Bach, Maximilian; Fischer, Mira

Working Paper

Understanding the Response to High-Stakes Incentives in Primary Education

IZA Discussion Papers, No. 13845

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Bach, Maximilian; Fischer, Mira (2020) : Understanding the Response to High-Stakes Incentives in Primary Education, IZA Discussion Papers, No. 13845, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/227372

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13845

Understanding the Response to High-Stakes Incentives in Primary Education

Maximilian Bach
Mira Fischer

NOVEMBER 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13845

Understanding the Response to High-Stakes Incentives in Primary Education

Maximilian Bach
ZEW Mannheim

Mira Fischer
WZB Berlin Social Science Center and IZA

NOVEMBER 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Understanding the Response to High-Stakes Incentives in Primary Education*

This paper studies responses to high-stakes incentives arising from early ability tracking. We use three complementary research designs exploiting differences in school track admission rules at the end of primary school in Germany's early ability tracking system. Our results show that the need to perform well to qualify for a better track raises students' math, reading, listening, and orthography skills in grade 4, the final grade before students are sorted into tracks. Evidence from self-reported behavior suggests that these effects are driven by greater study effort but not parental responses. However, we also observe that stronger incentives decrease student well-being and intrinsic motivation to study.

JEL Classification: I20, I28, I29
Keywords: student effort, tracking, incentives

Corresponding author:
Maximilian Bach
ZEW Mannheim
L 7, 1
68161 Mannheim
Germany
E-mail: maximilian.bach@zew.de

* Maximilian Bach acknowledges financial support from the Leibniz-Association through the Leibniz Competition 2019 project "Improving School Admissions for Diversity and Better Learning Outcomes". Mira Fischer acknowledges financial support by the German Science Foundation through CRC TRR 190 (project number 280092119). We benefited from helpful discussions with Matthias Parey, Friedhelm Pfeiffer, and Sönke Matthewes. This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort (SC) Kindergarten, doi:10.5157/NEPS:SC2:7.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the BMBF. As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. Data for the IQB National Assessment Study were made available by the Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB). The study also uses data from National Orientation Tests in grade 3 (VERA-3). Kathrin Edrich provided excellent research assistance.

1 Introduction

To what extent is educational achievement shaped by student incentives? Several studies have documented that students' investments into education at the secondary and post-secondary level strongly respond to various incentives, such as the returns to education (e.g., Jensen, 2010; Abramitzky and Lavy, 2014). However, salient and high-powered incentives often exist already in primary school, for example in the form of performance requirements for admission to selective schools or promotion to the next grade. Little is known about whether and to what extent these incentives affect educational investments and early skill acquisition.

This paper offers the first causal evidence on the implications of high-stakes incentives for primary school students. Specifically, we examine how primary school students respond to incentives arising from early ability tracking—the allocation of students to different secondary school tracks based on previous achievement. We address the following questions: Does the need to perform well to qualify for a better secondary school track affect the skill development of young children in primary school? What are the roles of study effort and parental responses? And, are there any downsides to high-stakes incentives, such as potentially harmful effects for student well-being or intrinsic motivation?

In order to shed light on these questions, we exploit the context of the German school system, a rigorous early tracking system where children are allocated to different hierarchically ordered school tracks at the end of primary school, as young as age 10. This setting is uniquely suited to analyze how children respond to incentives and whether this translates into achievement gains for two reasons: First, we will provide evidence that children view allocation to a secondary school track as important and exhibit strong preferences for higher tracks. Second, the specific rules governing the allocation to school tracks differ across German federal states and have undergone sharp changes in the past decade. While in some states students receive binding track recommendations at the end of primary school based on prior achievement, other states allow parents to freely choose a school track regardless of the recommended track. Granting parents free choice

effectively reduces incentives for children to perform well in primary school in order to be allowed to attend a higher track. This gives rise to between- and within-state variation in strong performance incentives for primary school children, which we can study.

We use the variation in performance incentives in three complementary research designs drawing on various data sources. The first is a difference-in-differences (DiD) design that exploits the fact that two states have repealed binding track recommendations in 2012. Specifically, we contrast student performance at the end of primary school in grade 4 (the final grade before tracking) in states that repealed binding track recommendations with states that did not change their rules, before and after the repeal. For this approach, we use student-level data from the National Assessment Study (NAS), which administered standardized math, reading, and listening tests to state-representative samples of around 25,000 grade 4 students covering all German federal states in 2011 and 2016.

In a second research design, we leverage the fact that student performance is not graded until grade 3 and track recommendations are only based on performance in grade 4. As a result, we expect the performance incentives from binding track recommendations to be less salient in grades below grade 4. To estimate effects, we thus compare test score gains between grades 2 and 4 in states with and without binding track recommendations. While this approach only makes use of between-state variation in rules, flexibly controlling for lower grade test scores allows us to account for potential confounding arising from differences in the student population and schooling environment between states. Importantly, it turns out that estimates remain virtually unchanged without lower grade test score controls as long as class-level averages of a limited set of students' family background characteristics (e.g., parental education and migration background) are included to control for differences in the socio-economic composition of classes. This strongly supports the assumption that our between-state comparison of test score gains is not biased by omitted variables correlated with free track choice.[1] The data for the second approach come from starting cohort 2 of the German National Educational Panel Study (NEPS). This is a student-level panel, following a nationally representative sample of students who

[1]We also show that, conditional on pre-determined family background controls, there are no test score differences in grades 1 and 2 between states with and without free track choice.

entered primary school in 2012 over time.

In a third approach, we combine the DiD and lagged valued-added approach to address any potentially remaining biases in the previous designs. For a subset of states, we have state-level data on the distribution of proficiency levels (measured in 5 categories) in math and reading in grade 3 for the same cohorts that were tested in grade 4 in the NAS. This allows for a difference-in-difference-in-differences (DiDiD) approach, which identifies effects from within-cohort changes in the distribution of proficiency levels between grades 3 and 4, across pre- and post-repeal cohorts, between repeal and no-repeal states.

We find across all research designs that tracking incentives have a substantial impact on academic achievement of primary school students. Estimates from the DiD and lagged value-added designs based on student-level data indicate that free track choice lowers student test scores in grade 4 by between 0.10-0.14 standard deviations in math, 0.06-0.08 in reading, 0.09 in listening, and 0.2 in orthography. With the DiDiD design, based on state-level data, we find that free track choice consistently lowers the share of students scoring at higher proficiency levels. For example, the share of students scoring at or above the 3rd of 5 proficiency level decreases in response to free track choice from grade 3 to 4 relative to unaffected cohorts by 7.59 percentage points in math and 6.45 percentage points in reading. Since none of the standardized tests we use are incentivized — the test scores do not count towards students' grades — we are confident that our estimates capture effects on students' acquired skills, rather than differences in test-taking behaviour (see, e.g., Jalava et al., 2015; Levitt et al., 2016).[2] Effects can be found throughout the entire achievement distribution, however, low-achieving students show the strongest response, suggesting that tracking incentives for primary school students can reduce gaps in early skill acquisition.

Detailed survey measures of student and parental behaviour in the NEPS data allow us to investigate the mechanisms that help to explain these results. In line with an effort-based explanation, we show that with binding track recommendations 4th graders invest, on average, 13 minutes more per day (a 25% difference) into homework and private study.

[2]Also note that neither the NEPS nor NAS test results are revealed to schools.

Again, low-achieving students show the strongest response. At the same time, there is no evidence that greater study effort is a result of parental behaviour; parents do not report to help students with homework or hire private tutors more frequently with binding track recommendations. Instead, parents enforce stricter rules for studying and monitor their children more closely when track recommendations are *not* binding. This behaviour possibly reflects parental responses to students' lower study effort and achievement when they are not induced to study more by binding track recommendations. We therefore conclude that the achievement effects are likely driven by students' effort responses to a type of high-stake incentive that arises in many education systems.

The achievement gains due to binding track recommendations, however, come at the cost of a reduction in students' self-reported well-being. We provide direct evidence that this results from 4th grade students being more anxious about their school performance and future. Moreover, external incentives appear to crowd-out intrinsic motivation, an effect that persists after grade 4, when students have been allocated to a secondary school track. We thus identify potentially harmful side-effects of high-stakes incentives that need to be traded off against their immediate achievement effects.

The present paper contributes to a growing literature on the role of incentives in education. Studies at the secondary and post-secondary level have demonstrated older students' responsiveness in educational decisions and achievement to a variety of incentives.[3] Yet little is known about whether the same applies to younger students in primary school for whom the benefits of education are less tangible and who are less patient (see Sutter et al., 2019, for a review). Knowledge about the responsiveness of primary school students is particularly important since achievement gaps open up early (see, e.g., Heck-

[3]For example, empirical studies demonstrate achievement effects in response to changes in the returns to education (Abramitzky and Lavy, 2014; Chadi et al., 2019), direct financial incentives for academic performance, such as merit scholarships or financial rewards for passing grades (Henry and Rubenstein, 2002; Kremer et al., 2009; Pallais, 2009; Angrist et al., 2009; Angrist and Lavy, 2009; Leuven et al., 2010; Jackson, 2010a; Behrman et al., 2015; Burgess et al., 2016; Barrow and Rouse, 2018; Montalban, 2019). There are also studies that point to the importance of non-financial incentives, such as differences in the value of leisure (Stinebrickner and Stinebrickner, 2008; Metcalfe et al., 2019), high school exit exams (Jürges et al., 2005), exogenous changes in one's GPA (Hvidman and Sievertsen, 2019), academic probation (Lindo et al., 2010), or high school campus leave policies that are conditional on academic performance (Lichtman-Sadot, 2016). Schildberg-Hörisch and Wagner (2020) review the experimental studies on non-financial incentives.

man, 2006) and earlier investments are typically assumed to be more effective because of the dynamic complementarity of skills (Heckman and Cunha, 2007).

To the best of our knowledge, only two previous studies have investigated achievement effects of incentives for primary school students. Fryer (2011) and Bettinger (2012) both study randomized cash incentives for primary school students in disadvantaged school districts in the US and find mixed results for incentives up to $60. While Bettinger (2012) finds an impact of financial incentives on math but no impact on reading, social science, or science outcomes, Fryer (2011) finds no effect on either math or reading test scores. Importantly, these experimental studies do not allow to discriminate between two potential explanations for the largely null findings, each carrying different policy implications: First, low-stakes or purely financial incentives may be insufficient to motivate young students to exert more effort. Second, primary school students may lack understanding of the educational production function. That is, increased study effort does not necessarily raise achievement due, for example, to young students not knowing how to learn effectively. The variation in high-stakes incentives in our setting – which is not viable in experimental studies due to ethical concerns – allows us to detect effects that are consistent with the first and refute the second explanation for previous null findings. Given strong enough incentives, primary school students exert more effort and greater effort is productive in the sense that it improves achievement. Our findings are particularly remarkable because they show that not only do young children respond to incentives, they are significantly motivated by a reward that lies half a year or more in the future.

