a more efficient allocation of students to the different school tracks. Over time, achievements gains from early incentive provision can develop further in secondary schools as skill formation is a dynamic process (e.g., Cunha et al., 2006; Cunha and Heckman, 2007). Both would also lead to improved educational outcomes for students in the longer run. On the other hand, hidden psychological costs from incentivizing students, e.g. crowding out of intrinsic motivation, may also dominate—especially in the longer run (see e.g., Benabou and Tirole, 2000). Likewise, parents may be better informed about their child's ability and use this information as basis for their decision. Then, binding recommendations would lead to worse educational outcomes.

To analyze the reform effects empirically, I combine information on state reforms which took place during the 1990s and 2000s, with several data sources. First, I use individual student-level data on fourth graders, stemming from the 2001 and 2006 extensions of the Progress in International Literacy Study (IGLU-E) as well as the 2011 and 2016 National Assessment Studies (NAS) to analyze short-term effects on students' achievement in primary school, i.e. *prior to* track assignment. Second, I use administrative school data from the German Statistical Offices to analyze medium-term effects on academic school attendance in grades five to nine, i.e. *after* track assignment. I complement the analyses with individual student-level data of ninth graders, stemming from the 2000, 2003 and 2006 extensions of the Programme for International Student Assessment (PISA-E) as well as the 2009, 2012 and 2015 NAS to investigate academic school attendance and performance of students attending ninth grade.

Using a differences-in-differences approach that controls for fixed differences between states and years, I investigate the effect of selective vs. non-selective tracking by comparing outcomes of students attending school in states that reformed the bindingness of teacher recommendations to outcomes of students attending school in states that did not implement such reforms. I

find that binding teacher recommendations have a sizable impact on student achievement in primary school before track assignment. Conditional on state and school-year fixed effects as well as a rich set of sociodemographic controls, reading (math) achievement is 5.6 (11.6) percent of a standard deviation higher for students who require a teacher recommendation for academic school attendance than for students who have free choice of secondary schools.

Detailed time-use data for 9–10 year old children from the German Socio-Economic Panel (GSOEP) allow me to shed some light on the mechanisms underlying these effects. I find that children spend significantly more time with reading as well as with activities deemed as being rather *conducive* for child development, such as doing sports or making music, which suggests that the achievement effects are likely due to increased time investments. Similarly, children spend more time with their family, suggesting that parents respond accordingly by supporting their child more often after the reforms. Parents also consult their child's teacher more frequently which likely indicates increased 'lobbying' for academic track recommendations.

Next, I show that binding teacher recommendations affect educational outcomes of students in secondary schools as well. Under binding recommendations, students are slightly less likely (albeit not statistically significantly) to attend academic schools in fifth grade. Beyond that, I show that these small and negative effects—measured immediately after tracking—mask important positive effects in the medium run. In particular, reform effects gradually *increase* throughout grades and ninth grade academic track attendance is significant 2 percentage points *higher* under binding recommendations. This pattern suggests that the reforms reduce the rates at which students transfer to lower track schools during their secondary school career. Less 'downtracking' eventually increase academic school attendance rates by grade nine. Further analyses show that ninth grade students also perform slightly better in standardized reading tests and have better grade point averages in the subjects German and math (albeit not statistically significant).

Finally, I explore effect heterogeneities by students' socioeconomic background. For the sub-sample of students with information on family background, reform effects are by and large homogeneous across different subgroups.

A series of robustness tests support the main results. Results remain robust to controlling for a rich set of contemporaneous school reforms, economic and education input factors. Moreover, additional analyses suggest that there are little significant differences in pre-trends between reforming and non-reforming states. I also show that results are robust to including state-level controls for government ideology. Furthermore, students who are exposed to binding teacher recommendations do not enter primary school later, nor are they more likely to strategically repeat a grade. Finally, I implement the diagnostic tools by de Chaisemartin and D'Haultfœuille (2020) to show that biases arising from negative weights are likely not an issue in my setting.

This paper mainly contributes to two strands of the existing literature. First, it extends the literature on school tracking. There is substantial heterogeneity across countries in the extent and age at which students are tracked (Betts, 2011). Most of the existing tracking literature investigates the effects of earlier vs. later tracking (e.g., Meghir and Palme, 2005; Hanushek and Woessmann, 2006; Malamud and Pop-Eleches, 2011; Pekkala Kerr et al., 2013; Borghans et al., 2020; Canaan, 2020) and finds that later tracking increases educational attainment and wages in adulthood, mostly for students from disadvantaged families.[2] Yet, very little is known about the impact of selective vs. non-selective tracking procedures.

Second, the paper relates to the literature on selective admission requirements which are ubiquitous in education. Decisions to receive a scholarship, to participate in more advanced courses, or to be promoted into the next grade often depend on some sort of prior student achievement. Consequently, researchers have examined the impact of admission requirements in various contexts. In the short run, empirical studies have found that selective admission requirements often incentivize students to meet the criteria, thereby increasing overall performance (e.g., Angrist and Lavy, 2009; Pallais, 2009; Jackson, 2010; Behrman et al., 2015; Barrow and Rouse, 2016; Lichtman-Sadot, 2016). Few studies find zero or negative effects for specific student subgroups, mostly from the lower end of the achievement distribution (e.g., Leuven et al., 2010; Lindo et al., 2010). In the longer run, positive effects of admission requirements can be dominated by hidden psychological costs (see e.g., Benabou and Tirole, 2000). Empirical findings are therefore rather inconclusive. While some studies find persistent positive effects of admission requirements (e.g., Jackson, 2010; Leuven et al., 2010) , others find zero (e.g., Abdulkadiroğlu et al., 2014) or even negative (Lindo et al., 2010) effects on students' educational or labor-market realizations.

This paper combines the two literature strands by examining selective admission policies in the context of school tracking. In that sense, closely related is the study by Guyon et al. (2012) who leverage a reform in Northern Ireland that increased the overall proportion of students in so-called *grammar schools* (the highest, most selective track). Before the reform, students were selected based on their performance on a national test whereas after the reform admission was left more to the parents' choice. The authors find large and significant improvements in students' long-term educational outcomes due to the reform. But it remains unclear to what extent the change in the selectivity of the tracking procedure affected short-term behavior as well.

[2]For Germany, Matthewes (2020) and Piopiunik (2021) find positive achievement effects from decreasing tracking intensity by combining the two lower tracks into one comprehensive track. Conversely, Piopiunik (2014) analyses a reform-induced increase in tracking intensity in the state of Bavaria and finds large achievement losses for students who are subject to more intense tracking after the reform. Finally, Dustmann et al. (2017) finds no effects of academic track exposure on educational attainment and earnings in the long run for students who are at the margin between two tracks, exploiting birth cut-off rules as an instrument.

Several other papers have investigated the German reforms on the bindingness of teacher recommendations. The sociological literature mainly focuses on evaluating single reforms that took place in one particular state, either in cross-section analyses (e.g., Neugebauer, 2010; Dollmann, 2011, 2016) or within a differences-in-differences framework (e.g., Jähnen and Helbig, 2015; Roth and Siegert, 2015, 2016). These studies generally find none to small negative effects of binding recommendations on academic school attendance, shortly after the transition to secondary schools (usually in grade five). Similarly, an economics paper by Osikominu et al. (2021) analyzes a single reform in the state of Baden-Wuerttemberg in 2011 and finds small negative effects of binding teacher recommendation on academic track attendance in fifth grade, the year after students have been tracked. Finally, the contemporaneous paper by Bach and Fischer (2020) investigates short-term effects on students in primary school using several identification strategies: (triple) differences-in-differences specifications that exploit the two latest reforms in Baden-Wuerttemberg and Saxony-Anhalt as well as value-added approaches that investigate effects of binding teacher recommendations more comprehensively. My results that binding teacher recommendations affect student achievement and incentivize study effort of fourth graders are in line with their findings.

I depart from previous studies in two important ways. Contentwise, I provide the first comprehensive analysis of binding teacher recommendations in the short *and* medium run. I find that binding teacher recommendations do not only improve educational outcomes among students in primary school, but effects extend to students in secondary schools. In terms of academic school enrollment, I show that small negative effects in fifth grade mask important positive attendance effects that evolve over time. Particularly, binding teacher recommendations seem to reduce the incidences of students transferring to lower track schools throughout their secondary school careers, resulting in positive attendance effects by grade nine. Methodologically, I apply a 'generalized' differences-in-differences approach that does not only identify from reforms in one or two states, but simultaneously exploits up to 10 state-level changes. Thus, I paint a detailed picture of how reforms affect students across the country. The analyses in this paper are built on a novel data base that combines data on fourth graders from IGLU-E and NAS as well as on ninth graders from PISA-E and NAS, respectively. As these studies draw representative samples of students in all states and mandate participation, estimated effects are informative for the average student body.

Overall, my results suggest that selective tracking policies—in form of binding teacher recommendations—can lead to persistent improvements of students' educational performance in the short and medium run. These findings have important implications for the scientific and political discourse by providing direct evidence that the particular design of tracking policies is important for human-capital formation.

The remainder of the paper is structured as follows: Section 2 provides detailed background information on the school system and the tracking procedure in Germany. Section 3 introduces the empirical model and the data. Section 4 presents main results on the short and medium-term effects of introducing binding teacher recommendations. Section 5 reports robustness checks, and section 6 concludes.

2 Institutional Background

This section first provides an overview of the German school system. It then describes the role of teacher recommendations in general as well as the reforms on the bindingness of teacher recommendations in particular.

2.1 German School System

In Germany, responsibility for the school system and therefore decisions regarding educational policies are vested in the 16 federal states. The German constitution even prohibits the federal government to exert influence on the educational policies of the states. Yet, a general assembly of all state ministers of education called *Kultusministerkonferenz (KMK)* aims to harmonize education policies countrywide. Consequently, the general outlook of the school system is fairly uniform, as are degrees or teacher employment conditions. However, some other education policies may still differ across states or may be implemented or abolished at different points in time.[3]

Figure 2 provides an overview of the German school system. Compulsory schooling extends from the age of five or six until the age of 18. The comprehensive primary school takes four years (in Berlin and Brandenburg six years) and provides basic education in mathematics, German and several other science and social subjects. In primary school, students are usually taught all main subjects by the same teacher.

Upon completion of primary school, students move on to secondary schools. At this point, children are assigned to one of three different tracks: the basic and intermediate track last though grades nine and ten, respectively, and prepare students for apprenticeship training or other forms of vocational education. The academic track ends with grade 13 (or 12) and leads to the university entrance qualification *Abitur*. While nowadays many different school types

[3]It is thus not surprising that several economic research papers have exploited the characteristics of the German federal system to evaluate educational reforms implemented over time (e.g., Pischke, 2007; Pischke and Wachter, 2008; Dustmann et al., 2017; Marcus and Zambre, 2019; Matthewes, 2020; Obergruber and Zierow, 2020)

(including comprehensive school models) incorporate both the basic and the intermediate track, the academic track is primarily offered by the academic school *Gymnasium*.[4]

Academic school attendance is relatively common in Germany. Overall, 37 percent of 14 to 16-years-old children attended a *Gymnasium* in 2015 (own calculations based on German Microcensus 2015[5]). In addition, students attending academic schools differ substantially in their sociodemographic characteristics. Appendix Table A1 provides descriptive results from regressing academic school attendance on various sociodemographic characteristics. For instance, children at the *Gymnasium* are more likely to be female, live in East Germany, live together with their parents or have parents with a university entrance degree. Switching to higher track schools is possible, but rather uncommon. In 2000, only 1.5 percent of students switched to a higher track at any grade throughout grades 5–9. However, the best students from the lower tracks often transfer to higher tracks after successful degree completion.

2.2 The Role of Teacher Recommendations

As school track decisions are made at the end of primary school, the transition from primary to secondary school marks an important milestone in the students' further educational careers. To formally structure this transition—particularly to academic schools—entrance examinations were conducted in most of the western German states until the 1960s. As public critique grew in the 1950s (Gass-Bolm, 2005), the state ministers of education decided to facilitate the selection process for the academic schools (Herrlitz et al., 2009). Over time, more and more states abolished entrance examinations and replaced them with teacher recommendations to guide school track decisions for students and their parents (Helbig and Nikolai, 2015). Teacher recommendations were also adopted by the East German states after reunification in 1991. Today, all children in Germany obtain a recommendation at the end of their primary school.

Recommendations are issued by the students' primary school teachers and entail explicit information on which school type she thinks the child should pursue. They are issued in the students' final year of primary school, shortly after the first semester, i.e. in February or March. Recommendations are mainly based on students' grades in their mid-term report card. In some states teachers additionally base their recommendation on their assessment of the student's socio-emotional maturity (Baumert et al., 2010).[6] Appendix Figure A1 reports which criteria determine the recommendations. Students' math and German grades are very strong

[4]Academic schools are also the most important school type on which students receive the university entrance qualification.

[5]Research Data Centres of the Federal Statistical Office and the statistical offices of the Laender, Microcensus, census year 2015.

[6]For instance, Bavaria uses the students' grade as the only criterion. The Bavarian regulation states that *[i]n order to obtain a recommendation for the Gymnasium (academic track) a student needs to have a GPA of at most 2.33 in the subjects German, math as well as science and local history* (with grades ranging from 1 to

or strong determinants for nearly all teachers. Among soft skills, commitment, concentration and self-reliance are ranked highest with again almost all teachers considering them as (very) strong criteria.

While teacher recommendations are issued to all students in Germany, their bindingness differs across states. In some states, teacher recommendations are non-binding, i.e. the content of the recommendations is purely informative and students can transfer to any school type regardless of the recommendation outcome. In other states, teacher recommendations are binding, i.e., students can only attend the academic schools if they have a recommendation to do so.[7] In these states, if students wish to transfer to the academic schools without an appropriate recommendation, they must pass additional entrance examinations or trial lessons (Kultusministerkonferenz, 2015). Conversely, if students wish to attend a lower track than recommended, they can always do so.

Under both regimes, it can happen that applications to particularly popular schools exceed their capacities. Under these circumstances, school admission depends on further criteria, such as proximity to school or sibling attendance. As a consequence, students and parents are not eligible for attending any particular school of choice, but they are for attending the general school type of choice. For the case of non-binding teacher recommendations this specifically means that children and parents have a legal claim for academic school attendance regardless of the recommendation outcome.

2.3 Reforms on the Bindingness of Teacher Recommendations

Since the general adoption of teacher recommendations, states have frequently reformed their binding nature: Several states have abolished binding recommendations to replace them with non-binding ones, and vice versa while other states have switched back and forth. Figure 3 and Appendix Table A2 present a summary of the state regulations for the time period considered in this paper since the 1990s.

State reforms are usually accompanied by emotional public debates. Arguments for binding recommendations come from the conservative or liberal camp and are based on the idea of the *Gymnasium* as an elite school (Fokken, 2020): Accordingly, the government should strengthen the academic schools and prevent a decline in performance due to the presence of unsuitable students (e.g., Die Welt, 2014). In addition, teachers are supposedly better able to evaluate the potential of their students than their parents (e.g., Breyton, 2018). Opposing arguments

6 and lower values indicating better grades). In Schleswig-Holstein, for instance, the *recommendation shall be based on students' maturity considering the students' current grades* (Kultusministerkonferenz, 2015).

[7] The bindingness mainly applies to the academic school *Gymnasium*: In almost all states with binding recommendations, students with a recommendation for the basic track can nevertheless attend an intermediate track school (Kultusministerkonferenz, 2015).

come from the left-leaning camp (Fokken, 2020). The prevailing view here is that non-binding recommendations can provide access to the academic schools for broad groups of the population. Moreover, the bindingness may put strong pressure on students in third and fourth grade and four years may be too short a time to make binding statements about students' future potential (e.g., Schenk, 2010; Otto and Schenk, 2011).