Our results also yield an important lesson for the broad literature on the implications of school systems that separate students based on prior achievement. According to PISA 2012, an average 43 percent of 15-year-old students in OECD countries attend schools where previous academic performance is considered for admission OECD (2013).[4] Two

[4]In the majority of OECD countries, separation takes the form of explicit between-school tracking (OECD, 2013), whereas some school systems mostly track students within schools (e.g., those of Canada and the US). Achievement-based separation also exists in the form of selective schools where admission is based on previous academic achievement (e.g., grammar schools in the UK or Boston and New York City's elite exams schools), fee-charging private schools that grant discounts based on students' previous academic achievement, or centralized school assignment mechanisms where priority at over-subscribed

main questions that researchers have asked concerning selective schools are: What is the benefit to the student of attending a selective school (Damon, 2010; Jackson, 2010b; Abdulkadiroğlu et al., 2014; Dobbie and Fryer, 2014; Clark and Del Bono, 2016; Dustmann et al., 2017; Borghans et al., 2019), and what is the benefit of a selective system as a whole on achievement after students have been allocated to different schools (see, e.g., Hanushek and Wössmann, 2006; Atkinson et al., 2006; Guyon et al., 2011; Matthewes, 2020). Our results point to an important effect that has been missed in these analyses, namely that achievement-based selection can improve student achievement *before* selection takes place by raising incentives to study.[5] This effect has to be taken into account when evaluating school systems with different degrees of selectivity. Our results suggest, for example, that even if students do not benefit from attending selective schools (see, e.g., Abdulkadiroğlu et al., 2014; Dustmann et al., 2017), the presence of selective schools alone can enhance skill acquisition if strong preferences for them induces greater study effort.

The paper proceeds as follows: Section 2 describes the institutional context of Germany's school system and Section 3 details the empirical strategies. Section 4 describes the data, Section 5 presents the results, and Section 6 concludes.

schools is partly based on academic achievement (e.g., Hungary, Boston, Chicago and New York City's public schools).

[5]Hanushek and Wössmann (2006) compare achievement gains between the end of primary school (before students have been tracked) and grades 7 and 8 in tracking and non-tracking school systems. Hence, their estimation strategy absorbs any achievement gains of tracking that occur before students are tracked. Matthewes (2020) follows a similar estimation strategy by comparing test score gains after students have been tracked between German federal states with three versus two separate school tracks. Guyon et al. (2011) exploit a reform in Northern Ireland that sharply increased the proportion of students admitted into elite schools after grade 6 from one year to the next. However, by comparing cohorts right before and after implementation of the reform, it is unclear whether these students anticipated the sudden increase in the probability to be selected for elite education. Hence, these students possibly only had limited ability to respond in earlier grades. Another related strand of the literature studies long-run effects (e.g., on educational attainment and wages in adulthood) of several de-tracking school reforms in Europe (see, e.g., Aakvik et al., 2010; Meghir and Palme, 2005; Pekkala Kerr et al., 2013; Canaan, 2020). Since these studies measure outcomes in adulthood and the de-tracking reforms did not abolish tracking but merely delayed it to higher grades in secondary school (and often coincided with other major changes to the school system), it is unclear how informative they are about the incentivizing effect of selective schools.

2 Institutional context

In this section, we will give a concise description of the German school system for the academic years 2010/11-2016/17, with a focus on those aspects that are most relevant for understanding the transition from primary to secondary school.

Figure 1 provides a stylised overview of primary and secondary education in Germany. Primary school covers grades 1 through 4, or in some federal states grades 1 to 6, and assignment is based solely on whether a student lives within a school's catchment area.[6] At the end of primary school, students are allocated to different, vertically ordered school tracks, which educational researchers consider one of the most important events in a person's educational biography (Baumert et al., 2010). The placement in a school track after primary school can be viewed as permanent, as few students move between school tracks (particularly in the upward direction) before they have completed a track.[7] There are two types of tracking systems in Germany. The first is a three-tiered system, with three school types: *Hauptschule*, *Realschule*, and *Gymnasium*. We designate these tracks 'basic', 'intermediate', and 'academic', respectively. Education in the academic track lasts eight to nine years (grades 5-12/13) and prepares students for higher education. Education in the basic and medium track, on the other hand, lasts five (grades 5-9) and six years (grades 5-10), respectively, is less academic and prepares students for an apprenticeship in blue-collar or white-collar occupations. The second system is a two-tiered system where the basic and intermediate track are combined into one comprehensive school. In grades 5 and 6, students in comprehensive schools are not grouped by ability. However, some comprehensive schools form track-specific classes from grade 7 onward and offer the same degrees as the basic, intermediate, and academic track schools.[8]

[Figure 1 about here]

[6] Shure (2019) reports that based on anecdotal evidence from speaking to relevant Ministries of Education in four German states, on average less than one percent of families request that their child attend a primary school that is not the school to which they were assigned. Unfortunately, no official statistics are collected.

[7] Dustmann et al. (2017) show that only 2% of students switch tracks between grades 5-9 based on the School Census for Bavaria and Hesse. However, once students complete one of the lower tracks (i.e., after grade 9 or 10) with good grades, it is quite common to continue in a higher track.

[8] See Matthewes (2020) for more details on comprehensive schools in Germany.

School tracks in both systems are clearly ordered hierarchically, a fact that is also reflected in teacher salaries, which tend to increase with the school track.[9] School tracks also imply differences in social status. According to sociological studies, students assigned to the lowest school track may generally be considered as socially disadvantaged (Bos et al., 2010), they also suffer from social contempt and lack of social recognition (Wellgraf, 2014) and self-stigmatize as losers (Knigge, 2009). Primary school students are aware of the importance of school track assignment and tend to have a strong preference for the academic track in particular. Figure 2 shows how primary school students view the different tracks.[10] More than 60% of 3rd graders expect the job prospects of academic track graduates to be very good, compared to less than 18% for the basic track (Panel A).[11] If these students could freely choose their school degree, 78% state that they would choose the academic degree and only around 6% the basic degree (Panel B).

[Figure 2 about here]

The rules governing which track children can attend after primary school differ across states and have undergone sharp changes in the last decade. However, the basic structure is the same across states. There are no formal exit exams at the end of primary school, rather the primary school issues a secondary school track recommendation for each student which is generally guided by the student's abilities and their performance in the last grade of primary school.[12] The main difference between states is whether this recommendation is binding or not. Whereas in some states students cannot attend a

[9]Teacher salaries differ from state to state, but in all states where teachers are still employed as civil servants (*Beamte*), academic track teachers are paid according to salary grade (*Besoldungsstufe*) A13. Basic track teachers are generally paid according to salary grade A12, but some states further differentiate between intermediate and basic track teachers. To get an sense of the salary differences across school tracks, consider North Rhine-Westphalia, the state which employs the most teachers. In 2017, the gross starting salary for a academic track teacher was 4,038 Euros and 3.459 Euros for an intermediate track teacher, a 16.7% salary differential.

[10]Figure 2 is based on data from starting cohort 2 of the NEPS, which will be introduced in Section 4.

[11]If we pool responses and standardize them to have mean zero and variance one, the average responses by school track are -.38 SD, -.24 SD, and .60 SD for the basic, intermediate, and academic track, respectively.

[12]The factors determining recommendations differ across states. Bavaria and Saxony, for example, have specific GPA cutoffs, while other states do not specify the exact requirements to get a recommendation for a particular school. For more details on these rules see Helbig and Nikolai (2015). For a discussion of the factors teachers consider for their recommendations see Baumert et al. (2010).

higher track than recommended,[13] other states allow parents to freely choose a secondary school track for their child regardless of the recommendation.[14] Table 1 shows which states had binding track recommendations for the school years 2010/11-2016/17. In 7 out of 16 states, the school's recommendation was binding in 2010, however, of these seven states, three changed the rules between 2011 and 2013 and decided to give parents free school track choice.[15] These are North Rhine-Westphalia, Baden-Württemberg, and Saxony-Anhalt.

[Table 1 about here]

3 Empirical approach

3.1 Difference-in-differences (DiD)

We employ three identification strategies to estimate responses to high-stakes incentives, which are tailored to the three available datasets that will be discussed in Section 4. The first identification strategy takes advantage of the repeal of the binding track assignment policy in some states in a difference-in-differences (DiD) framework. Our first difference compares student achievement in states that repealed binding track assignment (repeal states) to states that did not change the school track allocation rules (no-repeal states). The second difference compares cohorts of students before (pre-repeal) and after the binding track assignment were removed (post-repeal). In our baseline empirical exercise, we estimate the following regression:

$$Y_{isc} = \beta_1 TC_{sc} + \alpha_s + \alpha_c + \beta_2 X_{isc} + \epsilon_{isc} \tag{1}$$

where Y_{isc} is an outcome of student i in cohort c in state s, TC_{sc} is an indicator for

[13] In case of conflict between the recommendation and the parents' wishes, some states allow students to take a special test whose outcome determines whether a student is allowed to attend the higher track.

[14] Parents can always opt for a lower track than recommended, also in states with binding track recommendations.

[15] Note that even with free track choice, teachers still have to recommend tracks for each student. Therefore parents receive the same type of information on their children's academic achievement with or without free track choice. This is important as there is evidence that performance information interventions for parents can have positive impacts on children (Barrera-Osorio et al., 2020).

track choice which takes on the value 1 if state s has free track choice and varies at the state-cohort level. To account for differences across states and across time, we include state (α_s) and cohort (α_c) fixed effects.[16] X_{is} is a vector of various school-, class-, and student-level characteristics. These include age at test, a rich set of parental background variables, for example, parental education and occupation, books at home, and migration background. In the most parsimonious specifications, X_{isc} only includes test booklet fixed effects because test items differed across students.[17]

The identifying assumption is that the removal of binding track assignment is orthogonal to potential outcomes of students in repeal and no-repeal states in the post-repeal period. In other words, we assume that if binding track assignment would have stayed in place, student outcomes would have evolved similarly in states that did and did not change their track recommendation policy. A typical concern for the validity of this assumption is self-selection into treatment. However, track assignment policies vary at the federal state level and across-state mobility in Germany is relatively low. Hence, the potential for selection bias due to sorting of students based on track assignment policies is very small in this context. Nevertheless, we do find that repealing binding track assignment is correlated with several student characteristics. For this reason, we also estimate specifications where X_{is} includes a rich set of school-, class-, and student-level characteristics to ensure the comparability of students within states over time, because each cohort comes from a different sample of schools, which might differ due to sampling variability.