Given the lively political debate, one would expect to see reforms in the bindingness of teacher recommendations particularly after ideological changes in state governments. And indeed, seven of the ten reforms since the early 1990s were implemented after governmental changes (see Table A2). Among those seven, all four that involved a change from binding to non-binding recommendations were introduced after a more social government had replaced a more conservative one. Conversely, the remaining three reforms including a change from non-binding to binding recommendations were introduced after a more conservative government had replaced a more social one. To address the potential issue of non-random reform introduction, my main specification controls for a variety of different educational input factors which are observed at the state-year level (see section 3.1 for details). Section 5 also analyzes potential effects of government ideology on student outcomes. Reassuringly, the main effects of binding teacher recommendations are robust to controlling for state government ideology.

3 Empirical Strategy and Data

In this section, I first describe the empirical strategy and then present the data used for the analyses.

3.1 Empirical Strategy: Differences-in-Differences Approach

My identification strategy exploits variation in the implementation and abolition of binding teacher recommendations across German states and over time in a differences-in-differences framework. By controlling for fixed differences between states and years, I compare outcomes of students attending school in states that changed the bindingness of teacher recommendations to outcomes of students attending school in states that did not. The empirical model can be formalized by the following equation:

$$Y_{ist} = \alpha + \beta Bindrec_{st} + \gamma R_{st} + \delta E_{st} + \lambda X_{ist} + \eta_s + \mu_t + \varepsilon_{ist} \quad (1)$$

where Y_{ist} is the outcome of interest (e.g., student achievement) for student i who attends school in state s and is tracked into different schools in year t. $Bindrec_{st}$ is the treatment indicator which equals 1 if the recommendation is binding and varies at the state and year

level. To account for differences across states and over time, I include state (η_s) and year (μ_t) fixed effects. Since treatment varies at the state level, I use a conservative inference and cluster standard errors at the state level to account for potential correlation of error terms within states across years (Athey and Imbens, 2018).[8] Moreover, regressions are weighted by students' sampling probabilities, giving equal weight to each wave.

During the observation period, Germany has undergone several major education policy changes that potentially affect students' outcomes in primary and secondary schools. To rule out any biases arising from omitting these reforms, I include R_{st}, a vector of reform indicators that vary by year and state. The vector entails the duration of academic schools—whether the academic schools take eight or nine years—,[9] the intensity of school tracking,[10] the duration of primary school—whether primary schooling takes four or six years—,[11] the basis of the teacher recommendation—whether they are only based on students' grades or also on their socio-emotional maturity—,[12] as well as whether the recommendation has to be explicitly requested.[13] Overall, the pairwise correlations of the different reform indicators with my main treatment variable are relatively low and range from between ρ=-0.25 to ρ=0.36.[14] Therefore, I do not expect the contemporaneous occurrences of these reforms to severely bias the main results which is also confirmed by the fact that results below remain robust to including the corresponding reform vector.

To further avoid biases from the fact that the timing of recommendation reforms may not be random to the economic and educational performance of the states, I include E_{st} and X_{ist} as additional controls. E_{st} is a vector of economic and educational measures that vary by year and state. These entail GDP as *overall economic performance measure*, but also average school spending, the number of classes in primary school, average class size in primary school, average hours of instruction in primary school as well as share of full-time employed primary

[8] I additionally present wild cluster bootstrap p-values, relying on Roodman et al. (2019) in all main tables. These p-values account for a limited number of clusters when analyzing at most sixteen German states.

[9] Several states reduced the length of the academic school while simultaneously increasing the instruction hours in the remaining years. The effects of those reforms have been investigated by Andrietti and Su (2018) or Marcus and Zambre (2019).

[10] Several states reformed whether students attending the two lower tracks are taught comprehensively or further streamed into two separate tracks. The effects of such reforms have been investigated by Matthewes (2020).

[11] Except for the states Berlin and Brandenburg where primary school takes six years, students are tracked after grade four. A few sates experimented with 'later tracking' by introducing (and again abolishing) so-called 'orientation grades' where students were comprehensively taught until grade six and subsequently tracked into the different school types. These reforms are described by Helbig and Nikolai (2015).

[12] While most states have a standing rule on the criteria used for the outcome of the recommendations, in a few cases states have reformed those criteria together with the bindingness of recommendations.

[13] This was only the case in Bavaria until 2008. In all other states and years, students automatically receive a teacher recommendation before transition to secondary schools.

[14] The only exemption builds the correlation between the bindingness of teacher recommendation and the basis of the recommendation which is expectedly higher and amounts to ρ=-0.60.

school teachers as *school input factors*. The vector also entails the average share of students starting primary school late. X_{ist} is a vector of various school- and student-level characteristics, including student gender, student age, immigration background, parental education, parental occupation, and books at home as well as community location and public school status.

The key identifying assumption is the standard differences-in-differences assumption: Conditional on the rich set of included control variables at the student, school and state level, in the absence of reforms the change in student outcomes in states that reformed the bindingness of teacher recommendations would have been similar to the change in student outcomes in states that did not reform at a given point in time.[15] I will come back to a detailed discussion of potential violations of this assumption in section 5.

3.2 Data

To analyze the effects of binding teacher recommendations empirically, I combine information on state reforms with the following three data sources: (i) individual-level data on students in fourth and ninth grade from nation-wide assessment studies, (ii) individual-level data on 9–10 year old children from the German Socio-Economic Panel and (iii) administrative state-level data on the education system from the German Statistical Office. This section presents the four components in turn.

3.2.1 Data on State-wide Educational Reforms

I collect data on a series of state-wide educational reforms related to the tracking procedure. First, I compile information on the bindingness of teacher recommendation using the following sources: For the reforms before 2010, I draw on Helbig and Nikolai (2015). For the subsequent reforms, I gather information from Kultusministerkonferenz (2015), newspaper articles (e.g., Otto and Schenk, 2011), plenary protocols from sessions of the state parliaments as well as individual correspondences with the 16 state ministries of education.

Similarly, I compile information on several other educational reforms which took place during the observation period. The corresponding reform indicators serve as control variables and include the following: the basis of the teacher recommendations (in particular whether they are only based on students' performance or also on their general maturity), whether the recommendation has to be explicitly requested, the duration of primary schools (four vs. six years) as well as the intensity of school tracking (i.e., whether students attending the two lower

[15] Further relaxing the identifying assumption by including state-specific linear time trends is difficult given the data structure. As I am limited to four (in the fourth grade sample) and six (in the ninth grade sample) time-series observations for each state, adding a linear time trend for each state renders coefficients too imprecise for clear inference.

tracks are taught comprehensively or further tracked into two different school types). Finally, I use information on the duration of the academic schooling (eight vs. nine years) from Marcus and Zambre (2019).

3.2.2 Student Assessments

To analyze short-term reform effects on primary school students, I combine data from the German extension of the Progress in International Reading Literacy Study (IGLU-E) with data from the National Assessment Study (NAS).[16] Both studies are repeated cross-sections, testing students at the end of fourth grade (between April and July). While both studies assess students in German (reading), NAS additionally assesses math. The studies were administered in 2001 and 2006 (IGLU-E) as well as in 2011 and 2016 (NAS) and are representative for all German states.[17] Neither IGLU nor NAS follow individual students over time. However, repeated testing of fourth graders allow me to build a pseudo-panel of German states observed every five years. Reading score are generally comparable across tests and waves as NAS was explicitly designed to emulate IGLU tests (Pietsch et al., 2009; Bos et al., 2012).[18] In each study and wave, primary schools are randomly drawn and within each of those schools, a whole class then randomly participates in the tests (see Richter et al., 2014; Schipolowski et al., 2019, for more details on test administration).

To analyze medium-term effects on students in secondary school, I combine data from the German extension of the Programme for International Student Assessment (PISA-E) and the National Assessment Study (NAS).[19] Whereas the international version of the PISA test samples 15 year old students, the German extension tests ninth graders. PISA-E and NAS thus build repeated cross-sections, testing students at the end of grade nine (between May and July). The tests have been administered in 2000, 2003 and 2006 (PISA-E) as well as in 2009, 2012 and 2015 (NAS) and are representative for ninth graders in the 16 German states.[20] Again, neither PISA nor NAS follow individual students over time, but repeated testing allow me to build a pseudo-panel of German states observed every three years.

[16] For further details see Bos et al. (2007, 2010) and Stanat et al. (2014, 2019).

[17] While all 16 German states participate in all four waves, in 2001 only seven states draw larger sample sizes that were fully representative.

[18] In addition, Böhme et al. (2014) compare reading scores among students who have been tested in both studies and show a strong correlation (ρ=0.86) between the test scores produced by the IGLU and the NAS items.

[19] NAS replaced PISA-E after 2006. See Prenzel et al. (2007), Baumert et al. (2009), Prenzel et al. (2010), Köller et al. (2011), Pant et al. (2015), and Stanat et al. (2018) for details.

[20] In 2006, the KMK decided to replace the state-level representative samples of PISA-E by the NAS (Kultusministerkonferenz, 2006). Since then, PISA is still conducted in Germany but with much smaller sample sizes to only represent the overall student body of 15 year old in Germany.

While PISA regularly tests relevant skills in math *and* reading, NAS alternates tested domains every other wave. Consequently, NAS 2009 and 2015 assess reading while NAS 2012 assesses math. Therefore, all results on reading achievement are conducted without students tested in 2012.[21] Achievement scores are generally comparable across studies and waves as the NAS was explicitly designed to emulate the PISA-E testing procedure (Hartig and Frey, 2012; Böhme et al., 2014). In each study and wave, random samples of schools were drawn to be representative at the federal state level and within each school, one class randomly participated in the tests (see Baumert et al., 2001a, 2002, 2004; Sachse et al., 2012; Lenski et al., 2016; Schipolowski et al., 2018a,b, for more details on test administration).

Tables 1 and 2 present student-level descriptive statistics of fourth and ninth graders, respectively. I consider a student as subject to binding teacher recommendations if the recommendation was binding in her current state of school attendance in the year she was tracked into the secondary schools.[22] In addition, Appendix Figure A2 graphically represents an overview containing all reforms since the early 1990s alongside information about the cohorts covered in the different data sets. The fourth grade sample consists of approximately 70,000 students who were streamed into secondary schools between 2001 and 2018 and the ninth grade sample consists of more than 220,000 students who were streamed into the secondary schools between 1994 and 2012. Student assessments were accompanied by comprehensive school, student and parent questionnaires covering a wide range of questions on students' sociodemographic characteristics and family background. While test participation is always compulsory for students, completing accompanying family background questionnaires is voluntary. As a result, non-response rates for the background questionnaires are much larger than for the test items (for example, response rates to the parent questionnaire are 72 percent in the fourth grade sample and 80 in the ninth grade sample). For the final analyses, I select a core set of student and school controls available in each wave.[23] To avoid sample selection bias from non-response in the background questionnaires, I impute missing values by using the respective state-by-year mean and include a set of imputation dummy that equal one for imputed values and zero otherwise.

[21] While it is in principal also possible to evaluate math achievement for the subset of students tested in 2000, 2003, 2006 and 2012, I abstain from the respective analyses due to the following two reasons: (i) By excluding students assessed in 2009 and 2015, I loose a substantial amount of observations (more than 100,000). (ii) Unlike the reading test, the math test was re-scaled in 2003, which renders the comparability of PISA 2000 to the remaining waves unclear.

[22] In the fourth grade sample, a students' year of tracking depends on the duration of primary school. In the ninth grade sample, a students' year of tracking depends on the duration of primary school in her current state of school attendance as well as on self-reported grade retention. Important for identification, retention was not affected by reforms on the bindingness of teacher recommendations (results available upon request).

[23] Due to lack of availability I use school-level controls only for the analyses of the fourth graders, but not of the ninth graders.

3.2.3 The German Socio-Economic Panel

To further investigate behavioral responses of students and their parents, I use data from the German Socio-Economic Panel (GSOEP). Since 2010, the GSOEP contains a mother-child questionnaire administered to parents of 9–10 year old children (see Schröder et al., 2013, for further information). The questionnaire collects detailed information on children's daily lives and is supplemented with questions on background characteristics of children and parents.

I focus on the following variables: First, I exploit detailed time-use information on leisure activities:[24] (i) reading; (ii) watching TV; (ii) playing on the computer; (iv) surfing on the internet; (v) listening to music; (vi) making music; (vii) dancing or theater; (viii) doing sports; (ix) doing technical work; (x) drawing, and (xi) spending time with the family. On a 5-point scale parents could indicate how much time their children spend on each activity.

Following Grewenig et al. (2020), I group activities into four categories: reading (which is directly related to students' reading test scores), activities rather *detrimental* to child development (activities (ii)–(v)), activities rather *conducive* to child development (activities (vi)–(x)) and *family* activities (activity (xi)).[25] Second, I exploit information on whether parents report frequently consulting their childrens' teachers.[26] The sample consists of approximately 4,400 students who were streamed to secondary schools between 2005 and 2017.

3.2.4 Administrative School Data

Finally, I use data on general schools (*allgemeinbildende Schulen*) provided by the German Statistical Office (Statistisches Bundesamt, 1991-2016). The administrative data comprises annual state-level information on the number of students in each track and grade for the years 1991 to 2016. To obtain state-wide information on the share of academic track students, I divide the number of academic school students in each grade and school year by the total number of students in the respective grade-year cell. I enrich this data with the share of issued recommendations for the academic schools collected through personal correspondence with the 16 federal ministries of education.[27] The administrative data further includes comprehensive state-wide information on school input factors which are used as additional controls (see section 3.1 for details).

[24] I only use items which have been consistently asked throughout all GSOEP waves.

[25] Grewenig et al. (2020) show that categorization of *detrimental* and *conducive* activities reflects parental beliefs about how beneficial those activities are for child development.

[26] Teacher's consulting is elicited as follows: *'How often do you or other family member seek contact with the school?'* Respondents could tick a box if they *'frequently consult teachers outside of regular meeting hours.'*

[27] Overall, I receive data on recommendations for eight states and multiple school years (108 state-year observations).

4 Main Results

In this section, I discuss the main results. First, I consider short-term effects on students in primary school. Second, I estimate medium-term effects on students in secondary schools. Finally, I explore potential effect heterogeneities with respect to the students' family background.

4.1 Students in Primary School

This section sheds light on the academic performance of students in primary school *before* school track assignment. I first analyze how the reform affects achievement of students in fourth grade. Then, I turn to examining behavioral responses of students, parents and teachers.

4.1.1 Student Achievement

I start the discussion on short-term effects with student achievement in reading and math among fourth graders. If binding teacher recommendation indeed serve as incentive to improve academic performance prior to track assignment, one would expect to find positive reform effects on the outcomes discussed here.

Panel A of Table 3 presents the main differences-in-differences results on student achievement, using equation 1. For ease of interpretation, the dependent variables are z-standardized reading test scores in grade four. Column (1) shows the baseline results, controlling only for state and year of transition fixed effects. Reading achievement among students who require a respective teacher recommendation to attend academic schools is 6.4 percent of a standard deviation higher than that of students with free choice of secondary schools.

In columns (2)–(4), I gradually include controls that account for potential differences (i) in educational reforms that were contemporaneously implemented, (ii) in the overall economic condition and schooling input factors, and (iii) in the sociodemographic composition of the student body within states and over time. The estimates in column (4)—controlling for the full set of background characteristics—are slightly smaller in size than the baseline estimates, suggesting that student, school and economic controls are somewhat correlated with the reform indicator. Most importantly, all estimates stay positive and statistically significant. In the full specification, introducing binding recommendations is associated with a significant increase in reading achievement by 5.6 percent of a standard deviation.

To explore the dynamic of reforms effects over time, Panel A of Figure 4 additionally depicts non-parametric event-study estimates which are obtained by including an indicator for the first cohorts with binding teacher recommendations as well as lead and lag indicators besides state and year fixed effects. The depicted pre-reform effect is in line with the common trend assumption.

The coefficient on the lead dummy is economically and statistically insignificant, suggesting that students in states that switched to binding teacher recommendations were on similar pre-reform trends as students in states that remained with free choice of secondary schools. Though shy of significance, the first cohort exposed to binding teacher recommendations experiences an increase in reading test scores and the improvement seems to remain persistent over time.

Panel B of Table 3 presents differences-in-differences results on math achievement which is only assessed in 2011 and 2016. Consequently, identification here stems from changes in the achievement of students in states that reformed their teacher recommendations between 2011 and 2016 (Baden-Wuerttemberg and Saxony-Anhalt). I find that binding teacher recommendations substantially increase students' math scores. Reassuringly, reform-effect patterns through columns (1) to (5) are remarkably similar to those found for reading. In the full specification, math achievement among students who require a respective teacher recommendation to attend academic schools is 11.6 percent of a standard deviation higher than that of students with free choice of secondary schools.

In sum, the results in Table 3 suggest that selective tracking in form of binding teacher recommendations improve students' academic achievement in primary school even before track assignment takes place. Considering the rule of thumb that average student learning in a year is equivalent to about one-quarter to one-third of a standard deviation, the reform effects amount to what students roughly learn during a fifth (reading) to a third (math) of a school year.

4.1.2 Behavioral Responses

I now turn to investigating behavioral responses as potential mechanisms for the achievement gains revealed in the previous section. Behavioral responses could stem from students, parents and teachers which are investigated in turn.

First, I examine students' responses. One obvious explanation for the positive achievement effects is that binding teacher recommendations incentivize students to put more effort into studying. To explore this channel, I draw upon detailed time-use information on various leisure activities of 9–10 year old students collected as part of the GSOEP. Panel A of Table 4 depicts reform effects on children's time spent with reading, standardized to have mean zero and standard deviation one. When teacher recommendations become binding, children spend significantly more time with reading (column(4)), suggesting that the reform effects on students' test scores are indeed driven by increased time investments into the development of reading skills. Similarly, Panel B depicts reform effects on an index, summarizing child's time spent with other leisure activities deemed as being rather *conducive* for skill development, using the method by Kling et al. (2007). The index includes making music, dancing, sports,

technical work, and drawing. The results suggest that reforms also significantly increase the amount of time that children spend with these rather conducive activities. In contrast, Panel C depicts reform effects on time spent with activities deemed as being rather *detrimental* to child development, i.e., watching TV, playing computer, surfing on the internet or listening to music. The estimates indicate zero to slightly negative (albeit not significant) reform effects on time spent on these rather detrimental activities. Thus, the bindingness reforms do not only serve as incentives for children to spend more time with activities directly related to skills taught in school (e.g., reading), but also with other conducive leisure activities more indirectly related to skill development.

Next, I examine parental responses. Similar to students, binding recommendations may serve as incentives for parents to support the skill development of their children. To explore this channel, I investigate how much time children spend with their family. Panel D of Table 4 shows significant positive reform effects, suggesting that binding teacher recommendations also encourage parental time investments. Besides, the reforms may induce parents to exert influence on the content of the teacher recommendation. For instance, parents could seek out the child's teacher more frequently to 'lobby' for an academic school recommendation. Panel E of Table 4 therefore analyzes reform effects on whether parents consult their child's teacher on a regular basis. Column (4) reveals that parents are indeed 13 percentage points more likely to frequently consult their child's teacher when recommendations become binding which is a large and significant increase from a baseline share of 43.4 percent.

Finally turning to teachers' responses, I examine the number of recommendations for academic schools. There are several reasons why teachers would increase academic recommendations in response to the reform. First, they may reward the students' achievement gains discussed in section 4.1.1. Second, they may simply yield to the parents' lobbying efforts discussed above. Finally, they may become less stringent when their students' future educational path hinges more on their evaluation. Appendix Table A3 depicts reform effects on the number of academic school recommendations at the state level. Although not statistically significant, the table provides suggestive evidence that—if anything—the share of academic school recommendations increases.

In sum, results suggest that the achievement gains in primary school are likely due to increased time investments in students' skill development by children and parents. Parents also seek out the child's teacher more frequently. These efforts are then rewarded by teachers who tend to issue more academic school recommendations in response to the bindingness reforms.

4.2 Students in Secondary School

This section investigates educational outcomes of students in secondary school. I first analyze reform effects on academic school attendance. Then, I examine academic performance of students through ninth grade, i.e. several years *after* track assignment.

4.2.1 Academic School Attendance

Reform effects on educational outcomes in secondary school are ex ante less clear. On the one hand, students can be allocated to the different tracks more efficiently if teachers are better able to assess the child's academic potential than parents. Likewise, the positive achievement effects from primary school (see section 4.1.1) can spillover to secondary schools. Both would lead to improved educational outcomes in the medium run. On the other hand, psychological costs of incentivizing students early can dominate in the longer run or parents may hold superior information on the child's academic potential. In these cases, binding recommendations would lead to worse educational outcomes. With respect to school attendance, binding recommendation may also prevent some students from transitioning to the highest track, resulting in fewer students attending academic schools.

To investigate the relevance of the opposing effects, I first analyze reform impacts on attending academic schools, the most important school type where students can obtain a university entrance qualification. Several studies have shown that individuals with a university entrance degree do not only experience a large wage premium on gross earnings of around 42 to 44 percent (see Dodin et al. (2021) and Schmillen and Stüber (2014) for corresponding estimates), but they also show lower risk of unemployment (Hausner et al., 2015) and exhibit a higher life expectancy (Gärtner, 2002). Table 5 presents differences-in-differences estimates on academic school attendance throughout grades five to nine, using annual state-level data from the German Statistical Office. Since I only observe the average share of students attending academic school in a given state and over time, the corresponding estimates are based on an adjusted 'state-level' version of equation 1 which controls for potential differences (i) in educational reforms that were contemporaneously implemented and (ii) in the overall economic condition and schooling input factors.[28]

Column (1) shows negative, statistically insignificant reform effects of one percentage point on the share of fifth grade students, attending academic schools. Finding negative attendance effects directly after track assignment is consistent with earlier studies (e.g., Jähnen and Helbig, 2015; Osikominu et al., 2021). But I also show that the effects observed in fifth grade mask

[28] Basic differences-in-differences results, controlling for state and year fixed effects only, are very similar to those depicted in Table 5 (results available upon request).

important positive attendance effects in the medium run. Specifically investigating academic school attendance through grades five to nine in columns (2) to (5), I find that reform effects gradually increase across grades. In grade nine, the introduction of binding recommendations significantly ($p<0.1$) *increases* academic school attendance by 1.2 percentage points.[29] This pattern suggests that the bindingness reforms decrease the rates at which students transfer to lower track schools during their secondary school career.[30] The reduced incidences of students 'downgrading' school tracks subsequently manifest themselves in higher academic school attendance rates by grade nine.

Data from PISA-E and NAS allow me to additionally analyze academic school attendance at the individual student level. Panel A of Table 6 depicts the reform effects using equation 1. Overall, the estimates confirm significant positive effects of binding teacher recommendations on the share of ninth grade students attending academic schools. In all specifications, effects are highly significant ($p<0.01$) and amount to roughly 2 percentage points.[31] Panel B of Appendix Figure 4 additionally depicts the evolution of reform effects over time by applying a non-parametric event-study specification. Again, the depicted pre-reform patterns are mostly in line with the common trend assumption as coefficients on the most recent lead dummies (back to 10 years prior reform implementation) are economically and statistically insignificant. In contrast, the cohorts exposed to binding teacher recommendations experience an increase in ninth-grade academic school attendance and the corresponding increase seems to gradually phase in over time.

In sum, I find that positive effects of binding teacher recommendations indeed dominate in the medium run. This also suggests that selective tracking may not only lead to short-term improvements in educational outcomes but may be also beneficial for the students' future educational path.

[29] In addition, Appendix Figure A3 depicts corresponding event-study estimates for academic school enrollment in grade five (panel A) and nine (Panel B). See also section 5 for a corresponding discussion of Appendix Figure A3.

[30] Later transitioning to lower track school is rather common in Germany. In 2000, 13 percent of all students have switched to lower track schools by grade nine. Interestingly, the vast majority (70 percent) of those 'downgraders' has done so during grades seven to nine (i.e., after they have spend some years on the higher track schools)—which is consistent with the depicted reform effect patterns on academic school attendance. In contrast, 'upgrading' to higher track schools is very uncommon as only 1.5 percent have ever done so by grade nine.

[31] Attentive readers may notice that I lose a few observations when including economic controls in column (3). This is because I observe average hours of instruction in primary school (one of the included control variables) starting in 1992. The few dropped observations are students who transferred earlier due to multiple grade retention. Importantly, grade retention is not affected by the reforms (results available upon request).

4.2.2 Academic Performance

To complement the above analyses on academic track attendance, I now turn to investigating students' academic performance in ninth grade. The results in Panel B of Table 6 suggest that the implementation of binding teacher recommendations increases reading achievement among ninth graders. Effects are statistically indistinguishable from zero, but estimates are still sizable (2 to 5 percent of a standard deviation) and represent about 50 to 75 percent of the achievement gains found for primary school students. Panel C of Table 6 reports reform effects on students' grade point average (GPA) in German and math, standardized to have mean zero and standard deviation 1. Overall, GPA among students who require a respective teacher recommendation for academic school attendance seems to be roughly 4 percent of a standard deviation better than among students with free choice of school tracks (not always statistically significantly different from zero at conventional levels).

I conclude that binding recommendations lead to persistent improvements in students' educational outcomes in the medium run.

4.3 Heterogeneities by Socioeconomic Background

This section explores heterogeneities in reform effects by students' socioeconomic background. While the results so far suggest that binding teacher recommendations increase overall educational outcomes, it remains an empirical question whether reforms differently affect students from different family backgrounds. On the one hand, binding teacher recommendations may particularly favor advantaged students. Because parents with a high socioeconomic status (SES) have better resources (e.g., money, knowledge or time) to support the skill development of their children, high SES students may respond more strongly to the incentives provided by binding recommendations (see also Bach and Fischer, 2020). Similarly, high SES parents could push their children harder to receive a recommendation for the academic track because they aspire them to obtain a high socioeconomic status, as well (e.g., Osikominu et al., 2021). In the medium run, efficiency gains, arising from the changes in the allocation of students to school tracks, may be particularly pronounced among advantaged students since high SES parents overrule the outcome of the recommendation more frequently when there is free parental choice.

On the other hand, binding teacher recommendations may particularly favor disadvantaged students. Since they generally perform worse and invest less into skill development at baseline, any incentives arising from binding recommendations have more room to effectively improve academic achievement among low SES students. For the same reason, disadvantaged students can subsequently profit more from spillover effects throughout secondary school.

To analyze the empirical relevance of the opposing effects, I investigate whether reform effects differ along the following dimensions: whether the student has less than 100 books at home, whether the parents' occupational status according to Ganzeboom et al. (1992) is below the median, whether parents have a university (entrance) degree,[32] and whether the student is a first or second generation migrant. For each characteristic, I extend equation 1 to include a full interaction between the respective socioeconomic variable and the reform indicator.

The estimates for the fourth and ninth grade samples are presented in Table A4. As I can only estimate effects for the sub-sample of students with available information on their family background, findings regarding the heterogeneity should be interpreted with caution.[33] By and large, the table depicts homogeneous effects across student subgroups, confirming contemporaneous findings by Bach and Fischer (2020).[34] The coefficients on the interaction term between reform indicator and socioeconomic characteristic is statistically distinguishable from zero in three out of 20 cases. Reform effects on reading and math achievement in fourth grade are larger for students with less than 100 books at home and academic school attendance in ninth grade is smaller for children whose parents have a lower occupational status.

5 Robustness

This section challenges the robustness of the main results. It first discusses the validity of the identifying assumption. Then, it confirms the comparability of tests items used in PISA and NAS. Finally, it evaluates the properties of the differences-in-differences estimator by performing the diagnostic test proposed by de Chaisemartin and D'Haultfœuille (2020).

5.1 The Validity of the Identifying Assumption

The differences-in-differences model identifies the effect of binding teacher recommendations on students' educational outcomes from policy changes within states and over time. Accordingly, a causal interpretation of the estimated effects relies on the assumption that in the absence of

[32] Because data on parental education was not consistently collected in IGLU-E, PISA-E and NAS, I investigate heterogeneitities by whether parents have a university degree in the fourth grade sample and by whether parents have a university entrance degree (*Abitur*) in the ninth grade sample.

[33] While test-taking is mandatory for all students, providing further information on the family background is voluntary. Appendix Table A5 depicts differences in educational outcomes between students with and without available information on their socioeconomic background. Students with missing values (who are by definition excluded from the heterogeneity analyses) perform significantly worse across all depicted dimensions than students with non-missing values. Additional analyses also reveal that binding teacher recommendation are associated with a lower probability to provide information on the number of books at home and migration background in the fourth grade sample (results available upon request). A a result, the interpretation of heterogeneities with respect to these two characteristics are even less clear.

[34] Classifying students with missing information as students with a low socioeconomic background and re-running the heterogeneity analyses leads to similar conclusions (results available upon request).

reforms the changes in outcomes in states that reformed the bindingness of teacher recommendations would have been similar to the changes in outcomes in states that did not reform. Although the event-study graphs presented in section 4 depict only few pre-treatment data points and are, therefore, somewhat limited in their ability of paint a complete picture, they still provide first suggestive evidence in favor of the parallel trend assumption. In particular, there are generally small and insignificant differences in pre-trend outcomes between reforming and non-reforming states.

To further evaluate the validity of the identifying assumption, I next scrutinize how the share of students attending academic schools in grade five and nine evolves over time, using data from the Statistical Offices.[35] As the estimates in Panel A of Appendix Figure A3 show, there is no significant difference in the pre-trends between reforming and non-reforming states in fifth grade academic school enrollment. In fact, the effects move down and up without showing a clear pre-trend pattern. The same is true for academic track attendance in ninth grade. Panel B of Appendix Figure A3 depicts insignificant effects of binding teacher recommendation before the actual reform implementation. While I lack the statistical power to show significant reform effects for each of the depicted periods separately, the p-values from F-tests for joint significance of pre-reform effects amount to 0.364 and 0.300 respectively, thus confirming the visual impression.

Though reassuring, parallel trends cannot rule out the occurrence of contemporaneous policy changes. Any state-specific variation over time that (i) is correlated with the timing structure of the bindingess reforms and (ii) contemporaneously affects students' outcomes can still pose a thread to identification. Hence, the main specification also includes a variety of reform indicators that account for major educational reforms conducted during the observation period as well as a rich set of school and economic input factors. If the main results are driven by systematic differences in these inputs, the estimated coefficients on binding teacher recommendation should approach zero when further including them as controls. However, finding that reform effects remain stable throughout all specifications strongly supports the robustness of my findings (see section 4 for details).

Given that reforms to binding recommendations are mostly implemented by conservative governments (see Appendix Table A2), a remaining major concern is that conservative governments undertake other unobserved policy actions which coincide with the reform under study. To get a first impression for the relevance of potential government effects, I regress government ideology at the time of transition to secondary school on student outcomes in a differences-in-

[35] Academic school attendance—which is observed every year—is arguably more salient to policy-makers than student achievement which is observed only every 3–5 years. Hence, if reforms are implemented endogenously, it may be likely that the reforms are implemented in response to changes in observed academic track attendance rather than in response to students' academic performance.

differences setting. Appendix Tables A6 and A7 show that conservative prime ministers and ministers of education are associated with better educational outcomes of students' in fourth and ninth grade. The positive performance effects remain partly robust when including the full set of controls (see columns (2) and (4)). Next, I account for government ideology in the main differences-in-differences specification. Appendix Table A8 depicts the reform effects on students' educational outcomes in fourth and ninth grade, when controlling for the ideology of the prime minster (see column(1)) or the ideology of the education minister (see column(2)). Reassuringly, the estimated coefficients are of similar magnitude to those presented in section 4.

Another concern is strategic retention. To mitigate potential consequences of not meeting the requirements for the academic recommendation, students may either enroll late or strategically repeat grades. Both would lead to students being older at the time of test-taking. Panel A of Appendix Table A9 depicts respective reform effects. I find significant negative effects of about one month on age at test taking in fourth grade. The effects likely reflect recent attempts to lower school starting age.[36] For ninth grade students, I find small and insignificant effects on age at test taking. Thus—if anything—age effects run contrary to reported achievements gains from binding teacher recommendations, implying that strategic retention does not pose a major threat to identification.