Another concern is that the repeal of binding track assignment partly overlapped with other state-level changes, for example, due to other schooling reforms. During our study time period, three major education policies were implemented that could have affected primary student achievement differently across states. These are the expansion of early public childcare and full-day schooling, and the inclusion of special-needs students in

[16]Since both repeal states we consider removed binding track assignment in the same year and there were no changes in the no-repeal states, we do not run into the negative weights problem that can bias two-way fixed effects regressions in the presence of effect heterogeneity (see, e.g., Chaisemartin and D'Haultfoeuille, 2020).

[17]See Section 4 for more details on this.

regular schools (*Inklusion*).[18] The implementation of these reforms was relegated to the federal states, which, in turn, mostly relegated it to the municipalities or individual schools. As a result, the timing and extent of implementation differs widely, not only across but also within states.[19] To account for these policy changes, we include controls for years spent in public childcare, whether a school offers full-day schooling, and the share of special-needs students in class. Controlling for the potential policy effects this way does not change our conclusions.

3.2 Lagged valued-added model

The identification assumption for the DiD approach is that the repeal of binding track assignment was not accompanied by other changes to the education system that affected student achievement in repeal states differently relative to no-repeal states. We provide several pieces of evidence suggesting that the coefficient β_1 in equation (1) indeed measures the effects of free track choice, as opposed to other school policies. However, to further address concerns regarding such potential threats to identification, we also employ an alternative approach that relies on a source of variation less susceptible to bias arising from time-varying schooling policies. Specifically, we leverage within-student variation and compare test score gains between grades 2 and 4 in states with and without free track choice. The basic idea is that track assignment rules become more salient towards the end of primary school since students are not graded until grade 3 and track recommendations are supposed to only be based on grade 4 performance. Hence, assignment rules that govern the transition to secondary school tracks are less likely to affect student behaviour in the lower grades of primary school when students do not yet receive quantitative performance appraisals that make relative abilities salient. By conditioning on grade 2 achievement, we can account for potential confounding arising from differences in the student population or the schooling environment across choice and no-choice states.

[18] Note that other school reforms, such as the introduction of central exit exams in the academic track and the introduction (and subsequent abolition) of university tuition fees, were all implemented at the secondary school or university level. Hence, they are unlikely to affect achievement in primary school. For an overview of these reforms see Helbig and Nikolai (2015) or Marcus and Zambre (2019).

[19] In Appendix C, we provide more details on these policy reforms.

For this approach, we use longitudinal student-level data for the cohort that transitioned to secondary school in 2016. Importantly, the variation here stems from differences in 4th grade achievement between states which had binding track assignment in 2016 (Bavaria, Saxony, and Thuringia) and all states with free track choice in 2016. Free-track-choice states include Baden-Württemberg and Saxony-Anhalt, which repealed the binding track recommendation policy. The variation for this identification strategy is thus complementary to that used in the DiD design, which is mainly based on the comparison of achievement in pre- and post-repeal periods in repeal states. We will estimate the following *lagged valued-added model* that is often used in the economics of education literature:

$$Y_{is4} = \beta_1 TC_s + \lambda Y_{is2} + \beta_2 X_{is} + \epsilon_{is} \tag{2}$$

where Y_{is4} is the outcome variable of interest of student i in state s measured in grade 4, TC_{sc} is an indicator equal to 1 if state s has track choice and 0 otherwise, as in equation (1); Y_{is2} is a vector of lagged achievement measures from grade 2. Finally, X_{is} is a vector of various school-, class-, and student-level characteristics that includes, for example, the degree of urbanization of the school district, student-teacher ratio, age composition of the teaching staff, number of instruction hours in math and German, class size, student composition (e.g., the share of low- and high-SES students), and student-level demographics (parental education, reported books at home, number of siblings, migration background, gender, premature birth, timing of primary school enrolment, learning disabilities).

The key identifying assumption for this approach is that test score gains between grades 2 and 4 in states with free track choice form a valid counterfactual to test score gains in states with binding track assignment. To lend credibility to this assumption, we will show in Section 5 that there are virtually no baseline test score differences in grades 1 and 2 between choice and no-choice states once we condition on a set of pre-determined characteristics, such as the share of high-SES children in a school. Moreover, we do not find evidence for differences in test score gains in grade 2 (i.e., the difference in test scores

between grades 2 and 1) across choice and no-choice states.[20] This strongly suggests that student achievement levels and potential achievement gains in choice and no-choice states are comparable conditional on pre-determined controls for student demographics and the learning environment.[21]

However, even if choice and no-choice states are similar in terms of achievement levels and gains up to grade 2, these states could differ in their capacities to raise achievement in subsequent grades due to differences in the learning environment. To test whether potential differences in schooling inputs across binding and nonbinding states confound our effects, we draw on unusually rich data on the school environment. We have detailed information on the schooling environment, such as the time per week devoted to certain tasks in class, teachers' job satisfaction, their perception of the quality of the school management etc. In the robustness section we check that our results are not sensitive to the inclusion of these variables.[22]

Most of the tests in the data we use for this approach were administered relatively early in the school year and the timing of tests can differ widely within waves. In grade 4, for example, students were tested between November and January.[23] In addition, the beginning of the school year varies across states.[24] For example, in 2015 the earliest state (North Rhine-Westphalia) started school on August 11, while the last state (Bavaria)

[20]As a result, estimates of equation (2) with grade 4 test scores as outcomes are very similar regardless of whether one includes test scores from grade 1, grade 2, or both as controls.

[21]Note that the typical problems identified in the literature with lagged value-added models are less likely to matter in our case (see, e.g., Rothstein, 2010; Andrabi et al., 2011). One major concern is that the error term could include the ability to learn faster. Such unobserved heterogeneity in learning dynamics could result in a common individual-level component in the error term ϵ_{it}. However, our treatment varies at the state level where selection based on unobserved ability to learn is unlikely to play a role. This is supported by the fact that we find no differences in baseline grade 1 tests scores or grade 2 test scores gains once we condition on pre-determined characteristics. The second concern is that test scores are inherently noisy measures of latent achievement which attenuates the coefficients on lagged achievement and may bias the free-track-choice coefficient in the process. However, our results are virtually unchanged if we exclude lagged achievement measures from (2), despite the fact that they are strongly predictive of achievement in grade 4. This suggests that measurement error does not pose a problem for estimating the free-track-choice coefficient.

[22]Many of these variables are potentially endogenous as they could be influenced by student behaviour (e.g., time spent reviewing material in class). We therefore only consider them for robustness checks and not for our main results.

[23]See Table A.1 for the timing of tests and surveys in the NEPS. Figure A.1 shows the distribution of test dates by binding and nonbinding states.

[24]Table A.3 in the appendix reports the state-specific starts of the school years 2012/13, 2013/14, and 2015/16 when students were tested in grade 1,2 and 4, respectively.

did not start until September 15. Combined with differences in testing dates this means that some students will have experienced up to 5.5 months of schooling in grade 4 when tested, while others could have spent as little as 2.5 months in grade 4. To make tests comparable and ensure that this does not confound our estimates, we proceed as follows: We first regress tests scores on a cubic function of age at test, fixed effects for the month of the test,[25] and a linear control for the day of the school start in the school year that the test was administered. We then use the residuals from these regressions as outcome variables.

3.3 Difference-in-difference-differences (DiDiD)

In a third approach, we combine the DiD and lagged valued-added approach to address any remaining potential biases. For a subset of states, we have state-level data on the distribution of proficiency levels (measured in 5 categories) in math and reading in grade 3 for the same cohorts that were tested in grade 4 in the NAS (i.e., before and after the repeal).[26] This allows us to use each cohort's own grade 3 proficiency level distribution as a control for its grade 4 distribution in a difference-in-differences approach exploiting the repeal of binding track recommendations. In practice, we perform double differences: one within cohorts (across grades 3 and 4) and one across cohorts (pre- and post-repeal). The idea is again that binding track recommendations should have a stronger effect on incentives to perform well in grade 4 (where performance matters for recommendations) than in grade 3. Intuitively, the identifying assumption of this approach is that there are no other factors affecting pre- and post-repeal cohorts differently between grades 3 and 4. We will refer to this approach as between-grade difference-in-differences (BG-DiD) to distinguish it from the previous DiD approach that is only based on grade 4 performance measures. The BG-DiD estimate can be obtained by taking mean differences of the four

[25] We only have information on the month but not the exact day of testing. Hence, fixed effects for the month of the tests is the most flexible way to control for it.

[26] See Section 4 for more details on these data and their comparability to the NAS (which uses the same proficiency level classification).

cells defined by the two pre- and post-repeal cohorts in grades 3 and 4:

$$\hat{\Delta}_l^{BG-DiD} = (\bar{Y}_{l,Post,4} - \bar{Y}_{l,Pre,4}) - (\bar{Y}_{l,Post,3} - \bar{Y}_{l,Pre,3}) \tag{3}$$

where $\bar{Y}_{lcg}$ denotes the mean share of students scoring at or above proficiency level l of cohort c (pre- or post-repeal) in grade g in repeal states. This approach addresses potential confounds in the DiD design due to differences across pre- and post-repeal cohorts between repeal and no-repeal states. It also accounts for potential bias due to differences in potential achievement gains between states with and without free track choice in the lagged value-added model. It should be noted that if tracking incentives already affect student performance in grade 3, this will biase BG-DiD estimates towards zero.

The BG-DID approach can be made more robust by drawing on states that did not switch to free track choice as an additional control group. The expanded difference-in-difference-in-differences (DiDiD) version of (3) amounts to the difference between the BG-DID estimate for repeal and no-repeal states:

$$\begin{aligned}\hat{\Delta}_l^{DiDiD} = &(\bar{Y}_{l,Repeal,Post,4} - \bar{Y}_{l,Repeal,Pre,4}) - (\bar{Y}_{l,Repeal,Post,3} - (\bar{Y}_{l,Repeal,Pre,3}) \\ &- [(\bar{Y}_{l,No-repeal,Post,4} - \bar{Y}_{l,No-repeal,Pre,4}) - (\bar{Y}_{l,No-repeal,Post,3} - \bar{Y}_{l,No-repeal,Pre,3})]\end{aligned} \tag{4}$$

where the additional subscript s of $\bar{Y}_{lscg}$ indexes repeal and no-repeal states.

The second difference in the DiDiD design nets out any changes in proficiency levels between cohorts and across grade-levels that are unrelated to the repeal of binding track recommendations. This accounts, for example, for differences in the scaling and definition of proficiency levels across cohorts and grades.