5.2 Test Comparability

For the analyses on reading achievement, I combine data from various studies that assess students in Germany. The respective analyses produce meaningful results only if test scores are comparable across test and waves. By exploiting a specific feature of NAS 2012, I can shed some light on the comparability of PISA and NAS test scores. All PISA participants who attended ninth grade in 2012 were automatically sampled for NAS. For the sub-sample of students who participated in both tests, I can therefore merge additional information on student achievement as assessed by PISA (21 percent of the overall 2012 NAS sample)[37].

First, I compare the math achievement of students who participated in both studies in 2012. To the extent that the findings for math scores are informative for reading scores, I can explore whether the two tests produce similar results. I find a strong correlation (ρ=0.82) between the math scores from both studies. In addition, Appendix Figure A4 plots the distribution

[36] Between 2005 and 2010, several state governments (e.g., North Rhine-Westphalia, Baden-Wuerttemberg, Brandenburg, Berlin) changed the cutoff date for school enrollment. Students affected by the new cutoff dates are part of the fourth grade samples, but not yet of the ninth grade sample which also explains zero age effects for ninth graders.

[37] see Prenzel et al. (2015) for the PISA 2012 data

of within-student differences in test scores. Reassuringly, differences appear to be normally distributed around zero.

Second, the merged data set allows me to observe reading achievement for selected students in 2012, namely those who also participated in PISA. Appendix Table A10 reports reform effects on students' reading achievement, exploiting the merged data set. Reassuringly, the effects do not change substantially when the reading test scores of the 2012 PISA participants are included.

5.3 de Chaisemartin and D'Haultfoeuille Diagnostics

My empirical strategy falls into the category of two-way fixed effects differences-in-differences estimations whose estimates are a weighted sum of the average treatment effect in each state and year. When the treatment effect is constant across states and over time, the depicted regressions estimate the effect of binding teacher recommendations under the standard *common trends* assumption. However, the weighted sum may contain negative weights which is a problem when the average treatment effects (ATEs) are heterogeneous across states or years (see de Chaisemartin and D'Haultfœuille, 2020).

To evaluate the extent of negative weights in my setting, I perform the diagnostic test by de Chaisemartin and D'Haultfœuille (2020) on the simple model, which controls for state and year fixed effects. Appendix Table A11 shows the results on all student-level outcomes.[38] Overall, I obtain 22–84 ATTs with negative weights attached to 14–41 percent of them. Further exploring the causes of the negative weights, I conduct the same analyses without the 'always treated' students in the states Bavaria, Thuringia, and Saxony. Results show that in the fourth grade sample, 0 out of 10 ATTs receive a negative weight, in the ninth grade sample only 3 out of 35–36 ATTs (see Appendix Table A12). Reassuringly, after excluding the 'always treated' from the main specification, results remain robust (see Appendix Table A13). Although I loose some power due to decreased sample sizes, point estimates are similar in magnitude to those presented in section 4.

6 Conclusion

This paper studies how changes in the selectivity of school tracking influence students' educational outcomes in the short and medium run. To that end, I exploit state-level reforms in whether children require a respective recommendation from their primary school teacher to

[38] Since the analyses on math achievement in fourth grade exploit changes in test scores measured in 2011 and 2016 (two time periods only), negative weights cannot bias the results in the math specification. Consequently, Table A11 is performed on the remaining outcomes.

attend the academic schools (the highest track) or whether children and parents have free choice of secondary schools.

In the short run, I find that binding teacher recommendations have a substantial impact on academic achievement of students in primary school *before* track assignment. Conditional on state and school year fixed effects and a rich set of sociodemographic controls, reading (math) achievement is 5.6 (11.6) percent of a standard deviation higher for students who require a respective teacher recommendation for academic school attendance than for students who have free choice of secondary schools. Further analyses show that these achievement gains likely reflect increased time investments in the students' skill development undertaken by children and their parents.

Subsequently, I show that binding teacher recommendations also affect students' educational outcomes in secondary schools. Under binding teacher recommendations, students are slightly less likely to attend academic schools in fifth grade (not statistically significant at conventional levels). However, these small and negative effects which are measured directly after tracking has taken place, veil important positive effects in the medium run 2.0 percentage points higher under binding recommendations. Therefore, binding recommendations seem to reduce the incidence of students transferring to lower tracks during their secondary school career. As academic schools are the most crucial school type where students can obtain a university entrance degree and this degree is positively associated with many economic outcomes (e.g., Gärtner, 2002; Schmillen and Stüber, 2014; Hausner et al., 2015; Dodin et al., 2021), binding teacher recommendations can bear important consequences for students' labor market success.

The political debate in Germany around teacher recommendations has mostly revolved around the normative argument that broad groups in the population should be granted access to the academic schools. Consequently, the most recent reforms have abolished binding recommendations and guaranteed children and parents free choice of secondary schools. My findings, however, challenge this line of argumentation by providing evidence that free choice actually *reduces* academic school attendance in the medium run and can, therefore, harm students' academic performances.

As many school systems around the world employ some sort of (selective) tracking, the findings also allow for broader conclusions: Many 15 year old students in OECD countries are tracked into different school types and 58 percent of them attend selective schools which consider some sort of prior achievement for admission (see Figure 1 and OECD, 2013 for details). In this context, the results described in this paper present direct evidence that the selectivity of the tracking regimes plays an important role for human capital formation.

While the tracking literature has mostly focused on the effects of earlier vs. later tracking so far (e.g., Meghir and Palme, 2005; Hanushek and Woessmann, 2006; Malamud and Pop-

Eleches, 2011; Pekkala Kerr et al., 2013; Piopiunik, 2014; Borghans et al., 2020; Canaan, 2020; Matthewes, 2020), this paper argues that changes in the selectivity of tracking can influence students' educational performances as well. The findings particularly highlight that selective tracking does not only affect outcomes of students in the medium to long run, but they also induce behavioral changes among students, parents and teachers before track assignment takes place. Since the mentioned papers have mainly observed long-term educational outcomes and are thus limited in their ability to isolate short-term effects (see Bach and Fischer, 2020, as one exemption), investigating more systematically whether and how other aspects of the tracking procedure, e.g. the timing of tracking or the number of offered tracks, induce short-term responses is an interesting avenue for further research.

References

Abdulkadiroğlu, Atila, Joshua Angrist, and Parag Pathak (2014). "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools". *Econometrica* 82(1), 137–196.

Andrietti, Vincenzo and Xuejuan Su (2018). "The Impact of Schooling Intensity on Student Learning: Evidence from a Quasi-Experiment". *Education Finance and Policy* 14(4), 679–701.

Angrist, Joshua and Victor Lavy (2009). "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial". *American Economic Review* 99(4), 1384–1414.

Athey, Susan and Guido Imbens (2018). *Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption*. Cambridge, MA.

Bach, Maximilian and Mira Fischer (2020). "Understanding the Response to High-Stakes Incentives in Primary Education". *IZA Discussion Paper Series* 13845.

Barrow, Lisa and Cecilia Elena Rouse (2016). "Financial Incentives and Educational Investment: The Impact of Performance-based Scholarships on Student Time Use". *Education Finance and Policy* 13(4), 419–448.

Baumert, Jürgen, Cordula Artelt, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, and Manfred Weiß, eds. (2002). *PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

– eds. (2003). *PISA 2000 — Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*. Opladen: Leske + Budrich.

– (2009). *Programme for International Student Assessment 2000 (PISA 2000) (Version 1) [Data set]*. Berlin: IQB - Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2000_v1.

Baumert, Jürgen, Cordula Artelt, Ekhard Klieme, and Petra Stanat (2001a). *PISA -Programme for International Student Assessment: Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren. In: Leistungsmessungen in Schulen (285–310)*. Weinheim: Beltz.

Baumert, Jürgen, Werner Blum, Rainer Lehmann, Detlev Leutner, and Michael Neubrand (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland: Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.

Baumert, Jürgen, Cornelia Gresch, Nele McElvany, and Kai Maaz (2010). *Der Übergang von der Grundschule in die weiterführende Schule - Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten: Zusammenfassung der zentralen Befunde*. Berlin: Bundesministerium für bildung und Forschung.

Baumert, Jürgen, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Petra Stanat, Klaus-Jürgen Tillmann, and Manfred Weiß, eds. (2001b). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin (2015). "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools". *Journal of Political Economy* 123(2), 325–364.

Benabou, Roland and Jean Tirole (2000). *Self-Confidence and Social Interactions*. Cambridge, MA.

Betts, Julian R. (2011). "The Economics of Tracking in Education". *Handbook of the Economics of Education* (3), 341–381.

Böhme, Katrin, Dirk Richter, Sebastian Weirich, Nicole Haag, Heike Wendt, Wilfried Bos, Hans Pant, and Petra Stanat (2014). "Messen wir dasselbe? Zur Vergleichbarkeit des IQB-Ländervergleichs 2011 mit den internationalen Studien IGLU und TIMSS 2011". *Unterrichtswissenschaft* 42, 342–365.

Borghans, By Lex, Ron Diris, Wendy Smits, and Jannes de Vries (2020). "Should We Sort it Out Later? The Effect of Tracking Age on Long-Run Outcomes". *Economics of Education Review* 75, 101973.

Bos, Wilfied, Irmela Tarelli, Albert Bremerich-Vos, and Knut Schwippert (2012). *Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, Wilfried, Sabine Hornberg, Karl-Heinz Arnold, Gabriele Faust, Lilian Fried, Eva-Maria Lankes, Knut Schwippert, and Renate Valtin (2010). *Internationale Grundschul-Lese-Untersuchung 2006 (IGLU 2006) [Progress in International Reading Literacy Study 2006 (PIRLS 2006)] (Version 1) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_IGLU_2006_v1.

Bos, Wilfried, Eva-Maria Lankes, Manfred Prenzel, Knut Schwippert, Gerd Walther, and Renate Valtin (2007). *Internationale Grundschul-Lese-Untersuchung 2001 (IGLU 2001) [Progress in International Reading Literacy Study 2001 (PIRLS 2001)] (Version 1) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_IGLU_2001_v1.

Breyton, Ricarda (2018). "Grundschulempfehlung kommt zurück". *Die Welt* 2018.

Canaan, Serena (2020). "The Long-Run Effects of Reducing Early School Tracking". *Journal of Public Economics* 187, 104206.

Checchi, Daniele and Luca Flabbi (2013). "Intergenerational Mobility and Schooling Decisions in Germany and Italy: The Impact of Secondary School Tracks". *Rivista di Politica Economica* (3), 7–57.

Cunha, Flavio and James Heckman (2007). "The Technology of Skill Formation". *American Economic Review* 97(2), 31–47.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov (2006). "Interpreting the Evidence on Life Cycle Skill Formation." *Handbook of the Economics of Education* (1), 697–812.

de Chaisemartin, Clément and Xavier D'Haultfœuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". *American Economic Review* 110(9), 2964–2996.

Die Welt (2014). "Schülerfrust auf dem Gymnasium". *Die Welt*.

Dodin, Majed, Sebastian Findeisen, Lukas Henkel, Dominik Sachs, and Paul Schüle (2021). "Social Mobility in Germany". *CESifo Working Papers* 9200.

Dollmann, Jörg (2011). "Verbindliche und unverbindliche Grundschulempfehlungen und soziale Ungleichheiten am ersten Bildungsübergang". *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 63(4), 595–621.

– (2016). "Less Choice, Less Inequality? A Natural Experiment on Social and Ethnic Differences in Educational Decision-Making". *European Sociological Review* 32(2), 203–215.

Dustmann, Christian, Patrick A. Puhani, and Uta Schönberg (2017). "The Long–Term Effects of Early Track Choice". *The Economic Journal* 127(603), 1348–1380.

Fokken, Silke (2020). "Die Abschaffung des Gymnasiums wäre politischer Selbstmord". *Der Spiegel* 2020.

Ganzeboom, Harry, Paul Graaf, and Donald Treiman (1992). "A Standard International Socio-economic Index of Occupational Status". *Social Science Research* 21(1), 1–56.

Gärtner, Karla (2002). "Differentielle Sterblichkeit - Ergebnisse des Lebenserwartungssurveys". *Zeitschrift für Bevölkerungswissenschaft* 2(27), 185–211.

Gass-Bolm, Torsten (2005). *Das Gymnasium 1945 - 1980: Bildungsreform und gesellschaftlicher Wandel in Westdeutschland: Zugl.: Freiburg (Breisgau), Univ., Diss., 2004 u.d.T.: Gass-Bolm, Torsten: Das Ende der Penne*. Vol. 7. Moderne Zeit. Göttingen: Wallstein.

Grewenig, Elisabeth, Philipp Lergetporer, Katharina Werner, Ludger Wößmann, and Larissa Zierow (2020). "COVID-19 and EducationalInequality: How School Closures Affect Low- and High-Achieving Students". *CESifo Working Papers* 8648.

Guyon, Nina, Eric Maurin, and Sandra McNally (2012). "The Effect of Tracking Students by Ability into Different Schools: A Natural Experiment". *Journal of Human Resources* 47(3), 684–721.

Hanushek, Eric A. and Ludger Woessmann (2006). "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries". *The Economic Journal* 116(510), C63–C76.

Hartig, Johannes and Andreas Frey (2012). "Validität des Tests zur Überprüfung des Erreichens der Bildungsstandards in Mathematik: Zusammenhänge mit den bei PISA gemessenen Kompetenzen und Varianz zwischen Schulen und Schulformen". *Diagnostica* (58), 3–14.

Hausner, Karl Heinz, Doris Söhnlein, Brigitte Weber, and Enzo Weber (2015). "Bessere Chancen mit mehr Bildung". *IAB-Kurzbericht* 11.

Helbig, Marcel and Rita Nikolai (2015). *Die Unvergleichbaren: Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949*. Bad Heilbrunn: Verlag Julius Klinkhardt.

Herrlitz, Hans-Georg, Wulf Hopf, Hartmut Titze, and Ernst Cloer (2009). *Deutsche Schulgeschichte von 1800 bis zur Gegenwart: Eine Einführung*. 5., aktualisierte Aufl. Weinheim: Juventa.

Jackson, C. Kirabo (2010). "A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program". *Journal of Human Resources* 45(3), 591–639.

Jähnen, Stefanie and Marcel Helbig (2015). "Der Einfluss schulrechtlicher Reformen auf Bildungsungleichheiten zwischen den deutschen Bundesländern". *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 67(3), 539–571.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz (2007). "Experimental Analysis of Neighborhood Effects". *Econometrica* 75(1), 83–119.

Köller, Olaf, Michel Knigge, and Bernd Tesch (2011). *IQB-Ländervergleich Sprachen 2008/2009 (IQB-LV 2008-9) [IQB National Assessment Study 2008/2009 (IQB-LV 2008-9)] (Version 2) [Data set]*. Berlin: Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2008_v2.

Kultusministerkonferenz (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: Wolters Kluwer.

– (2015). *Übergang von der Grundschule in Schulen des Sekundarbereichs I und Förderung, Beobachtung und Orientierung in den Jahrgangsstufen 5 und 6 (sog. Orientierungsstufe)*. Berlin: Sekretariat der Ständigen Konferenz der Kultusminister der Länderin der Bundesrepublik Deutschland.

Kunter, Mareike, Gundel Schümer, Cordula Artelt, Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Petra Stanat, Klaus-Jürgen Tillmann, and Manfred Weiß (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente (Materialien aus der Bildungsforschung Bd. [vol.] 72)*. Berlin: Max-Planck-Inst. für Bildungsforschung.

Lenski, Anna E., Martin Hecht, Christiane Penk, Felix Milles, Manuel Mezger, Patricia Heitmann, Petra Stanat, and Hans A. Pant (2016). *IQB-Ländervergleich 2012. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.20386/HUB-42547.

Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw (2010). "The Effect of Financial Rewards on Students' Achivement: evidence from a Randomized Experiment". *Journal of the European Economic Association* 8(6), 1243–1265.

Lichtman-Sadot, Shirlee (2016). "Improving Academic Performance through Conditional Benefits: Open/Closed Campus Policies in High School and Student Outcomes". *Economics of Education Review* 54, 95–112.

Lindo, Jason M., Nicholas J. Sanders, and Philip Oreopoulos (2010). "Ability, Gender, and Performance Standards: Evidence from Academic Probation". *American Economic Journal: Applied Economics* 2(2), 95–117.