For the BG-DiD and DiDiD approach only state-level data is available. Since we lack additional covariates that vary within-cohorts at the state-level, four (eight) means have to be estimated for the BG-DiD (DiDiD) approach. Regressions produce a perfect fit

with no residual variance in this case, which rules out inference.[27]

4 Data

4.1 National Assessment Study (NAS)

This study draws on three main sources for student-level data. First, for the DiD and the DiDiD designs we use the National Assessment Study (NAS), which is designed to produce representative test score data for all 16 German federal states.[28] It is a repeated cross-section and has been administered in 2011 and 2016, testing students at the end of grade 4 (between May and July) in math and German (reading and listening). Tests were administered by external staff to around 25,000 students in both years, with each state contributing roughly a similar number of students. Around 2,600 schools participated in both waves combined. In each wave, random samples of schools were drawn to be at the federal state level and within each school, one class was randomly selected for testing. Not all students answered the same set of questions. Instead, schools were randomly assigned booklets with different blocks of questions which contained a subset of the complete item pool (also called 'block design').[29] However, test items are spread over booklets in a way that allows for the transformation of all tests on a common scale (for more details, see Stanat et al., 2012). In addition to test score data, the NAS includes information collected through surveys of the participating students, their parents, teachers and school principals. While participation in the competence tests was mandatory in all sampled classes in public schools, completion of the student questionnaire was mandatory only in some states, and participation in the parent questionnaire was voluntary in all states. As a result, participation rates for the student and parent questionnaire (83% and 74%, respectively) are considerably lower than test participation, which is 98% and 94% for

[27]Randomization inference has limited applicability in this context since we only have data for one repeal state (Baden Württemberg) and three no-repeal states (Schleswig-Holstein, Berlin, and Brandenburg). With only three possible permutations of the data, the p-value is only interval-identified and the lowest attainable interval is 0 to 0.25 (which our estimates always attain).

[28]For more details on the NAS data, see Stanat et al. (2012, 2017).

[29]For math the booklets were randomized within classes. For the language tests this was not possible since the listening test required to play an audio recording to the entire class.

waves 2011 and 2016, respectively. Table 2 presents student-level summary statistics, pooling the data from 2011 and 2016.

4.2 German National Educational Panel Study - Starting Cohort 2 (NEPS)

The second data source for the lagged value-added approach is starting cohort 2 of the German National Educational Panel Study (NEPS), an individual-level panel, following a nationally representative sample of students who entered primary school in 2012 over time. The data contain test scores and detailed information on family background and the schooling environment, collected through student, teacher, principal and parent questionnaires between 2010 and 2018.[30] Table A.1 shows the structure and timing of the tests and surveys. Competence tests for the NEPS were administered in schools by external staff and participation was voluntary. We will rely mostly on test scores in grade 1 (math, language, and science), grade 2 (math, reading, science, and non-verbal cognition) and grade 4 (math, reading, and orthography).[31]

We have very detailed information on teachers and school principals. In addition to demographic characteristics, these include teachers' teaching philosophies, pedagogical approaches, and classroom practices. We explain how we construct summary measures based on this information to control for differences in the schooling environment for robustness checks in Appendix D.

4.3 National Orientation Tests (VERA-3)

Since 2008, students in all German federal states have taken the nationally-standardized VERA-3 test in math and German language at the end of grade 3 (April-May) every

[30] See Blossfeld et al. (2011) for an overview. The NEPS originally started with a sample from the population of children attending day-care facilities (*Kindergarten*) two years prior to regular school enrollment in the year 2010/2011. However, the great majority of these children could not be followed into primary school. Therefore, a new refreshment sample of first grade students was drawn in 2012/13. It is the refreshment sample that we work with.

[31] Unfortunately, higher grade test scores are not available and in grade 3 neither math, reading, nor orthography were tested.

year.[32] These tests are administered and rated by teachers based on detailed guidelines. Test results do not count towards students' grade. Instead, the stated goal of these tests is to provide schools and teachers with feedback on their students' competences compared to other classes and schools (KMK, 2020). VERA-3 student-level data is generally not made available to researchers, but some federal states publish results aggregated at the state level. These are Schleswig-Holstein, North Rhein-Westphalia, Baden-Württemberg, Berlin, and Brandenburg.[33] The aggregated data contain information on the fraction of students scoring at one of five proficiency levels (below minimum level, minimum level, regular level, regular plus level, optimal level) for each subject (KMK, 2013a,b,c). We will use waves 2010 and 2015, which correspond to the same cohorts for which we have NAS test scores in grade 4. Importantly, the NAS publishes comparable state-level data on the distribution of students across proficiency levels based on the same scaling and proficiency-level classification as VERA-3.[34]

4.4 Sample selection

Our analytic samples form the NAS, NEPS, and VERA-3 data are selected in the following way.

Since survey participation was mostly voluntary for both the NEPS and the NAS, background characteristics are missing for a considerable share of students, especially in the NAS data. To increase sample size, we keep all observations with non-missing test scores. Missing control variables are dealt with in following way: We create a separate missing category for all categorical variables (e.g., parents' work status). Missing values for control variables that enter linearly in our estimations (e.g., parents' years of educa-

[32]Students with diagnosed special educational needs and students who have lived in Germany for less than a year are exempted.

[33]A request for state-level data for the other federal states was unsuccessful. Note that states' performance in VERA-3 is a politically sensitive topic as it allows to compare the performance of each federal state's educational system over time (recall that each federal state runs its own education system). For this reason, most federal states do not published their VERA-3 results to prevent unwanted discussion of their school policies. This is also reflected in data usage policies for the NEPS and NAS microdata, which explicitly prohibt the publication of results that allow the identification of individual federal states. Results can only be published for aggregated groups of federal states where at least two states are combined to form a single meaningful group.

[34]The results are published in Stanat et al. (2012, 2017).

tion) are imputed by the respective school-level average. In case a school has only missing values for a variable (e.g., for some variables derived from the school principal or teacher questionnaire), we impute missing values by the respective state-year average. We do not impute any test scores or other variables that are used as outcome variables.

From the NAS and VERA-3 data, we exclude the state of North Rhine-Westphalia because the decision to allow free track choice was made at the end of 2010, just a few months before the first-wave tests were administered in 2011. Hence, it is unclear when exactly students became aware of this change and whether they would have had enough time to adjust their behaviour. Furthermore, the NAS administered easier tests for students with severe mental or physical disabilities. We exclude these students because it is unlikely that they respond to track assignment rules. From both datasets, we also exclude classes with an unusually high share of special-needs students,[35] as teaching can be expected to be very different in these classes.

From the NEPS data, we further drop all observations from the state of Brandenburg. Brandenburg has binding track recommendations, but besides Berlin it is the only state where students transition to secondary school tracks after grade 6 and not after grade 4 as in all other states. Hence, it is their performance in grade 6 that matters most for the track recommendation and whether one should expect student responses in grade 4 is unclear.

4.5 Descriptive statistics

Table 2 and 3 show descriptive statistics for the estimation samples for the NAS and NEPS data, respectively.[36] The NAS sample for the DiD analysis includes slightly less than 48,000 students, out of which 11 percent are from repeal states. The NEPS sample for the lagged value-added approach includes around 4,800 students, out of which 25 percent are from states without free track choice. Figure 4 shows the distribution of

[35]Specifically, we exclude classes with a special-needs student share above the 99% percentile, which are classes with more than 40% special-needs students.

[36]Since some students did not participate in all grade 4 tests, the estimation samples for each outcome will differ slightly. Here we report descriptives for all students with at least one non-missing test score in grade 4

proficiency levels in grades 3 and 4 for the cohorts of students who were in grade 4 in the school years 2010/2011 and 2015/2016. Grade 3 data comes from VERA-3 and grade 4 data from the NAS. As can be seen, students are distributed roughly evenly across the five proficiency levels, but the highest proficiency level represents the smallest group.

[Table 2 about here]

[Table 3 about here]

[Figure 4 about here]

5 Results

5.1 Does binding track assignment limit school track choice?

In this section, we provide direct evidence that binding track recommendations limit students' track choice by examining how track enrollment changes in response to granting free choice. Figure 3 plots transition rates into the academic- and basic track based on full-population, administrative data from the Federal Statistical Office (German Federal Statistical Office, 2018).[37] The solid line corresponds to the group of states who repealed binding track recommendations in 2011 and the dashed line to states without changes to their track assignment rules between 2005 and 2018.[38]

[Figure 3 about here]

Figure 3 shows that transition rates for the academic and basic track differ across repeal and no-repeal states but evolve in parallel in the years leading up to the repeal of binding track assignment. In 2012, the year that parents were given free choice in repeal

[37]German Federal Statistical Office (2015a). Allgemeinbildende Schulen: Fachserie 11, Reihe 1.https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/BroschureSchulenBlick.html.

[38]This figure excludes North Rhine-Westphalia because it changed its track recommendation policy in 2010 and it is also excluded from the DiD analysis below. Furthermore, in pooling the data from different states all states receive the same weight. This is done to mirror the NAS sampling where all states contribute approximately the same number of observations. Weighting states by student population gives almost identical results.

states, transitions into the academic track (basic track) noticeably increase (decrease) in repeal states, with no corresponding changes in the no-repeal states. For the academic track, the repeal seems to result in a one-time permanent rise in the share of students attending the academic track, with no further changes in subsequent years. In contrast, transitions into the basic track continue to decline in the post-repeal years.

Overall, Figure 1 shows that transition rates change mostly in repeal states and that these changes coincide with the repeal of the binding track assignment. This is important for our analysis for two reasons. First, binding track recommendations seem to be binding constraints in the sense that their removal increases enrolment in the most preferred track while reducing enrolment in the least desired track.[39] This highlights their high-stake nature. Second, since track decisions should be reflective of student achievement, similar growth patterns in transition rates across repeal and no-repeal states prior to the repeal are consistent with the common trend assumption of our DiD design.

5.2 Test score results

5.2.1 Difference-in-differences results

In Table 4, we present the main DiD results of estimating equation (1) based on the NAS data. The dependent variables are grade 4 tests scores in math, reading, and listening, which have all been standardized to have mean zero and unit variance to facilitate interpretation. Column 1 shows results if we only control for state, year, and test booklet fixed effects. These suggest that repealing binding track recommendations is associated with a decline in student performance in grade 4 of 0.17 standard deviations in math and 0.14 in reading and listening.

In column 2, we add several class-level controls to account for potential differences in the student composition over time. These include class-level averages of the following variables: Reported books at home, parental years of education, and their highest Inter-

[39]Osikominu et al. (2020) provide further evidence that binding track assignment constrains secondary track choice by showing that parents were more likely to send their child to a higher track than recommended by teachers after the repeal of binding track assignment in Baden-Württemberg.

national Socio-Economic Index of Occupation Status (ISEI, Ganzeboom et al., 1992),[40] indicators for migration background, non-native students, special-needs status, grade repeater, and any learning disability (e.g., ADHD or dyslexia). The estimates in column 2 are somewhat smaller in size than the estimates in the first column, suggesting that unadjusted DiD effect sizes are slightly upward biased due to a relative decline in the socio-economic composition of the sample of students in states that repealed binding track assignment. However, all estimates remain negative and statistically significant. Free track choice is now estimated to lower test scores by 0.11 standard deviations in math (p-value$<$0.01), 0.08 in reading (p-value$<$0.1), and 0.082 in listening (p-value$<$0.01).

[Table 4 about here]

In column 3, we control for potential differences in schooling inputs within states over time. These include controls for whether the school offers a full-day program, is a private school, instruction hours in math and German, school enrolment, and the work experience of the teacher in each respective subject. Adding these controls has no effect on the estimates and their standard errors, suggesting that potential changes in the schooling inputs do no confound our estimates.

Finally, in column 4, we add individual student-level controls. These include, in addition to those used for the class-level averages, quadratic age controls, years spent in public childcare, dummies for whether a student has skipped a grade, fathers' and mothers' working status in four categories (not employed, part-time, full-time, or other), and their birth countries (aggregated into five categories).[41] Again, estimates are nearly identical to those of column 2 and 3, suggesting that our class level controls are sufficient to remove any bias from differences in the overall student composition over time.

In Appendix B, we provide additional robustness checks to rule out that our results are caused by compositional differences in the student population. We further check that our results remain statistically significant when inference is based on standard errors clustered at the state level.

[40] The ISEI measure of occupational prestige ranges from 16 (e.g., cleaning personnel) to 90 (judges).

[41] We create categories for the four most common countries of origin (Germany, Turkey, former Soviet countries, and Poland) and pool all remaining countries into one category to avoid too small cell sizes.

5.2.2 Value-added results

We now turn to the lagged valued-added model. Before we present our main results based on this approach, we first check its validity. Specifically, one may worry about other differences between binding and non-binding states, even after controlling for grade 2 achievement. To address this concern, we provide two sets of tests for potential bias based on the idea that binding track recommendations should only become salient towards the end of primary school and therefore not affect achievement in lower primary school grades. First, we test for baseline differences in lower-grade test scores between states with and without free track choice after regression-adjusting for pre-determined characteristics. Second, we test for differential lower grade test score gains between choice and no-choice states.[42]

Table 5 presents the results of these specification checks for average test scores in Panel A and separately for each available test in the lower panels.[43] Columns 1-2 report the results from regressions of test scores in grade 1 on an indicator for free track choice with and without further controls. Columns 3-4 show corresponding estimates for grade 2 test scores. Without any controls (columns 1 and 3), students in choice states have consistently lower test scores in grades 1 and 2. These differences are substantial, ranging from 0.06 to 0.20 standard deviations, and are mostly statistically significant. However, once we add school-, class-, and student-level controls (excluding lagged test scores) in column 2 and 4, these test score gaps become small in magnitude, are never statistically significant, and often reverse sign—consistent with no systematic differences between choice and no-choice states conditional on controls.

While this shows that there are no baseline test score differences after regression-adjusting, choice and no-choice states should also not differ in their potential test score gains. As a falsification check, we test for differences in grade 2 test score gains by conditioning on cubic functions for all grade 1 test scores in column 5. Again, the estimated

[42]This test is similar in spirit to those proposed by Rothstein (2010).

[43]The samples in columns 1-2 include all observations with non-missing grade 1 test scores in the respective subject. For the samples in columns 3-5, we further restrict the sample to observations with non-missing grade 2 test scores in the respective subject. We do this to increase power, but results are virtually the same if we use the smaller analytic sample for our main results that excludes students with missing test scores in grade 4.

coefficients are small in magnitude, ranging from -0.034 to 0.027, and are never statistically significant. In sum, we find choice and no-choice states to be balanced in the level and growth of student achievement up to grade 2 after regression-adjusting.

In Table 6, we present our main results for the lagged value-added model of the effect of free track choice on test scores in grade 4. Results for math, reading, and orthography are reported in panels A, B, and C, respectively. The sample in each panel includes all students with non-missing grade 2 test scores and the respective outcome variable. In column 1, we report estimates for specifications which include only the baseline school-, class-, and student-level controls but no lagged test scores. 4th grade students in no-choice states are estimated to outperform those in choice states by 0.14 standard deviations in math, 0.07 standard deviations in reading, and 0.23 standard deviations in orthography.

Estimates for our preferred specification, which conditions on cubic functions for all test scores and teacher ratings on students' competences in grade 2,[44] are reported in column 2. As a result of the balancing documented in Table 5, these estimates are very similar to those without test score controls in column 1 but have substantially smaller standard errors. Our estimates suggest that binding track recommendations raises grade 4 achievement by 0.14 standard deviations in math (p-value$<$0.001), 0.06 standard deviations in reading (p-value$<$0.1), and 0.21 standard deviations in orthography (p-value$<$0.001). The result for reading is remarkably close to the estimate from our preferred specification based on the DiD design reported in Table 4, while the math coefficient is slightly larger than the DiD estimate. In column 3, we also report results with grade 1 test score controls. Adding these controls reduces the sample size by around 10% but gives very similar results. We therefore view the specification in column 2 as our preferred specification.

In Appendix B, we provide several further robustness checks for the results in Table

[44] Assessed competences include social skills, persistence and ability to concentrate, language, science, and math. These subjective assessments were elicited by asking teachers to assess the skills and abilities of a student in comparison to all other students of the same age on a 5-point scale ranging from "much worse" to "much better". We also control for teachers' responses to the question: "From today's perspective, what school type would you recommend for this child?" in grade 2. Omitting controls for students' assessed competences by their teacher leaves the coefficients in column 2 virtually unchanged. This is further evidence for the assumption that students' unobserved characteristics are uncorrelated with track-choice policies conditional on our controls.

6. These include a replication of the specification in column 1 with the same cohort of students from the NAS data, re-estimation of the results dropping one state at a time, inference based on standard errors clustered at the state level and wild cluster bootstrap procedure, and specifications with more extensive (but potentially endogeneous) controls for the schooling environment.

5.2.3 Difference-in-difference-in-differences results

Finally, we present estimates based on the BG-DiD and DiDiD designs which exploit within-cohort variation across grades and between pre- and post-repeal cohorts. As noted in Section 3, an important caveat is that all of the following estimates are based on only 4-8 cell means. We therefore make no attempt at inference. To provide a fuller picture of where differences emerge, Figure 5 shows post-pre repeal differences in the share of students scoring at or above specific proficiency levels separately by proficiency levels, grades, and repeal and no-repeal states. Panel A reveals no consist differences in the distribution of grade 3 math proficiency levels between pre- and post-repeal cohorts in the repeal state (grey bars). For example, while the share of students scoring at or above the 2nd lowest math level decreases from 2010 to 2015 by 5.35 percentage points, the share at or above the 2nd highest level (level 4) increases by 5.4 percentage points. In contrast, the grade 4 post-pre repeal differences are consistently negative, ranging from -10.10 percentage points ($\geq$ level 4) to -6 percentage pints ($\geq$ level 2 and $\geq$ Level 5). Panel B shows a similar pattern but with much smaller post-pre repeal differences for no-repeal states.

Figure 5, Panels C and D report analogous results for reading. Similar to the results for math, Panel C shows that the share of students scoring at higher reading proficiency levels in grade 4 decreases after the repeal of binding track assignment in the repeal state. For no-repeal states (Panel D), there are no post-pre repeal differences in grade 4. It should be noted that Panel C also reveals large positive post-pre repeal differences (up to 16 percentage points) in grade 3 in the repeal state. However, as the same pattern can be observed for the no-repeal states, this is most likely due to differences in the definition

of grade 3 proficiency levels across cohorts (i.e., lower-proficiency students being classified into higher proficiency levels in the pre-repeal cohort), rather than genuine improvements in post-repeal students' reading proficiency in grade 3.

Figure 6 combines the results of Figure 5 by reporting BG-DiD and DiDiD results based on equations (3) and (4), respectively. For math (Panel A), the BG-DiD results (grey bars) indicate no change in the the share of students with the lowest math proficiency level in response to free track choice in the repeal state. However, the share scoring at or above a specific higher proficiency level increases for all proficiency levels above the second lowest level from grade 3 to grade 4 for the post-repeal cohort relative to the pre-repeal cohort. Results for the DiDiD results (black bars), which are obtained by taking the difference of BG-DiD estimates between repeal and no-repeal states, yield very similar results, ranging from -7.59 to -12.05 percentage points.

Figure 6, Panel B shows the corresponding BG-DiD and DiDiD results for reading. As for math, both approaches indicate a relative decline in 4th graders' reading proficiency in response to free track choice. However, here the two approaches yield very different effect sizes. Estimates for the BG-DiD range from -4.9 to -21.60 percentage points, while the DiDiD results are much smaller and exhibit less variation across proficiency levels, ranging from -4.03 ($\geq$ level 2) to -6.74 percentage points ($\geq$ level 4). The larger BG-DiD effects are most likely due to a more generous proficiency level classification for the post-repeal cohort in grade 3, as alluded to in our discussion of Figure 5. The resulting increase in grade 3 proficiency levels for the post-repeal cohort inflates BG-DiD estimates. The DiDiD design nets out this effect, as post-repeal cohorts in no-repeal states exhibit a similar increase in their grade 3 proficiency levels, which explains the smaller DiDiD estimates.

In sum, the BG-DiD and DiDiD results provide further evidence for a decline in students' academic achievement in response to free track choice.[45]

[45]These results have to be interpreted in terms of changes in the share of students scoring at or above specific proficiency levels. As such, they are not directly comparable to the DiD and lagged value-added estimates, which measure average test score changes. However, to test the comparability of our results, we checked that the DiDiD results deliver similar results as those based on a data generating process with the following three features: (i) normally distributed student skills, (ii) shifts in the skill distribution in response to free track choice corresponding to the DiD and lagged value-added estimates in Tables 4

5.3 Effects on well-being and intrinsic motivation

In order to fully assess the implications of high-stakes incentives, it is important to consider other potential outcomes besides academic achievement. In particular, the richness of the NEPS data permit us to explore two potential downsides of high-stakes incentives: Their detrimental effects on student well-being and intrinsic motivation. To this end, we again use equation (2) using our preferred specification with full controls and different measures of student well-being and a proxy for students' intrinsic motivation as outcome variables. The results are reported in Table 7. Each column refers to a different outcome, which are all standardized to have mean zero and unit variance. Outcomes are mainly available for grade 4, but when available for grade 3 or 5, we also report estimates for these.

An important question is whether the increased pressure associated with high-stakes incentives reduces student well-being. Performance settings that contain both uncontrollable and social-evaluative elements induce the largest increase in physiological stress responses (Dickerson and Kemeny, 2004), which are associated with worse mental health and well-being outcomes in students (Shankar and Park, 2016). As a measure of subjective well-being, the NEPS survey includes the life satisfaction question: "How satisfied are you overall with your life?", which students answer on a 7-point scale ranging from "completely unsatisfied" to "completely satisfied." This is one of the most widely used measures of subjective well-being (Kahneman and Deaton, 2010).[46] Note that student interviews in grade 4 and 5 took place between October and January of the respective school year (see Table A.1). That is, we are measuring students' well-being before they know their track placement in grade 4 and after their transition to a secondary school track in grade 5. Hence, survey responses in grade 4 should reflect the uncertainty and pressure associated with having to qualify for a higher track in no-choice states rather than differences in eventual track placements.

and 6, and (iii) a skill-based classification of students into five proficiency levels matching the empirical distributions in Figure 4.