Malamud, Ofer and Cristian Pop-Eleches (2011). "School Tracking and Access to Higher Education among Disadvantaged Groups". *Journal of Public Economics* 95(11-12), 1538–1549.

Marcus, Jan and Vaishali Zambre (2019). "The Effect of Increasing Education Efficiency on University Enrollment". *Journal of Human Resources* 54(2), 468–502.

Matthewes, Sönke Hendrik (2020). "Better Together? Heterogeneous Effects of Tracking on Student Achievement". *The Economic Journal*.

Meghir, Costas and Mårten Palme (2005). "Educational Reform, Ability, and Family Background". *The American Economic Review* 95(1), 414–424.

Neugebauer, Martin (2010). "Bildungsungleichheit und Grundschulempfehlung beim Übergang auf das Gymnasium: Eine Dekomposition primärer und sekundärer Herkunftseffekte". *Zeitschrift für Soziologie* 39(3), 202–214.

Obergruber, Natalie and Larissa Zierow (2020). "Students' Behavioural Responses to a Fallback Option - Evidence from Introducing Interim Degrees in German Schools". *Economics of Education Review* 75, 101956.

OECD (2013). *PISA 2012 Results: What Makes Schools Successful (Volume IV)*. Paris: OECD.

Osikominu, Aderonke, Gregor Pfeifer, and Kristina Strohmaier (2021). "The Effects of Free Secondary School Track Choice: A Disaggregated Synthetic Control Approach". *IZA Discussion Paper Series* 14033.

Otto, Otto and Arnfrid Schenk (2011). "Am Ende entscheiden die Eltern". *Die Zeit* 2011(51/2011).

Pallais, Amanda (2009). "Taking a Chance on College: Is the Tennessee Education Lottery Scholarship Program a Winner?" *Journal of Human Resources* 44(1), 199–222.

Pant, Hans Anand, Petra Stanat, Martin Hecht, Patricia Heitmann, Malte Jansen, Anna Eva Lenski, Chrisitiane Penk, Claudia Pöhlmann, Alexander Roppelt, Ulrich Schroeders, and Thilo Siegle (2015). *IQB-Ländervergleich Mathematik und Naturwissenschaften 2012 (IQB-LV 2012) [IQB National Assessment Study2012 (IQB-LV 2012] (Version 4) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2012_v4.

Pekkala Kerr, Sari, Tuomas Pekkarinen, and Roope Uusitalo (2013). "School Tracking and Development of Cognitive Skills". *Journal of Labor Economics* 31(3), 577–602.

Pietsch, Marcus, Kathrin Böhme, Alexander Robitzsch, and Tobias Stubbe (2009). *Das Stufenmodell zur Lesekompetenz der länderübergreifenden Bildungsstandards im Vergleich zu IGLU 2006: In: Bildungsstandards Deutsch und Mathematik*. Weinheim: Beltz-Pädagogik.

Piopiunik, Marc (2014). "The Effects of Early Tracking on Student Performance: Evidence from a School Reform in Bavaria". *Economics of Education Review* 42, 12–33.

– (2021). "How Does Reducing the Intensity of Tracking Affect Student Achievement? Evidence from German State Reforms". *CESifo Working Papers* 9214.

Pischke, Jörn–Steffen (2007). "The Impact of Length of the School Year on Student Performance and Earnings: Evidence from the German Short School Years". *The Economic Journal* 117(523), 1216–1242.

Pischke, Jörn-Steffen and Till von Wachter (2008). "Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation". *The Review of Economics and Statistics* 90(3), 592–598.

Prenzel, Manfred, Cordula Artelt, Jürgen Baumert, Werner Blum, Marcus Hammann, Eckhard Klieme, and Reinhard Pekrun (2010). *Programme for International Student Assessment 2006 (PISA 2006) (Version 1) [Data set]*. Berlin: Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2006_v1.

Prenzel, Manfred, Jürgen Baumert, Werner Blum, Rainer Lehmann, Detlev Leutner, Michael Neubrand, Reinhard Pekrun, Hans-Günter Rolff, Jürgen Rost, and Ulrich Schiefele (2007). *Programme for International Student Assessment 2003 (PISA 2003) (Ver-*

sion 1) [Data set]. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2003_v1.

Prenzel, Manfred, Christine Sälzer, Eckhard Klieme, Olaf Köller, Julia Mang, Jörg-Henrik Heine, Anja Schiepe-Tiska, and Katharina Müller (2015). *Programme for International Student Assessment 2012 (PISA 2012) (Version 5) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2012_v5.

Richter, Dirk, Katrin Böhme, Jana Bastian-Wurzel, Hans A. Pant, and Petra Stanat (2014). *IQB Ländervergleich 2011. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.18452/3127.

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb (2019). "Fast and wild: Bootstrap inference in Stata using boottest". *The Stata Journal* 19(1), 4–60.

Roth, Tobias and Manuel Siegert (2015). "Freiheit versus Gleichheit?Der Einfluss der Verbindlichkeit der Übergangsempfehlung auf die soziale Ungleichheit in der Sekundarstufe". *Zeitschrift für Soziologie* 44(2), 118–136.

– (2016). "Does the Selectivity of an Educational System Affect Social Inequality in Educational Attainment? Empirical Findings for the Transition from Primary to Secondary Level in Germany". *European Sociological Review* 32(6), 779–791.

Sachse, Karoline A., Julia Kretschmann, Aleksander Kocaj, Olaf Köller, Michel Knigge, and Bernd Tesch (2012). *IQB-Ländervergleich 2008/2009. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.20386/HUB-42659.

Schenk, Arnfrid (2010). "Elternwille: Fehlanzeige". *Die Zeit* 2010(31/2010).

Schipolowski, Stefan, Johanna Busse, Camilla Rjosk, Nicole Mahler, Benjamin Becker, and Petra Stanat (2019). *IQB-Bildungstrend 2016. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente in den Fächern Deutsch und Mathematik*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen.

Schipolowski, Stefan, Nicole Haag, Felix Milles, Stefanie Pietz, and Petra Stanat (2018a). *IQB-Bildungstrend 2015. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente im Fach Französisch*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.18452/19998.

– (2018b). *IQB-Bildungstrend 2015. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente in den Fächern Deutsch und Englisch*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.18452/19997.

Schmillen, Achim and Heiko Stüber (2014). "Bildung lohnt sich ein Leben lang". *IAB-Kurzbericht* 1.

Schröder, Mathis, Rainer Siegers, and Katharina Spieß (2013). "Familien in Deutschland - FiD". *Schmollers Jahrbuch* (133), 595–606.

Stanat, Petra, Katrin Böhme, Stefan Schipolowski, Nicole Haag, Sebastian Weirich, Karoline A. Sachse, Lars Hoffmann, and Felicitas Federlein (2018). *IQB-Bildungstrend Sprachen 2015 (IQB-BT 2015) [IQB Trends in Student Achievement 2015 (IQB-BT 2015)] (Version 5) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_BT_2015_v5.

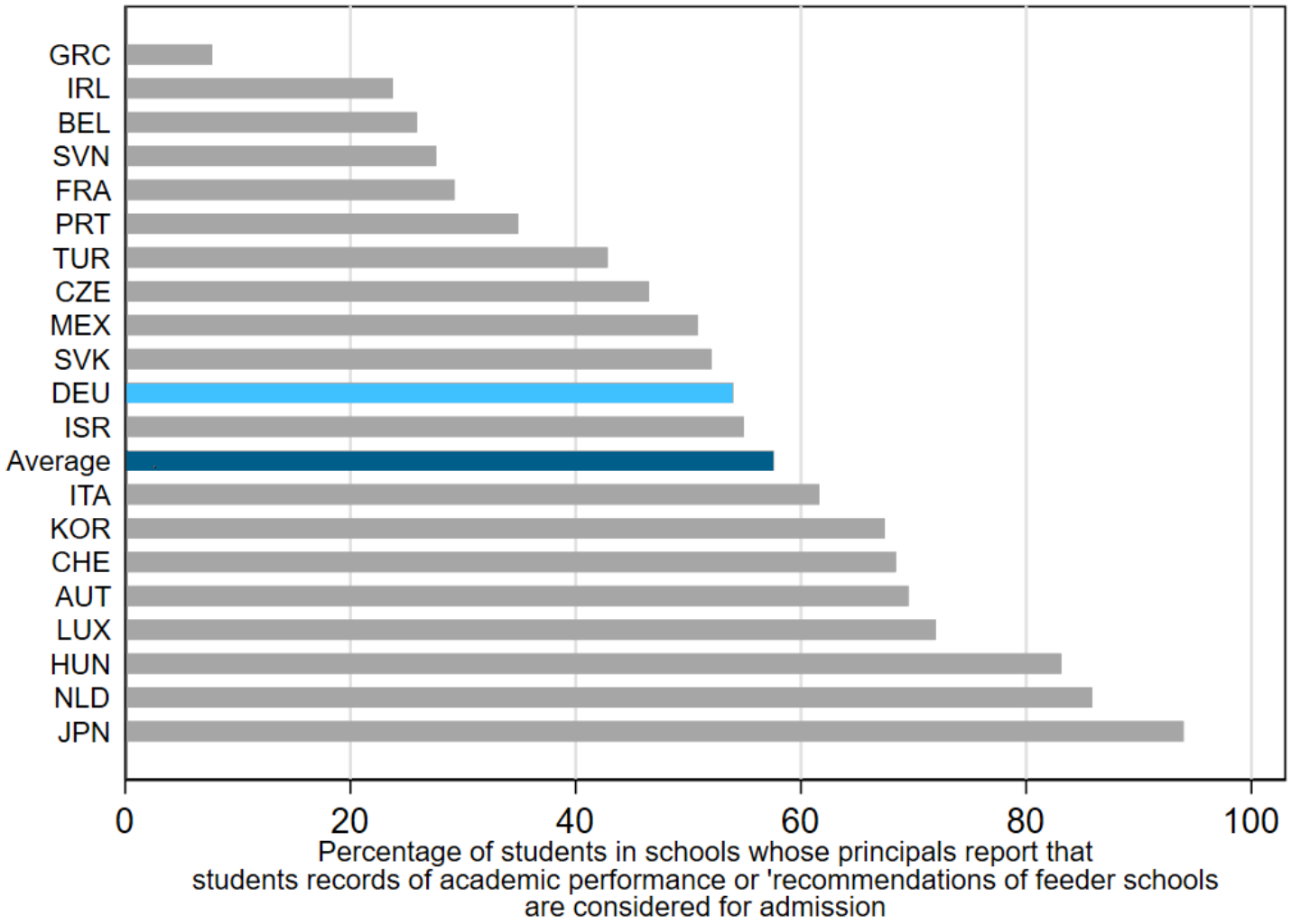
Stanat, Petra, Hans Anand Pant, Katrin Böhme, Dirk Richter, Sebastian Weirich, Nicole Haag, Alexander Roppelt, Maria Engelbert, and Heino Reimers (2014). *IQB-Ländervergleich Primarstufe 2011 (IQB-LV 2011) [IQB National Assessment Study 2011 (IQB-LV 2011)] (Version 3) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2011_v3.

Stanat, Petra, Stefan Schipolowski, Camilla Rjosk, Sebastian Weirich, Nicole Mahler, Pauline Kohrt, and Julia Wittig (2019). *IQB-Bildungstrend Primarstufe 2016 (IQB-BT 2016) [IQB Trends in Student Achievement 2016 (IQB-BT 2016)] (Version 1) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_BT_2016_v1.

Statistisches Bundesamt (1991-2016). *Allgemeinbildende Schulen. Fachserie 11 Reihe 1.* Wiesbaden: Statistisches Bundesamt.

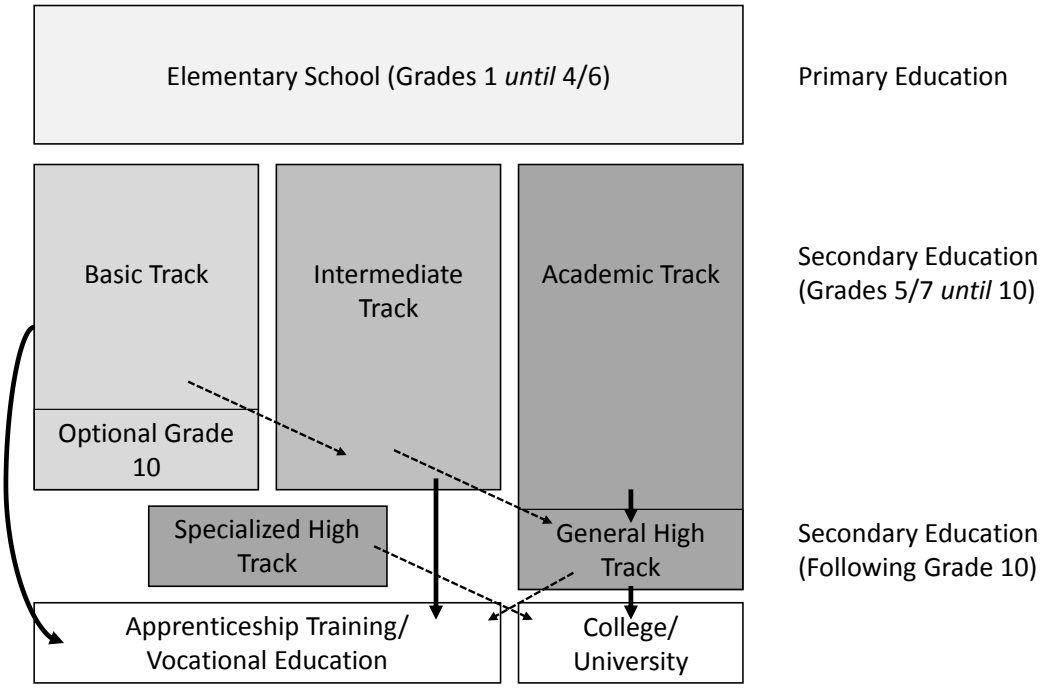
Timmermans, A. C., H. de Boer, H. T. A. Amsing, and M. P. C. van der Werf (2018). "Track Recommendation Bias: Gender, Migration Background and SES Bias over a 20-year Period in the Dutch Context". *British Educational Research Journal* 44(5), 847–874.

Figure 1: Selective Tracking across OECD Countries



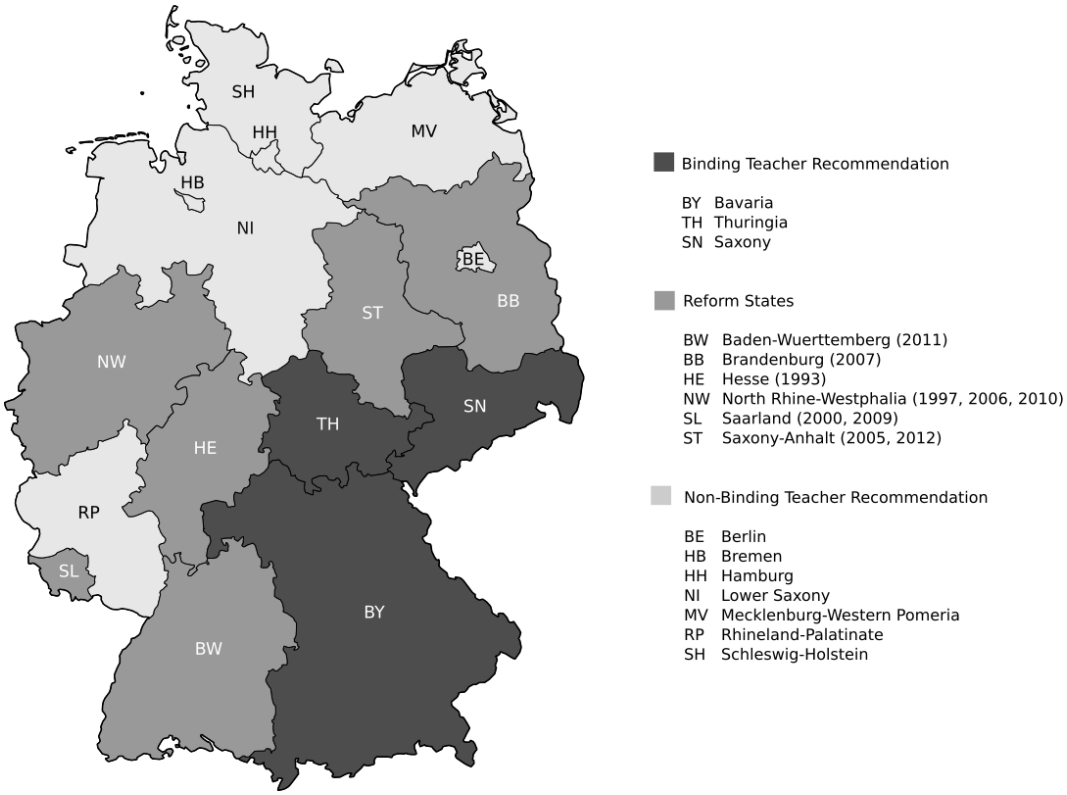
Notes: The figure gives an overview of the selectivity of tracking systems across OECD countries. Sample: OECD countries with more than one education programme available to students at age 15 following OECD (2013) which covers 53 percent of OECD students. Data source: Program for International Student Assessment (PISA-I) 2012.