[46] Despite some well-documented limitations of subjective well-being measures, there is ample evidence that they contain significant information about the individual's true well-being. Subjective well-being is, for example, positively correlated to objective measures of well-being, such as emotional expressions (Sandvik et al., 2009), and activity in the pleasure centers of the brain (Urry et al., 2004).

As shown in the first column of Panel B, subjective well-being in grade 4, when students should be most concerned about their track prospects, is 0.14 standard deviations higher in choice states relative to no-choice states (p-value$<$0.001). This effect persists through grade 5, after students have entered a secondary school track (0.12 standard deviations, p-value$<$0.05), indicative for a permanent reduction in student well-being in no-choice states. The fact that we do not find any effect in grade 3 (Panel A)—when track placements are not yet salient—underscores a causal interpretation of the results. Columns 2-4 in Panel B further indicate that the reduction in well-being in grade 4 is not health related but, in line with an explanation based on academic stress, driven by school performance anxiety and concerns about one's future.[47] Note that we find these negative effects for school performance anxiety despite the fact that students from no-choice states have better test scores.

Research in psychology has debated for over three decades whether external incentives inhibit students' subsequent intrinsic motivation (see, e.g., Deci et al., 1999). We test for potential crowding-out effects in the last column of Table 7. Our measure of intrinsic motivation is the first principal component of several student and parent questionnaire items asking, for example how much students enjoy learning or whether they feel bored at school.[48] Mirroring the well-being results, we find no effect of binding track recommendations on intrinsic motivation in grade 3 (Panel A), which is a useful specification check. By grade 4, however, intrinsic motivation is 0.13 standard deviations lower in binding states (p-value$<$0.05, Panel B). It remains low and decreases even further by grade 5 (Panel C) when students are not subject to tracking incentives anymore.

A caveat with the results for grade 5 is that they might reflect differences in track placements due to greater track choice in non-binding states. To address this concern, we report in Table A.5 in the appendix results where we control for the mediating role of

[47]Satisfaction with health is measured on a 7-point scale by the question "How satisfied are you with your health?" Anxiety about grades and students' own future is measured on a 5-point scale by agreement to the statements "In my last school week I was afraid of getting poor grades" and "In my last school week I worried about my future."

[48]The items and results for the principal-components analysis are reported in Table D.1 in the appendix.

track placements by conditioning on the track that students attend in grade 5.[49] These estimates are very similar to those in Panel C of Table 7. The documented effects in grade 5 thus seem to capture the persistent reduction of student well-being and intrinsic motivation due to the high-stakes incentives arising from binding track recommendations, as opposed to differences in track placements.

As far we know, ours is the first study to demonstrate the potentially harmful effects of high-stakes incentives on students' subjective well-being and intrinsic motivation. This contrasts with previous economic studies, which mostly find no evidence for crowding-out effects of external incentives in educational settings (see, e.g., Kremer et al., 2009; Fryer, 2011; Bettinger, 2012; Levitt et al., 2016; Barrow and Rouse, 2018).[50] However, most of these studies looked at relatively low-stakes incentives. One interpretation of this finding is that only sufficiently strong external incentives crowd-out intrinsic motivation.

5.4 Discussion of potential mechanisms

The documented achievement effects of binding track assignment could result from student and/or parental responses. In this subsection, we try to assess their respective role in driving the achievement effects. We also discuss possible distortions of incentives through teachers. Table 8 presents results using the same specification as in Table 6 with several proxies for study effort and parental behavior in grade 4 as outcome variables.[51]

As a proxy for students' study effort we use the reported time students spend on doing homework and other school exercises outside the classroom. Column 1 shows that students in states with binding track assignment spend 13 more minutes per day studying outside the classroom (p-value<0.001)—a 25% difference against the mean private study time of 51 minutes per day. One explanation for the substantial increase in study effort

[49]The validity of this mediation analysis depends on the assumption that track choice is independent conditional on our controls. To further test the plausibility of this assumption, Table A.5 also reports estimates where we additionally control for parents' stated track preferences in grade 2, which gives very similar results.

[50]The only study we could find that demonstrates crowding-out effects is by Visaria et al. (2016). They show that a scheme to reward school attendance in India lowered post-incentive school attendance, test scores, and intrinsic motivation, but only among students with low baseline attendance.

[51]Except for the outcome in the first column (self-study), all outcomes have been standardized to facilitate interpretation.

could be that parents encourage their children to study more in no-choice states. The NEPS includes several questions intended to measure this type of parental behavior, for example, whether parents pay attention to how much time their child spends on homework. We performed a principal components analysis on all these items and generated a parenting-style factor which loads positively on stricter rules and closer parental monitoring.[52] Column 2 shows the estimated effect if we use the resulting parenting-style factor as outcome variable. Contrary to what one would expect if the increase in study effort in no-choice states were explained by stricter rules and closer parental monitoring, the estimated effect is significantly positive. The most likely explanation for this finding is that parents set stricter rules and monitor their children more closely in response to students' lower study effort and achievement in choice states. This strongly suggests that the increase in study effort in no-choice states is driven by students themselves rather than enforced by their parents.

Another explanation for the relative achievement gains in no-choice states could be greater parental investments in the form of homework assistance or private tutoring. We investigate this possibility in columns 3-5 with several measures of parental investments as outcome variables. There is no evidence for differences in parental investments across choice and no-choice states; parents in no-choice states do not report to assist their children with homework or other school exercises more often (columns 3-4), neither do they invest more into private tutoring (column 5). The NAS data contain information on private tutoring as well. Table A.6 presents results based on the DiD specification with private tutoring in grade 4 as the dependent variable. Again, contrary to what one would expect if the achievement gains due to tracking incentives were driven by greater parental investments, repealing binding track assignment appears to decrease private tutoring. This is further evidence that parents respond to lower student effort under free track choice by investing more into private tutoring.

Furthermore, we can test whether teachers adjust the strictness of their track recommendations depending on whether they are binding or not. For example, teachers

[52]See Tables D.2 and D.3 for the questionnaire items and the results of the principal-components analysis. We use the first principal component as outcome variable in Table 8.

could be more lenient when a student's further academic career directly depends on their assessment. This would imply that teachers attenuate the incentive effect of binding recommendations. We test this with the NAS data using the DiD approach. Table A.4, column 4 shows that teachers make fewer academic track recommendations when they are non-binding, consistent with lower student achievement in the absence of high-stakes incentives. However, once test score controls are included to account for the fact that student achievement declines with free track choice, column 5 shows that this effect disappears. This suggests that teachers equally reward performance under both conditions and do not weaken students' and parents' investment incentives arising from binding track recommendations.[53]

Taken together, the results in this section suggest that the documented achievement gains due to a binding track assignment policy are driven by students' effort responses to high-stakes incentives.

5.5 Allowing for heterogeneous effects

In this section, we explore whether students differ in their response to high-stakes incentives by interacting the free-track-choice indicator with several student characteristics. The coefficients for these interaction terms are reported in Table 9. All regressions in Table 9 are based on the lagged valued-added model and include the same controls as those of our preferred specification from column 2 in Table 6.

Several studies have found that girls respond stronger to incentives in educational settings (see, e.g. Angrist et al., 2009; Angrist and Lavy, 2009; Kremer et al., 2009; Bettinger, 2012; Hvidman and Sievertsen, 2019) and are more patient than boys (Sutter et al., 2019), so should in particular respond more strongly to delayed rewards. The results in Panel A confirm this: Girls spend 15 more minutes per day studying on their own in no-choice relative to choice states. For boys the effort response is with 10 minutes somewhat smaller. These gender differences in study-effort responses are also reflected in achieve-

[53]Additionally, Table B.3 in the appendix shows that teacher behavior does not explain any achievement effects of tracking incentives as the results are unchanged when controlling for classroom activities, teaching styles, teacher school involvement etc.

ment differences, which are always larger for girls than for boys. In fact, boys' reading achievement appears to be unaffected by track-assignment policies, while girls score 0.13 standard deviations higher on the reading test in no-choice states (p-value$<$0.001). In math and orthography, boys from no-choice states also outperform those from choice states, but less so than girls.

Next, we investigate distributional effects. Panel B displays results by quartiles of the distribution of average test scores in grade 2. The point estimates suggest that students across the entire achievement distribution spend more time studying when track assignment is binding and see their test scores increase as a result. However, estimated test score effects are always largest for students from the bottom quartile. Hence, binding track assignment seems not only to shift the entire achievement distribution but also to disproportionately benefit low-performing students.

If parental behavior was the main explanation for the observed test score differences between choice and no-choice states, we would expect larger effects for parents with more resources (academic knowledge, time, and money) to invest in their children. To check this, we estimate effects separately by educational background in Panel C. Overall, we find very similar effects across family backgrounds. Only for students from parents without an academic school degree do we not find a significant effect for reading. All other estimated effects are similar to the average effects reported in Table 6 and statistically significant. This pattern is consistent with the lack of evidence for parental responses in Table 8 and corroborates the idea that our estimated effects are driven by student responses rather than parental behavior.

Table A.7 in the appendix reports heterogeneous effects by gender and parental education for the NAS data based on the DiD design. The results are qualitatively very similar to those in Table 9. Repealing binding track assignment has a somewhat stronger effect on girls for all tests. Results by parental education are also roughly similar, which again speaks against parental behavior as an explanation for the strong student responses.

6 Conclusion

In this paper, we study responses to high-stakes incentives in primary school arising from track assignment policies in a rigorous early ability tracking system. Across three research designs and data sets we find free secondary school track choice, as opposed to binding track assignment, to decrease students' study effort and academic achievement at the end of primary school. These effects can be found throughout the achievement distribution. However, low-achieving students show the strongest responses, both in terms of study effort and acquired skills, suggesting that ability tracking incentives potentially decrease educational inequalities in primary school.

Yet we also identify an important trade-off with regard to high-stakes incentives. The achievement gains due to higher pressure to perform are associated with a reduction in students' well-being and intrinsic motivation to study.

Importantly, we find no evidence that parental behavior drives students' responses to tracking incentives. Plausible hypotheses about parental responses to tracking incentives, such as more homework assistance, stricter study rules and closer monitoring, or private tutoring, all seem to contradict our rich survey data in important ways. Our effects are therefore best accommodated by an explanation where ability tracking directly incentivizes primary school students to exert more effort.

Our results might at first sight appear to contradict a large body of research documenting negative effects of early tracking on achievement and educational equity. However, most of these studies are somewhat limited in their ability to identify the effort inducing effects of tracking that we find in this study. This is for two reasons. First, many studies identify tracking effects by comparing test score gains from lower grade (before tracking) to higher grade (after tracking) between tracking and no-tracking systems, thereby absorbing any pre-tracking gains (see, e.g. Hanushek and Wössmann, 2006; Ruhose and Schwerdt, 2016; Matthewes, 2020).[54] Second, much of the remaining evidence comes from studies looking at the long-run effects of postponing the age of tracking (Meghir

[54]This drawback of the test score gain design has been pointed out before (see, e.g., Hanushek and Wössmann, 2006; Manning and Pischke, 2006; Waldinger, 2007).

and Palme, 2005; Aakvik et al., 2010; Pekkala Kerr et al., 2013; Borghans et al., 2020; Canaan, 2020). Students are still tracked at some point during secondary school in these contexts. As such, these estimates tell us little about the effects of tracking incentives. This is not to say that our results imply that tracking has positive net effects. To the contrary, to the extent that we find positive effects at the end of primary school, previous estimates of the long-run effects of tracking most likely underestimate the harmful effect of students' experiences in tracked secondary schools.

While Germany's rigorous ability tracking system is unique in that it tracks students as young as age 10, our findings have broad implications since many school systems feature similar high-stakes settings which create immediate incentives for primary and secondary students. Examples include grade retention policies, admission procedures for selective primary and middle schools, or merit scholarships. Our findings suggest that these features play an important role in the acquisition of human capital by inducing greater study effort in young students who are otherwise unlikely to be aware of the long-run returns to their present educational investments (Oreopoulos, 2007; Gneezy et al., 2011). More broadly, our study demonstrates that high-stakes incentives, even those with some delay in rewards, motivate children to invest learning effort, and that this effort, indeed, improves their academic performance.

This suggests that policy levers to raise effort through the use of incentives are worth considering seriously. Policy-makers should, however, also consider that while raising educational attainment is associated with large positive effects, e.g. on later employment, earnings (Heckman, 2006), and health (Grossman, 2006), they might come at a cost to student well-being in the short run.

References

Aakvik, A., Salvanes, K. G., and Vaage, K. (2010). Measuring Heterogeneity in the Returns to Education using an Education Reform. *European Economic Review*, 54(4):483–500.

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

Abramitzky, R. and Lavy, V. (2014). How Responsive Is Investment in Schooling to Changes in Redistributive Policies and in Returns? *Econometrica*, 82.

Andrabi, T., Das, J., Khwaja, A. I., and Zajonc, T. (2011). Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *American Economic Journal: Applied Economics*, 3(3):29–54.

Angrist, J., Lang, D., and Oreopoulos, P. (2009). Incentives and Services for College Achievement: Evidence from a Randomized Trial. *American Economic Journal: Applied Economics*, 1(1):136–63.

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

Atkinson, A., Gregg, P., and McConnell, B. (2006). The Result of 11 Plus Selection: An Investigation into Opportunities and Outcomes for Pupils in Selective LEAs. The Centre for Market and Public Organisation 06/150, Department of Economics, University of Bristol, UK.

Barrera-Osorio, F., Gonzalez, K., Lagos, F., and Deming, D. J. (2020). Providing Performance Information in Education: An Experimental Evaluation in Colombia. *Journal of Public Economics*, 186:104185.

Barrow, L. and Rouse, C. (2018). Financial Incentives and Educational Investment: The Impact of Performance-based Scholarships on Student Time Use. *Education Finance and Policy*, 13(4):419–448.

Baumert, J., Maaz, K., Gresch, C., McElvany, N., Anders, Y., Jonkmann, K., Neumann, M., and Watermann, R. (2010). Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten. Zusammenfassung der zentralen Befunde. In Maaz, K., Baumert, J., Gresch, C., and McElvany, N., editors, *Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten*, Bildungsforschung. 34, pages [5–22]. Bundesministerium für Bildung und Forschung, Referat Bildungsforschung, Bonn u.a.

Behrman, J., Parker, S., Todd, P. E., and Wolpin, K. I. (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.

Bettinger, E. P. (2012). Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *The Review of Economics and Statistics*, 94(3):686–698.

Blossfeld, H.-P., von Maurice, J., and Schneider, T. (2011). The National Educational Panel Study: Need, Main Features and Research Potential. *Zeitrschrift für Erziehungswissenschaft*, 14:5–17.

Blundell, R., Dias, M. C., Meghir, C., and van Reenen, J. (2004). Evaluating the Employment Impact of a Mandatory Job Search Program. *Journal of the European Economic Association*, 2(4):569–606.

Borghans, B. L., Diris, R., Smits, W., and de Vries, J. (2020). Should We Sort It Out Later? The Effect of Tracking Age on Long-run Outcomes. *Economics of Education Review*, 75:101973.

Borghans, L., Diris, R., Smits, W., and de Vries, J. (2019). The Long-run Effects of Secondary School Track Assignment. *PLOS ONE*, 14(10):1–29.

Bos, W., Müller, S., and Stubbe, T. C. (2010). Abgehängte Bildungsinstitutionen: Hauptschulen und Förderschulen. In *Bildungsverlierer*, pages 375–397. Springer.

Burgess, S., Metcalfe, R., and Sadoff, S. (2016). Understanding the Response to Financial and Non-Financial Incentives in Education: Field Experimental Evidence Using High-Stakes Assessments. IZA Discussion Papers 10284, Institute of Labor Economics (IZA).

Canaan, S. (2020). The Long-run Effects of Reducing Early School Tracking. *Journal of Public Economics*, 187:104206.

Chadi, A., de Pinto, M., and Schultze, G. (2019). Young, Gifted and Lazy? The Role of Ability and Labor Market Prospects in Student Effort Decisions. *Economics of Education Review*, 72:66 – 79.

Chaisemartin, C. and D'Haultfoeuille, X. ((forthcoming) 2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*.

Clark, D. and Del Bono, E. (2016). The Long-Run Effects of Attending an Elite School: Evidence from the United Kingdom. *American Economic Journal: Applied Economics*, 8(1):150–76.

Damon, C. (2010). Selective Schools and Academic Achievement. *The B.E. Journal of Economic Analysis & Policy*, 10(1):1–40.

Deci, E., Koestner, R., and Ryan, R. (1999). A Meta-Analytic Review of Experiments Examining the Effect of Extrinsic Rewards on Intrinsic Motivation. *Psychological bulletin*, 125:627–68; discussion 692.

Dickerson, S. S. and Kemeny, M. E. (2004). Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research. *Psychological bulletin*, 130(3):355.

Dobbie, W. and Fryer, Roland G., J. (2014). The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools. *American Economic Journal: Applied Economics*, 6(3):58–75.

Dustmann, C., Puhani, P. A., and Schönberg, U. (2017). The Long-term Effects of Early Track Choice. *The Economic Journal*, 127(603):1348–1380.

Fryer, Roland G., J. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *The Quarterly Journal of Economics*, 126(4):1755–1798.

Ganzeboom, H. B., Graaf, P. M. D., and Treiman, D. J. (1992). A Standard International Socio-economic Index of Occupational Status. *Social Science Research*, 21(1).

German Federal Statistical Office (2018). Allgemeinbildende Schulen: Fachserie 11, Reihe 1.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives*, 25(4):191–210.

Grossman, M. (2006). Education and Nonmarket Outcomes. *Handbook of the Economics of Education*, 1:577–633.

Guyon, N., Maurin, E., and McNally, S. (2011). The Effect of Tracking Students by Ability into Different Schools: A Natural Experiment. *Journal of Human Resources*, 47.

Hanushek, E. A. and Wössmann, L. (2006). Does Educational Tracking Affect Performance and Inequality? Differences- in-Differences Evidence Across Countries. *Economic Journal*, 116(510):63–76.

Heckman, J. and Cunha, F. (2007). The Technology of Skill Formation. *American Economic Review Papers and Proceeding*, 97.

Heckman, J. J. (2006). Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, 312(5782):1900–1902.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4):605–654.

Helbig, M. and Nikolai, R. (2015). *Die Unvergleichbaren. Der Wandel der Schulsysteme in den 16 deutschen Bundesländern.*

Henry, G. and Rubenstein, R. (2002). Paying for Grades: Impact of Merit-Based Financial Aid on Educational Quality. *Journal of Policy Analysis and Management*, 21:93–109.

Hvidman, U. and Sievertsen, H. H. (2019). High-Stakes Grades and Student Behavior. *Journal of Human Resources*.

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

Jackson, C. K. (2010b). Do Students Benefit from Attending Better Schools? Evidence from Rule-based Student Assignments in Trinidad and Tobago. *Economic Journal*, 120:1399–1429.

Jalava, N., Schrøter Joensen, J., and Pellas, E. (2015). Grades and Rank: Impacts of Non-financial Incentives on Test Performance. *Journal of Economic Behavior Organization*, 115:161 – 196. Behavioral Economics of Education.

Jensen, R. (2010). The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*, 125:515–548.

Jürges, H., Schneider, K., and Büchel, F. (2005). The Effect Of Central Exit Examinations On Student Achievement: Quasi-Experimental Evidence From TIMSS Germany. *Journal of the European Economic Association*, 3(5):1134–1155.

Kahneman, D. and Deaton, A. (2010). High Income Improves Evaluation of Life But Not Emotional Well-Being. *Proceedings of the National Academy of Sciences of the United States of America*, 107:16489–93.

KMK (2013a). Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich Lesen – mit Texten und Medien umgehen“ – Primarbereich. *Kultusministerkonferenz*.

KMK (2013b). Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich Schreiben“, Teilbereich Rechtschreibung“ – Primarbereich. *Kultusministerkonferenz*.

KMK (2013c). Kompetenzstufenmodell zu den Bildungsstandards im Fach Mathematik für den Primarbereich (Jahrgangsstufe 4). *Kultusministerkonferenz*.

KMK (2020). VERA 3 und VERA 8 (Vergleichsarbeiten in den Jahrgangsstufen 3 und 8): Fragen und Antworten für Schulen und Lehrkräfte. *Kultusministerkonferenz*.

Knauf, H. and Knauf, M. (2019). Schulische Inklusion in Deutschland 2009-2017. Eine bildungsstatistische Analyse aus Anlass des 10. Jahrestags des Inkrafttretens der UN Behindertenrechtskonvention am 26. März 2019. Bielefeld working paper. 1, Bielefeld.

Knigge, M. (2009). *Hauptschüler als Bildungsverlierer?: Eine Studie zu Stigma und selbstbezogenem Wissen bei einer gesellschaftlichen Problemgruppe*. Waxmann Verlag.

Kremer, M., Miguel, E., and Thornton, R. (2009). Incentives to Learn. *The Review of Economics and Statistics*, 91(3):437–456.