Figure 2: The German School System



Notes: The figure gives an overview of the school system in Germany. After elementary school which takes 4 years (only in a few states 6 years), students are tracked into three different school types: the basic and intermediate track last to grades 9 and 10, respectively, and prepare students for apprentice-ship training or other sorts of vocational education. The academic track ends with grade 13 (or 12) and leads to the university entrance qualification. Later track switching is possible, enabling graduates from the basic and intermediate track to continue on the next higher track, respectively, and/or obtaining their university entrance qualification via the specialized high track.

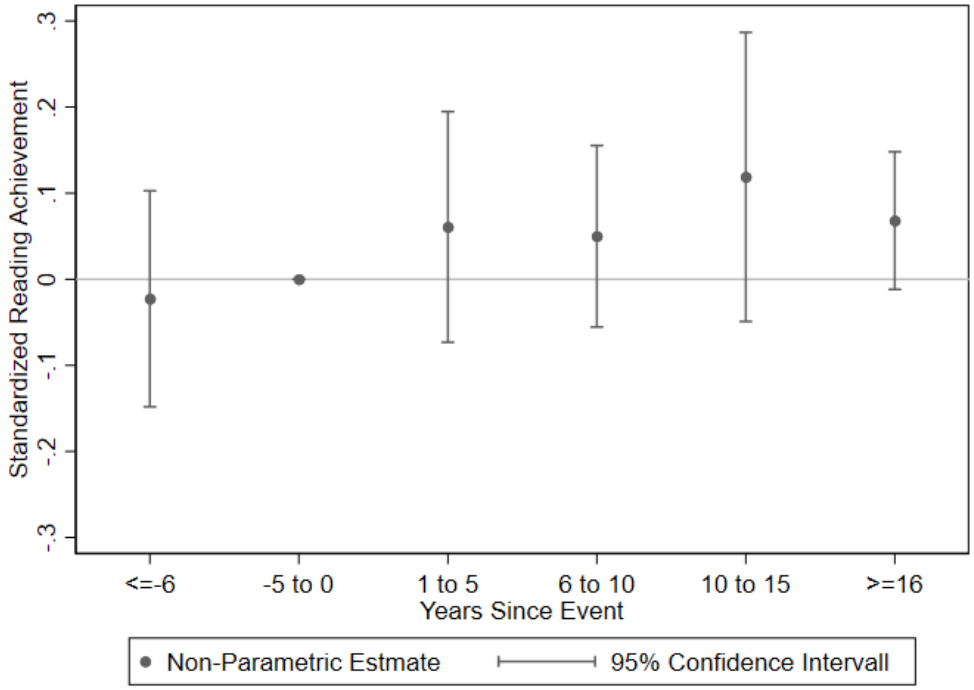
Figure 3: State Variation in the Bindingness of Teacher Recommendations



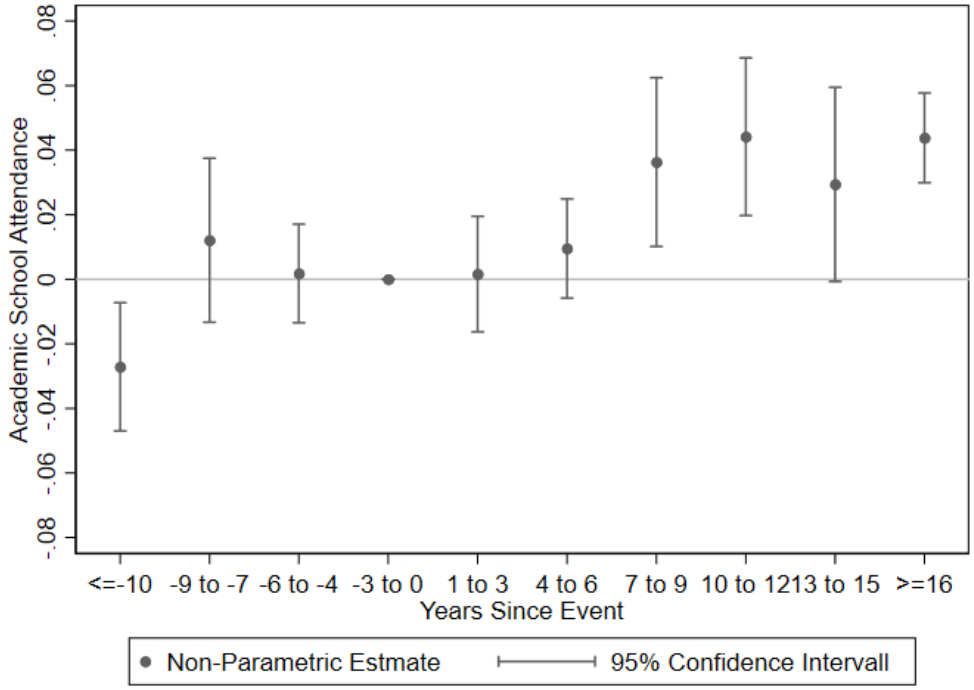
Notes: The figure gives an overview of the variation in the bindingness of teacher recommendations in Germany (for the period 1990 until 2017).

Figure 4: Non-Parametric Event-Study Estimates on Students' Academic Performance

Panel A: Standardized Reading Achievement in Fourth Grade



Panel B: Academic School Attendance in Ninth Grade



Notes: Coefficients from non-parametric event-study regressions and their 95 percent confidence intervals weighted by students' sampling probability, including state and school year of transition fixed effects. Dependent variables: (Panel A) Standardized reading achievement in fourth grade. (Panel B) Academic school attendance in ninth grade. Inference: standard clustering at state level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table 1: Descriptive Statistics: Fourth Grade Students

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Reform Indicator					
Binding Recommendation	0.299	0.458	0	1	76,886
Outcome Variables					
Reading Score (Raw)	500.165	99.050	-36.91487	852.1359	70,615
Math Score (Raw)	481.125	108.163	-94.592	869.943	60,941
Student and School Characteristics					
Age (in Months)	124.262	8.021	73	177	73,217
Female	0.490	0.500	0	1	72,820
More than 100 Books at Home	0.598	0.490	0	1	66,175
Highest Occupational Status (ISEI) of Parents	50.389	18.054	10	90	53,335
First or Second Generation Migrant	0.255	0.436	0	1	64,675
Parents with University Degree	0.299	0.458	0	1	49,848
School located in Urban Area	0.839	0.367	0	1	71,554
Public School	0.968	0.177	0	1	72,341
Reform Controls					
Recommendation Only Based on Students' Grades	0.481	0.500	0	1	76,886
Four Years of Primary School	0.775	0.418	0	1	76,886
Comprehensive School Besides Academic School	0.587	0.492	0	1	76,886
Academic School Takes Eight Years	0.725	0.447	0	1	76,886
Economic Controls					
GDP per Capita (in 1000 Euros)	32.930	9.731	16.323	61.045	76,886
Average School Spending per Capita	1044	340	562	1934	76,886
Share of Students Enrolling late into Primary School	0.074	0.047	0.008	0.215	76,886
Number of Classes in Primary School	9042	8697	995	34,237	76,886
Average Class Size in Primary School	21.031	1.504	16.968	24.607	76,886
Average Number of Lessons in Primary School	31.017	3.469	24.241	43.148	76,886
Share of Full-Time Employed Primary School Teachers	0.491	0.138	0.060	0.726	76,886

Notes: Descriptive statistics (mean, standard deviation, minimum and maximum) for treatment, outcome and control variables. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assesment Study (NAS) 2011 and 2016.

Table 2: Descriptive Statistics: Ninth Grade Students

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Reform Indicator					
Binding Recommendation	0.338	0.473	0	1	212,706
Outcome Variables					
Academic School Attendance	0.328	0.469	0	1	212,706
Reading Score (Raw)	493.594	96.298	8.484	976.180	164,261
German Grade	3.039	0.886	1	6	201,905
Math Grade	3.17	1.045	1	6	201,387
Student Characteristics					
Age (in Months)	187.489	6.813	139	256	212,412
Female	0.493	0.500	0	1	212,124
More than 100 Books at Home	0.486	0.998	0	1	175,913
Highest Occupational Status (ISEI) of Parents	49.833	17.394	11.01	90	172,980
First or Second Generation Migrant	0.245	0.430	0	1	199,054
Parents with University Entrance Degree	0.414	0.492	0	1	169,687
Reform Controls					
Recommendation Only Based on Students' Grades	0.488	0.500	0	1	212,706
Four Years of Primary School	0.782	0.413	0	1	212,706
Intensity of School Tracking	0.289	0.453	0	1	212,706
Academic School Takes Eight Years	0.429	0.495	0	1	212,706
Economic and Education Controls					
GDP per Capita (in 1000 Euros)	26.980	8.767	8.896	53.644	212,706
Share of Students Enrolling Late into Primary School	0.072	0.031	0.006	0.183	212,606
Number of Classes in Primary School	26.980	8.767	8.896	53.644	212,706
Average Class Size in Primary School	21.734	1.693	16.968	25.082	212,706
Average Number of Lessons in Primary School	8.431	2.901	22.782	41.099	212,370
Share of Full-Time Employed Primary School Teachers	0.450	0.157	0.060	0.894	212,706

Notes: Descriptive statistics (mean, standard deviation, minimum and maximum) for treatment, outcome and control variables. Data source: Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table 3: Reform Effects on Reading Achievement among Fourth Grade Students

	(1)	(2)	(3)	(4)
Panel A: Standardized Reading Achievement				
Binding Recommendation	0.064*	0.068*	0.100**	0.056**
	(0.031)	(0.035)	(0.041)	(0.026)
	[0.212]	[0.135]	[0.191]	[0.174]
Control Mean	-0.067	-0.067	-0.067	-0.067
Observations	70,615	70,615	70,615	70,615
R-squared	0.015	0.016	0.016	0.226
Panel B: Standardized Math Achievement				
Binding Recommendation	0.170***	0.145***	0.208***	0.116***
	(0.048)	(0.030)	(0.057)	(0.067)
	[0.195]	[0.083]	[0.160]	[0.141]
Control Mean	-0.084	-0.084	-0.084	-0.084
Observations	60,856	60,856	60,856	60,856
R-squared	0.024	0.028	0.029	0.201
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school (see Appendix Figure A2 for utilized reform variation). Dependent variable: (Panel A) Standardized test scores in reading. (Panel B) Standardized test scores in math. Control mean: mean of the outcome variable for students not subject to binding teacher recommendations. Control variables: reform controls include basis for recommendation, whether recommendation needs to be requested explicitly, duration of primary schooling, intensity of tracking system, and duration of academic track school. Economic controls include GDP, average school spending, average share of students enrolling late into primary school, number of classes in primary school, average class size in primary school, average hours of instruction in primary school, and share of full-time employed primary school teachers. Individual controls include gender, age, migration background,parental education, parental occupation, number of books at home, community location, public school status, wave fixed effects and imputation dummies. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assesment Study (NAS) 2011 and 2016.

Table 4: Reform Effects on Time Invested into Students' Skill Development (at the Age of 10)

	(1)	(2)	(3)	(4)
Panel A: Child's Time Spent with Reading				
Binding Recommendation	0.052	0.048	0.130	0.174**
	(0.087)	(0.065)	(0.075)	(0.068)
	[0.789]	[0.817]	[0.527]	[0.371]
Observations	4,324	4,324	4,324	4,324
R-squared	0.016	0.016	0.017	0.058
Panel B: Child's Time Spent with Conducive Activities				
Binding Recommendation	0.087**	0.092*	0.090*	0.178***
	(0.032)	(0.046)	(0.047)	(0.053)
	[0.154]	[0.198]	[0.161]	[0.129]
Observations	4,215	4,215	4,215	4,215
R-squared	0.021	0.021	0.024	0.118
Panel C: Child's Time Spent with Detrimental Activities				
Binding Recommendation	0.011	0.001	-0.069	-0.085
	(0.156)	(0.150)	(0.178)	(0.165)
	[0.955]	[1.000]	[0.791]	[0.817]
Observations	4,148	4,148	4,148	4,148
R-squared	0.011	0.012	0.014	0.025
Panel D: Child's Time Spent with Family				
Binding Recommendation	0.305***	0.312***	0.370***	0.389***
	(0.054)	(0.060)	(0.059)	(0.057)
	[0.134]	[0.174]	[0.079]	[0.099]
Observations	4,318	4,318	4,318	4,318
R-squared	0.013	0.013	0.014	0.049
Panel E: Parents Frequently Consult Teacher				
Binding Recommendation	0.136**	0.134**	0.135***	0.130***
	(0.053)	(0.053)	(0.042)	(0.040)
	[0.123]	[0.149]	[0.150]	[0.141]
Control Mean	0.434	0.434	0.434	0.434
Observations	4,345	4,345	4,345	4,345
R-squared	0.027	0.027	0.029	0.041
State & School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions with state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school (see Appendix Figure A2 for utilized reform variation). Dependent variables: time invested, the higher the value the more invested. (Panel A) Time spent with reading. (Panel B) Index summarizing child's time spent with making music, dancing or theatre, doing sports, doing technical work, or drawing, following Kling et al. (2007). (Panel C) Index summarizing child's time spent with watching TV, playing on the computer, surfing on the internet or listening to music, following Kling et al. (2007). (Panel D) Child's time spent together with family, standardized. (Panel E) Dummy variable (=1 if parent regularly meets teacher). Control mean: mean of the outcome variable for students not subject to binding teacher recommendations. Control variables: see Table 3 for included reform and economic controls. Individual controls include child's gender, child's age, respondents' gender, whether child is respondent's own child, migration background, parental education, and imputation dummies. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: German Socioeconomic Panel (GSOEP), stacked mother-child questionnaires for 9–10 year old children.

Table 5: Reform Effects on Academic School Attendance

	Share of Students Attending Academic Schools				
	5th Grade (1)	6th Grade (2)	7th Grade (3)	8th Grade (4)	9th Grade (5)
Binding Recommendation	-0.010	-0.003	0.008	0.010*	0.012*
	(0.012)	(0.013)	(0.006)	(0.005)	(0.006)
	[0.470]	[0.859]	[0.227]	[0.152]	[0.144]
Control Mean	0.377	0.359	0.355	0.336	0.324
Observations	309	297	376	360	344
Observations (Federal States)	14	14	16	16	16
R-squared	0.774	0.739	0.855	0.867	0.889
State FEs	Yes	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes	Yes
Reform & Economic Controls	Yes	Yes	Yes	Yes	Yes

Notes: Differences-in-differences regressions with state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school (see Appendix Figure A2 for utilized reform variation). Dependent variables: share of students attending academic schools. Control variables: see Table 3 for included reform and economic controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: German Statistical Office 1991-2016.

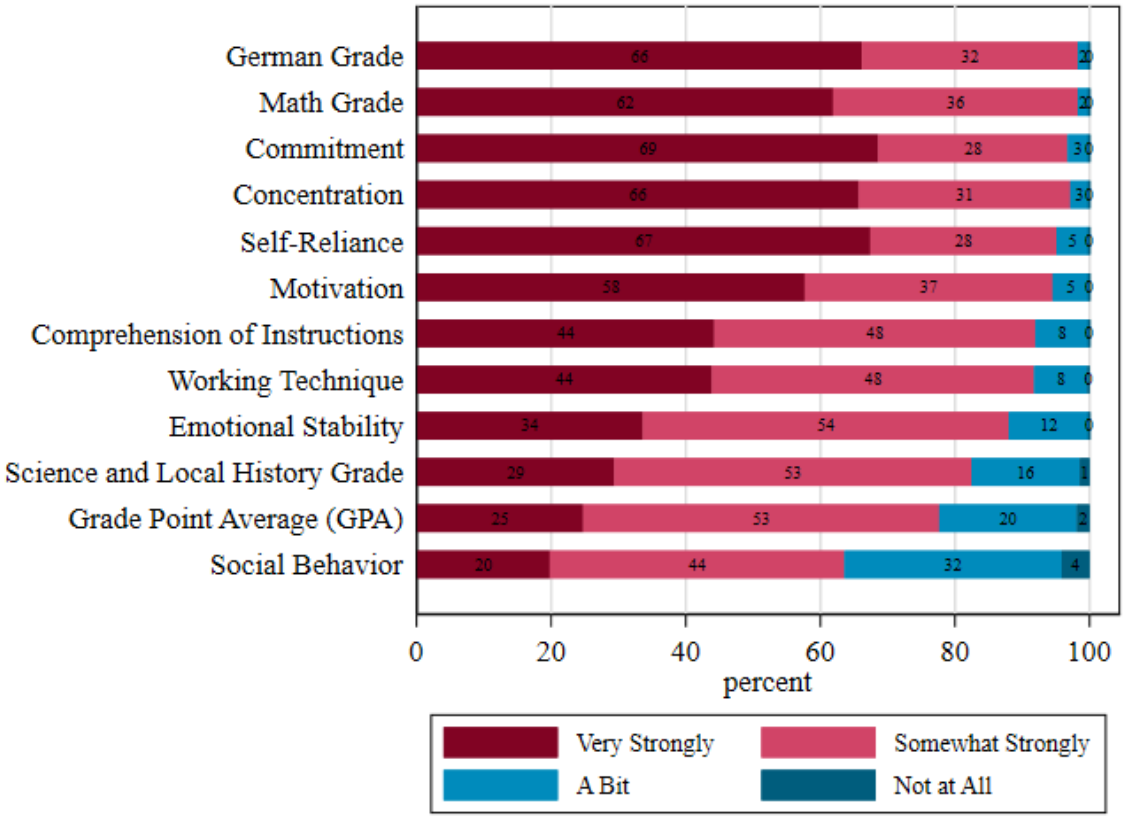
Table 6: Reform Effects on Academic School Attendance and Academic Performance among Ninth Grade Students

	(1)	(2)	(3)	(4)
Panel A: Academic School Attendance				
Binding Recommendation	0.017***	0.025***	0.023***	0.020***
	(0.004)	(0.005)	(0.008)	(0.006)
	[0.130]	[0.088]	[0.239]	[0.099]
Control Mean	0.335	0.335	0.335	0.335
Observations	208,405	208,405	207,969	207,969
R-squared	0.014	0.015	0.014	0.229
Panel B: Standardized Reading Achievement				
Binding Recommendation	0.041	0.055	0.046	0.019
	(0.033)	(0.043)	(0.036)	(0.025)
	[0.283]	[0.316]	[0.591]	[0.520]
Control Mean	0.009	0.009	0.009	0.009
Observations	163,346	163,346	162,940	162,940
R-squared	0.014	0.014	0.014	0.273
Panel C: Grade Point Average				
Binding Recommendation	0.041	0.049*	0.050*	0.041
	(0.025)	(0.023)	(0.025)	(0.024)
	[0.478]	[0.238]	[0.235]	[0.248]
Control Mean	3.186	3.186	3.186	3.186
Observations	197,252	197,252	196,846	196,846
R-squared	0.046	0.047	0.047	0.101
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school (see Appendix Figure A2 for utilized reform variation). Dependent variables. (Panel A) Academic school attendance. (Panel B) Standardized test score in reading. (Panel C) Standardized grade point average, the higher the value the better the GPA. Control mean: mean of the outcome variable for students not subject to binding teacher recommendations. Control variables: see Table 3 for included reform and economic controls. Individual controls include gender, age, migration background, parental education, parental occupation, number of books at home, wave fixed effects, and imputation dummies. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assesment Study (NAS) 2009, 2012, and 2015.

Appendix A: Appendix Figures and Tables

Figure A1: The Basis for Teacher Recommendations



Notes: Response to the following question: *'Thinking about the different students: How strongly do the following factors determine your recommendation?'* Sample: fourth grade teachers in German. Weighted responses. Source: IGLU 2006.

Figure A2: Reforms on the Bindingness of Teacher Recommendations and Used Datasets

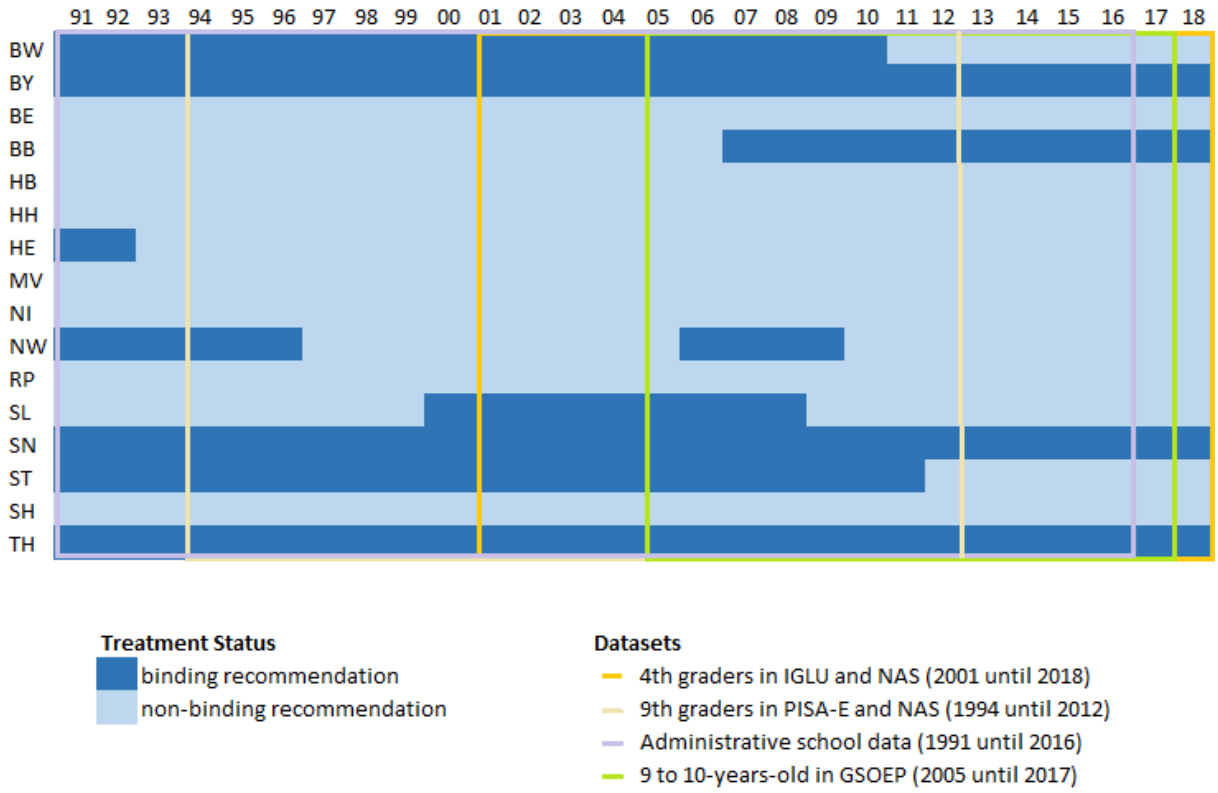
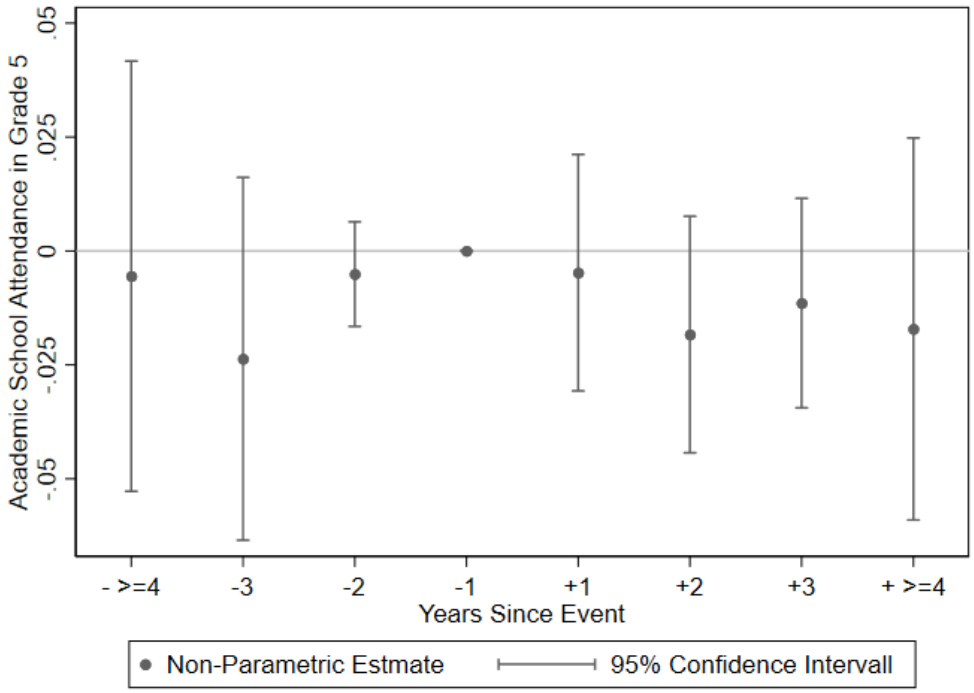
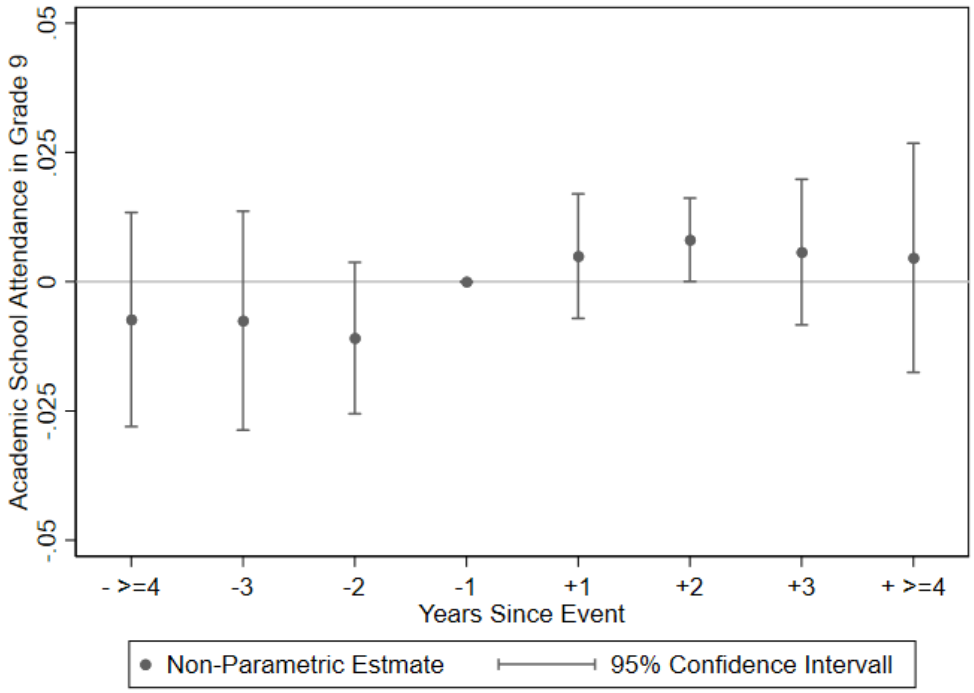


Figure A3: Non-Parametric Event-Study Estimates on Academic School Attendance

Panel B: Academic School Attendance in Fifth Grade

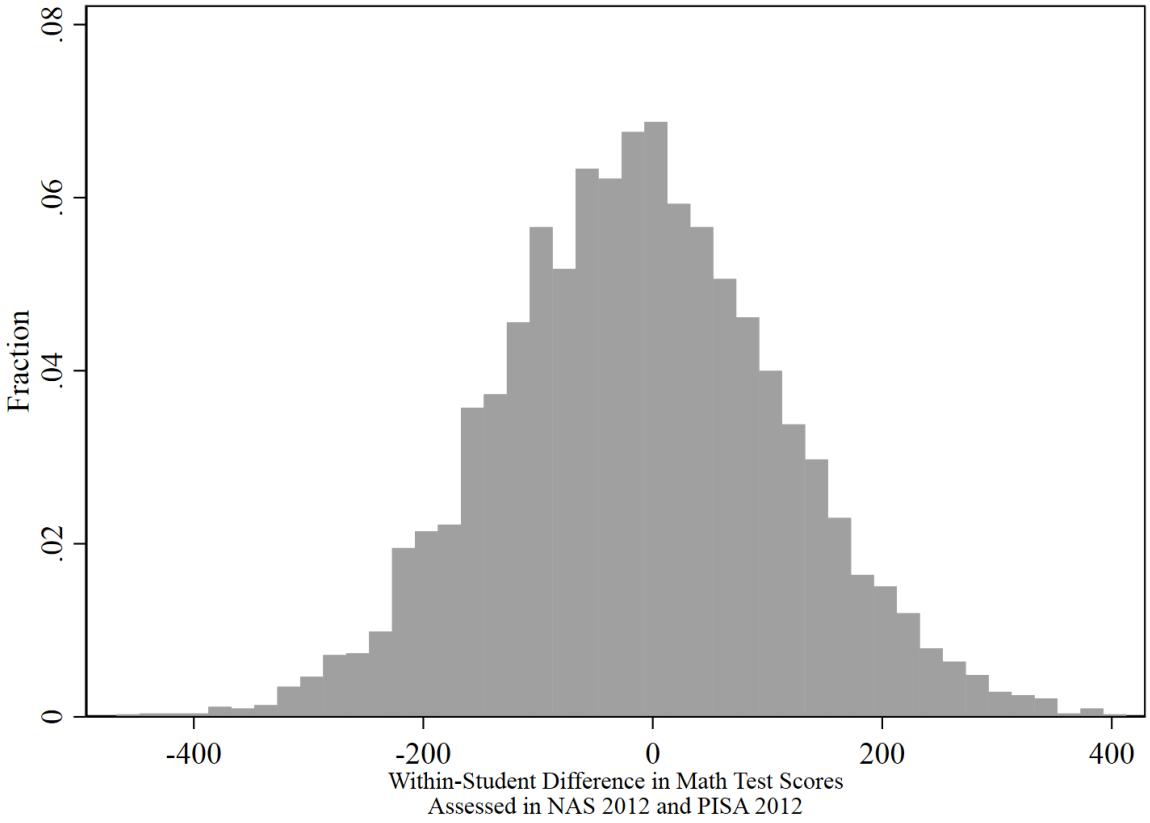


Panel B: Academic School Attendance in Ninth Grade



Notes: Coefficients from non-parametric event-study regressions and their 95 percent confidence intervals. Dependent variables: share of students attending academic schools. Control variables: see table 5 for included reform and economic controls. Inference: standard clustering at state level. The p-values from F-tests testing for joint significance of pre-trends [post-trends] are 0.364 [0.658] (Panel A) and 0.300 [0.09] (Panel B), respectively. Data source: German Statistical Office.

Figure A4: Distribution of Math Test Scores in NAS and PISA



Notes: In each of the two tests, achievement is mapped on a scale with mean of 500 and a standard deviation of 100 test-score points. Outcome: within-student difference between math scores assessed by NAS 2012 and PISA 2012 ('0' indicates no difference in test scores). Sample: students who participated in NAS 2012 and PISA 2012.

Table A1: Who Attends Academic Schools? Descriptive Evidence.

	Child Attends Academic School	
	(1)	
Child is Female	0.073***	(0.008)
Child's Age	-0.004	(0.005)
Child Lives in West Germany	-0.114***	(0.011)
Child Lives with Both Parents	0.054***	(0.010)
Mother's Age	0.008***	(0.001)
Mother with University Entrance Degree	0.254***	(0.010)
Mother Works Full-Time	-0.003	(0.010)
Mother Does Not Work	-0.083***	(0.009)
Father's Age	-0.000	(0.001)
Father with University Entrance Degree	0.207***	(0.011)
Father Works Full-time	0.061***	(0.022)
Father Does Not Work	-0.028	(0.026)
Constant	-0.056	(0.081)
Observations	13,145	
R-squared	0.189	

Notes: Ordinary least square regressions. Dependent variables. (1) Child attends academic school (*Gymnasium*). Background characteristics of parents are imputed to the median when corresponding information is missing (mainly due to children not living together with parents in one household). Inference: robust standard errors in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Sample: Students aged between 14 and 16 years. Data source: German Microcensus 2015.

Table A2: Reforms on the Bindingness of Teacher Recommendations and Ruling Parties, by State

State	Year of Reform	Reform Type: Recommendat. Change to	Ruling Parties in Legislation Period	
			Before the Reform	of the Reform
Hesse	1993	non-binding	CDU/FDP (1987-1991)	SPD/Gruene (1991-1995)
North Rhine-Westphalia	1997	non-binding	SPD /Gruene (1990-1995)	SPD/Gruene (1995-2000)
Saarland	2000	binding	SPD (1994-1999)	CDU (1999-2004)
Saxony-Anhalt	2005	binding	SPD (1998-2002)	CDU/FDP (2002-2006)
North Rhine-Westphalia	2006	binding	SPD/Gruene (2000-2005)	CDU/FDP (2005-2010)
Brandenburg	2007	binding	SPD/CDU (1999-2004)	SPD/CDU (2004-2009)
Saarland	2009	non-binding	CDU (2004-2009)	CDU/FDP/SPD/Gruene (2009-2012)
North Rhine-Westphalia	2010	non-binding	CDU/FDP (2005-2010)	SPD/Gruene (2010-2012)
Baden-Wuerttemberg	2011	non-binding	CDU/FDP/DVP (2006-2011)	Gruene/SPD (2011-2016)
Saxony-Anhalt	2012	non-binding	CDU/SPD (2006-2011)	CDU/SPD (2011-2016)

Table A3: Reform Effects on Issued Recommendations for Academic Schools

	Share of Students with Recommendation for Academic Schools		
	(1)	(2)	(3)
Binding Recommendation	0.060	0.064	0.051
	(0.040)	(0.050)	(0.044)
	[0.287]	[0.453]	[0.348]
Control Mean	0.374	0.374	0.374
Observations	108	108	108
Observations (Federal States)	8	8	8
R-squared	0.741	0.741	0.813
State FEs	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes
Reform Controls	No	Yes	Yes
Economic Controls	No	No	Yes

Notes: Differences-in differences regressions with state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variable: recommendations issued for the academic school as share of all recommendations issued. Control variables: see Table 3 for included reform and economic controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Sample: Baden-Wuerttemberg, Bavaria, Berlin, Mecklenburg Western Pomerania, North Rhine Westphalia, Rhineland Palatinate, Saxony. Data source: various State Ministries of Education, Germany.

Table A4: Heterogeneous Reform Effects by Socioeconomic Background

	Fourth Grade Students		Ninth Grade Students		
	Read. Achiev. (1)	Math Achiev. (2)	Acad. School (3)	Read. Achiev. (4)	GPA (5)
Panel A: Books at Home					
Binding Recommendation	-0.015	0.076	0.017**	0.049*	0.034
	(0.037)	(0.055)	(0.006)	(0.025)	(0.026)
Bind. Recomm. x Less than 100 Books	0.059**	0.097***	-0.003	-0.033	0.029
	(0.022)	(0.028)	(0.006)	(0.027)	(0.038)
Obseravtions	63,772	55,083	173,969	129,259	165,439
Panel B: Highest Occupational Status (ISEI) of Parents					
Binding Recommendation	0.024	0.079	0.035***	0.037	0.014
	(0.032)	(0.055)	(0.005)	(0.027)	(0.024)
Bind. Recomm. x ISEI below Median	0.008	0.046	-0.015**	-0.008	0.055
	(0.028)	(0.031)	(0.006)	(0.024)	(0.046)
Observations	51,654	44,149	171,014	145,717	160,219
Panel C: Parental Education					
Binding Recommendation	0.017	0.142**	0.050***	0.012	0.001
	(0.035)	(0.063)	(0.013)	(0.034)	(0.031)
Bind. Recomm. x Par. w\o Uni Degree	0.012	0.015	-0.034*	0.021	0.052
	(0.023)	(0.032)	(0.017)	(0.019)	(0.050)
Observations	48,372	41,409	168,188	137,097	160,219
Panel D: Migration Status (First or Second Generation Migrant)					
Binding Recommendation	0.029	0.076**	0.023***	0.037	0.044*
	(0.028)	(0.030)	(0.006)	(0.024)	(0.025)
Bind. Recomm.n x Migrant	-0.012	0.017	0.005	-0.015	-0.042*
	(0.044)	(0.032)	(0.008)	(0.028)	(0.021)
Observations	62,352	53,868	197,065	152,354	187,452
State FEs	Yes	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes	Yes
Reform & Economic & Individual Controls	Yes	Yes	Yes	Yes	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables. (1) Standardized test scores in reading, fourth grade. (2) Academic school attendance, ninth grade. (3) Standardized test scores in reading, ninth grade. (4) Standardized grade point average, the higher the value the better the GPA, ninth grade. Control variables: see Tables 3 and 6 for included reform, education and individual controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A5: Academic Performance and Missing Information on Socioeconomic Background

	Fourth Grade Students		Ninth Grade Students		
	Read. Achiev. (1)	Math Achiev. (2)	Acad. School (3)	Read. Achiev. (4)	GPA (5)
Panel A: Books at Home					
Missing: Books at home	-0.702***	-0.515***	-0.097***	-0.405***	-0.054***
	(0.028)	(0.023)	(0.006)	(0.014)	(0.013)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Panel B: Highest ISEI					
Missing: HISEI	-0.476***	-0.444***	-0.187***	-0.663***	-0.264***
	(0.016)	(0.015)	(0.004)	(0.013)	(0.009)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Panel C: Parental Education					
Missing: Parental Education	-0.352***	-0.330***	-0.140***	-0.456***	-0.188***
	(0.015)	(0.014)	(0.004)	(0.011)	(0.009)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Panel D: Migration Status					
Missing: Migration Status	-0.634***	-0.500***	-0.175***	-0.653***	-0.246***
	(0.024)	(0.022)	(0.005)	(0.017)	(0.015)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Study FEs	Yes	Yes	Yes	Yes	Yes

Notes: Ordinary least square regressions weighted by students' sampling probability, including wave fixed effects. Dependent variables. (1) Standardized test scores in reading, fourth grade. (2) Standardized test scores in math, fourth grade. (3) Academic school attendance, ninth grade. (4) Standardized test scores in reading, ninth grade. (5) Standardized grade point average, the higher the value the better the GPA, ninth grade. Inference: robust standard errors in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A6: Effects of Government Ideology on Students' Academic Performance in Fourth Grade

	(1)	(2)	(3)	(4)
Panel A: Standardized Reading Achievement				
Conservative Prime Minister	0.058	0.027		
	(0.036)	(0.033)		
	[0.224]	[0.548]		
Conservative Education Minister			0.058*	0.017
			(0.028)	(0.027)
			[0.069]	[0.660]
Observations	70,615	70,615	70,615	70,615
R-squared	0.015	0.226	0.015	0.226
Panel B: Standardized Math Achievement				
Conservative Prime Minister	0.133**	-0.021		
	(0.061)	(0.043)		
	[0.169]	[0.724]		
Conservative Education Minister			0.132**	-0.011
			(0.055)	(0.039)
			[0.079]	[0.823]
Observations	60,856	60,856	60,856	60,856
R-squared	0.024	0.200	0.024	0.200
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform & Economic & Individual Controls	No	Yes	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Dependent variables: (Panel A) Standardized test scores in reading. (Panel B) Standardized test scores in math. Control variables: see Table 3 for included reform, education and individual controls. Inference: Standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016.

Table A7: Effects of Government Ideology on Students' Academic Performance in Ninth Grade

	(1)	(2)	(3)	(4)
Panel A: Actual Academic School Attendance				
Conservative Prime Minister	0.006	0.010		
	(0.008)	(0.007)		
	[0.487]	[0.240]		
Conservative Education Minister			0.003	0.005
			(0.006)	(0.008)
			[0.740]	[0.639]
Observations	208,405	207,969	208,405	207,969
R-squared	0.014	0.229	0.014	0.229
Panel B: Standardized Reading Achievement				
Conservative Prime Minister	0.058**	0.030		
	(0.027)	(0.043)		
	[0.095]	[0.618]		
Conservative Education Minister			0.033	-0.004
			(0.029)	(0.040)
			[0.307]	[0.938]
Observations	163,346	162,940	163,346	162,940
R-squared	0.014	0.273	0.014	0.273
Panel C: Grade Point Average				
Conservative Prime Minister	0.062***	0.045		
	(0.018)	(0.028)		
	[0.036]	[0.329]		
Conservative Education Minister			0.069***	0.047
			(0.015)	(0.031)
			[0.002]	[0.345]
Observations	197,252	196,846	197,252	196,846
R-squared	0.047	0.101	0.047	0.101
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform & Economic & Individual Controls	No	Yes	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Dependent variables: (Panel A) Academic school attendance. (Panel B) Standardized test scores in reading. (Panel C) Standardized grade point average, the higher the value the better the GPA. Control variables: see Table 6 for included reform, education and individual controls. Inference: Standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: PProgram for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A8: Reform Effects on Students' Academic Performance, Controlling for Government Ideology

	(1)	(2)
Panel A: Standardized Reading Achievement (Fourth Grade)		
Binding Recommendation	0.052*	0.047
	(0.028)	(0.030)
	[0.230]	[0.264]
Observations	70,615	70,615
R-squared	0.226	0.226
Panel B: Standardized Math Achievement (Fourth Grade)		
Binding Recommendation	0.147**	0.179**
	(0.050)	(0.063)
	[0.211]	[0.172]
Observations	60,856	60,856
R-squared	0.201	0.201
Panel C: Actual Academic School Attendance (Ninth Grade)		
Binding Recommendation	0.019***	0.022***
	(0.006)	(0.006)
	[0.110]	[0.075]
Observations	207,969	207,969
R-squared	0.229	0.229
Panel D: Reading Achievement (Ninth Grade)		
Binding Recommendation	0.016	0.027
	(0.029)	(0.029)
	[0.609]	[0.424]
Observations	162,940	162,940
R-squared	0.273	0.273
Panel E: Grade Point Average (Ninth Grade)		
Binding Recommendation	0.026	0.026
	(0.022)	(0.024)
	[0.354]	[0.437]
Observations	196,846	196,846
R-squared	0.101	0.101
State & School Year FEs	Yes	Yes
Reform & Economic & Individual Controls	Yes	Yes
Conservative Prime Minister	Yes	No
Conservative Education Minister	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables. (Panel A) Standardized test scores in reading, fourth grade. Panel B) Standardized test scores in math, fourth grade. (Panel C) Academic school attendance, ninth grade. (Panel D) Standardized test scores in reading, ninth grade. (Panel E) Standardized grade point average. Control variables: see Tables 3 and 6 for included reform, education and individual controls. Inference: Standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A9: Reform Effects on Students' Age

	Age in Months			
	(1)	(2)	(3)	(4)
Panel A: Fourth Grade Students				
Binding Recommendation	-0.352	-0.531	-0.779**	-0.774**
	(0.395)	(0.394)	(0.342)	(0.318)
	[0.458]	[0.480]	[0.131]	[0.194]
Control Mean	124.833	124.833	124.833	124.833
Observations	68,135	68,135	68,135	68,135
R-squared	0.102	0.102	0.104	0.144
Panel B: Ninth Grade Students				
Binding Recommendation	0.188	0.159	0.155	0.335
	(0.182)	(0.174)	(0.279)	(0.275)
	[0.391]	[0.387]	[0.753]	[0.480]
Control Mean	187.680	187.680	187.680	187.680
Observations	208,247	208,247	207,812	207,812
R-squared	0.220	0.221	0.219	0.270
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables: age in months. Control variables: see Tables 3 and 6 for included reform, education and individual controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A10: Reform Effects on Ninth Grade Reading Achievement (Including PISA 2012 Scores)

	Reading Achievement in Ninth Grade			
	(1)	(2)	(3)	(4)
Binding Recommendation	0.041	0.068	0.027	0.024
	(0.037)	(0.041)	(0.036)	(0.027)
	[0.274]	[0.170]	[0.540]	[0.437]
Control Mean	-0.055	-0.055	-0.055	-0.055
Observations	172,686	172,686	172,279	172,279
R-squared	0.016	0.017	0.017	0.227
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables: standardized test scores in reading, ninth grade. Control variables: see Table 6 for included reform, education and indivudal controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Program for International Student Assessment (PISA-E) 2000, 2003, 2006 and 2012 and National Assessment Study (NAS) 2009 and 2015.

Table A11: de Chaisemartin and D'Haultfœuille (2020): Differences-in-Differences Diagnostics

	# ATTs	# of ATTs with Negative Weight	Sum of Negative Weights
	(1)	(2)	(3)
Panel A: Reading Achievement (Fourth Grade)			
	22	9	-0.218
Panel B: Actual Academic School Attendance (Ninth Grade)			
	84	12	-0.153
Panel C: Reading Achievement (Ninth Grade)			
	65	16	-0.109
Panel D: Grade Point Average (Ninth Grade)			
	84	12	-0.160

Notes: Results from de Chaisemartin and D'Haultfœuille (2020) diagnostics test implemented using STATA *twowayfeweights* command. Estimated weights of all group-period clusters in the basic model, controlling for state and school-year fixed effects.

Table A12: de Chaisemartin and D'Haultfœuille (2020): Differences-in-Differences Diagnostics without 'Always-Treated'

	# ATTs	# of ATTs with negative weight	Sum of negative weights
	(1)	(2)	(3)
Panel A: Reading Achievement (Fourth Grade)			
	10	0	–
Panel B: Actual Academic School Attendance (Ninth Grade)			
	36	3	-0.050
Panel C: Reading Achievement (Ninth Grade)			
	35	3	-0.064
Panel D: Grade Point Average (Ninth Grade)			
	36	3	-0.053

Notes: Results from de Chaisemartin and D'Haultfœuille (2020) diagnostics test implemented using STATA *twowayfeweights* command. Estimated weights of all group-period clusters in the simple model, controlling for state and school-year fixed effects. Sample, excluding the states of Bavaria, Thuringia and Saxony.

Table A13: Reform Effects on Students' Academic Performance without 'Always-Treated'

	Fourth Grade Students	Ninth Grade Students		
	Read. Achievement	Acad. School	Achievem.	GPA
	(1)	(2)	(3)	(4)
Binding Recommendation	0.058	0.013**	0.035	0.022
	(0.033)	(0.005)	(0.021)	(0.026)
	[0.392]	[0.121]	[0.090]	[0.540]
Observations	58,206	171,036	133,760	161,346
R-squared	0.232	0.231	0.278	0.105
State Fes	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform & Economic & Indivdual Controls	Yes	Yes	Yes	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables. (1) Standardized test scores in reading, fourth grade. (2) Academic school attendance, ninth grade. (3) Standardized test scores in reading, ninth grade. (4) Standardized grade point average, the higher the value the better the GPA, ninth grade. Control variables: see Tables 3 and 6 for included reform, education and individual controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015—excluding the states of Bavaria, Thuringia and Saxony.