Leuven, E., Oosterbeek, H., and Klaauw, B. (2010). The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment. *Journal of the European Economic Association*, 8.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016). The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

Lichtman-Sadot, S. (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, J., Sanders, N., and Oreopoulos, P. (2010). Ability, Gender, and Performance Standards: Evidence from Academic Probation. *American Economic Journal: Applied Economics*, 2.

Manning, A. and Pischke, J.-S. (2006). Comprehensive versus Selective Schooling in England in Wales: What Do We Know?

Marcus, J. and Zambre, V. (2019). The Effect of Increasing Education Efficiency on University Enrollment: Evidence from Administrative Data and an Unusual Schooling Reform in Germany. *Journal of Human Resources*, 54(2):468–502.

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

Meghir, C. and Palme, M. (2005). Educational Reform, Ability, and Family Background. *American Economic Review*, 95(1):414–424.

Metcalfe, R., Burgess, S., and Proud, S. (2019). Students' Effort and Educational Achievement: Using the Timing of the World Cup to Vary the Value of Leisure. *Journal of Public Economics*, 172(C):111–126.

Montalban, J. (2019). Countering Moral Hazard in Higher Education: The Role of Performance Incentives in Need-based Grants. Working papers, HAL.

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

Oreopoulos, P. (2007). Do Dropouts Drop out Too Soon? Wealth, Health and Happiness from Compulsory Schooling. *Journal of public Economics*, 91(11-12):2213–2229.

Osikominu, A., Pfeifer, G., and Strohmaier, K. (2020). The Effects of Free Secondary School Track Choice: A Disaggregated Synthetic Control Approach. Technical report.

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

Pekkala Kerr, S., Pekkarinen, T., and Uusitalo, R. (2013). School Tracking and Development of Cognitive Skills. *Journal of Labor Economics*, 31(3):577–602.

Roodman, D., MacKinnon, J. G., Webb, M. D., and Nielsen, M. (2018). Fast and Wild: Bootstrap Inference in Stata Using Boottest. Working Paper 1406, Economics Department, Queen's University.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics*, 125(1):175–214.

Ruhose, J. and Schwerdt, G. (2016). Does Early Educational Tracking Increase Migrant-native Achievement Gaps? Differences-in-differences Evidence across Countries. *Economics of Education Review*, 52.

Sandvik, E., Diener, E., and Seidlitz, L. (2009). Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures. *Journal of Personality*, 61:119–138.

Schildberg-Hörisch, H. and Wagner, V. (2020). Chapter 19 - Monetary and Non-monetary Incentives for Educational Attainment: Design and Effectiveness. In Bradley, S. and Green, C., editors, *The Economics of Education (Second Edition)*, pages 249 – 268. Academic Press, second edition edition.

Shankar, N. L. and Park, C. L. (2016). Effects of Stress on Students' Physical and Mental Health and Academic Success. *International Journal of School & Educational Psychology*, 4(1):5–9.

Shure, N. (2019). School Hours and Maternal Labor Supply. *Kyklos*, 72(1):118–151.

Stanat, P., Böhme, H., Weirich, S., Haag, N., Engelbert, M., and Reimers, H. (2014). IQB-Ländervergleich Primarstufe 2011 (IQB-LV 2011) [IQB National Assessment Study 2011] (Version 3) [Data set].

Stanat, P., Pant, H., Böhme, K., and Richter, D. (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011. Münster: Waxmann.*

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S., and Haag, N. (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich.*

Stanat, P., Schipolowski, S., Weirich, S., Mahler, N., and Wittig, J. (2019). IQB-Bildungstrend Primarstufe 2016 (IQB-BT 2016) [IQB Trends in Student Achievement 2016 (IQB-BT 2016) (Version 1) [Data set].

Stinebrickner, R. and Stinebrickner, T. R. (2008). The Causal Effect of Studying on Academic Performance. *The B.E. Journal of Economic Analysis & Policy*, 8(1):1–55.

Sutter, M., Zoller, C., and Glätzle-Rützler, D. (2019). Economic Behavior of Children and Adolescents – A first Survey of Experimental Economics Results. *European Economic Review*, 111:98 – 121.

Urry, H. L., Nitschke, J. B., Dolski, I., Jackson, D. C., Dalton, K. M., Mueller, C. J., Rosenkranz, M. A., Ryff, C. D., Singer, B. H., and Davidson, R. J. (2004). Making a Life Worth Living: Neural Correlates of Well-Being. *Psychological Science*, 15(6):367–372.

Villa, J. (2016). diff: Simplifying the estimation of difference-in-differences treatment effects. *Stata Journal*, 16(1):52–71.

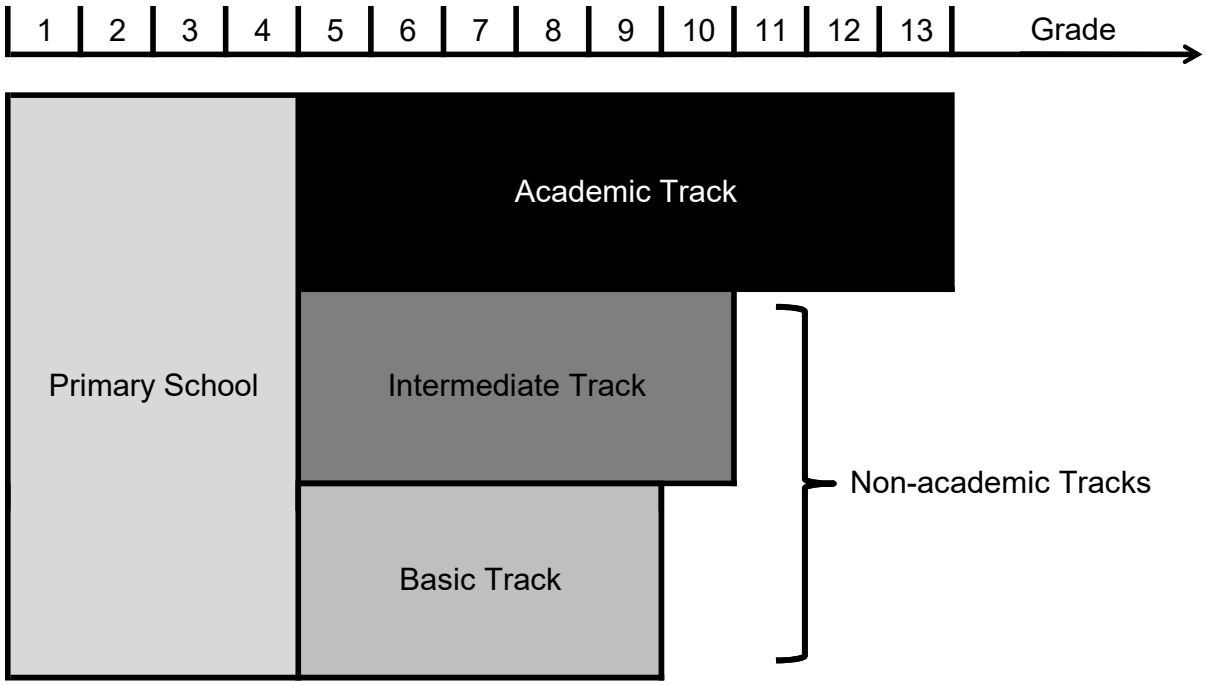
Visaria, S., Dehejia, R., Chao, M. M., and Mukhopadhyay, A. (2016). Unintended Consequences of Rewards for Student Attendance: Results from a Field Experiment in Indian Classrooms. *Economics of Education Review*, 54:173 – 184.

Waldinger, F. (2007). Does Tracking Affect the Importance of Family Background on Students' Test Scores? London School of Economics Working Paper.

Wellgraf, S. (2014). *Hauptschüler: Zur gesellschaftlichen Produktion von Verachtung.* Transcript Verlag.

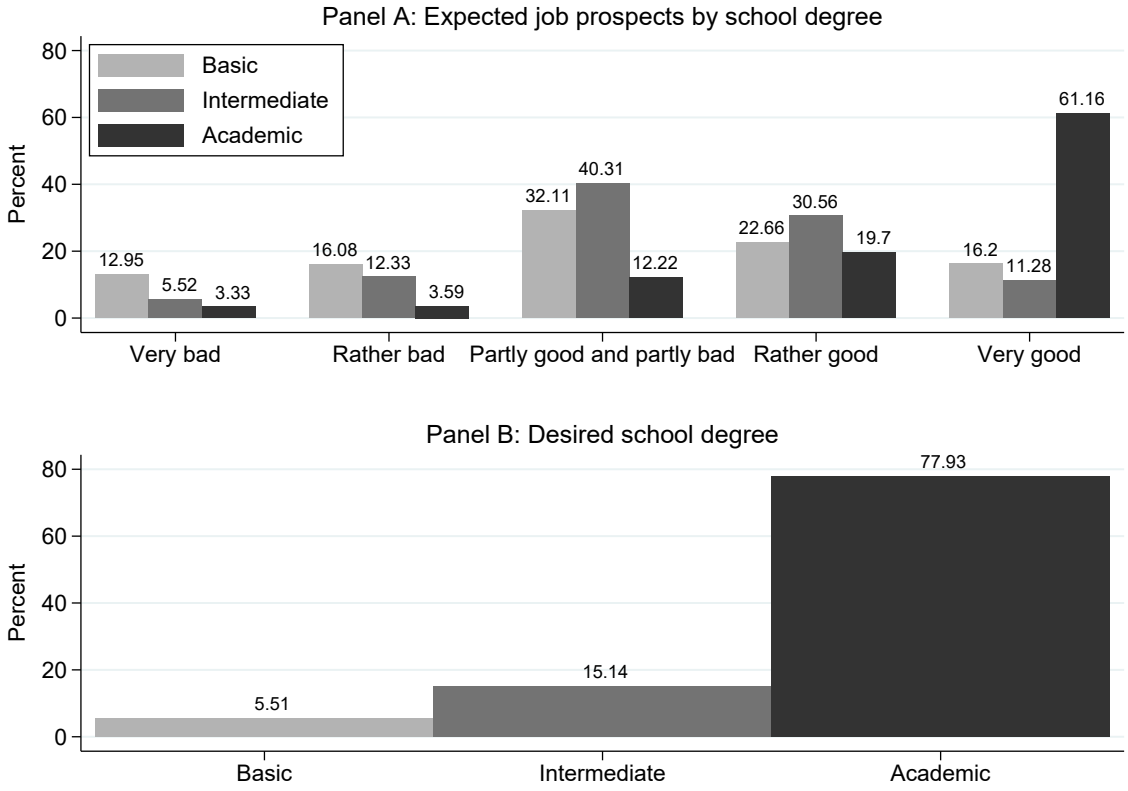
Figures and tables

Figure 1: Schematic overview of the tracking system in Germany



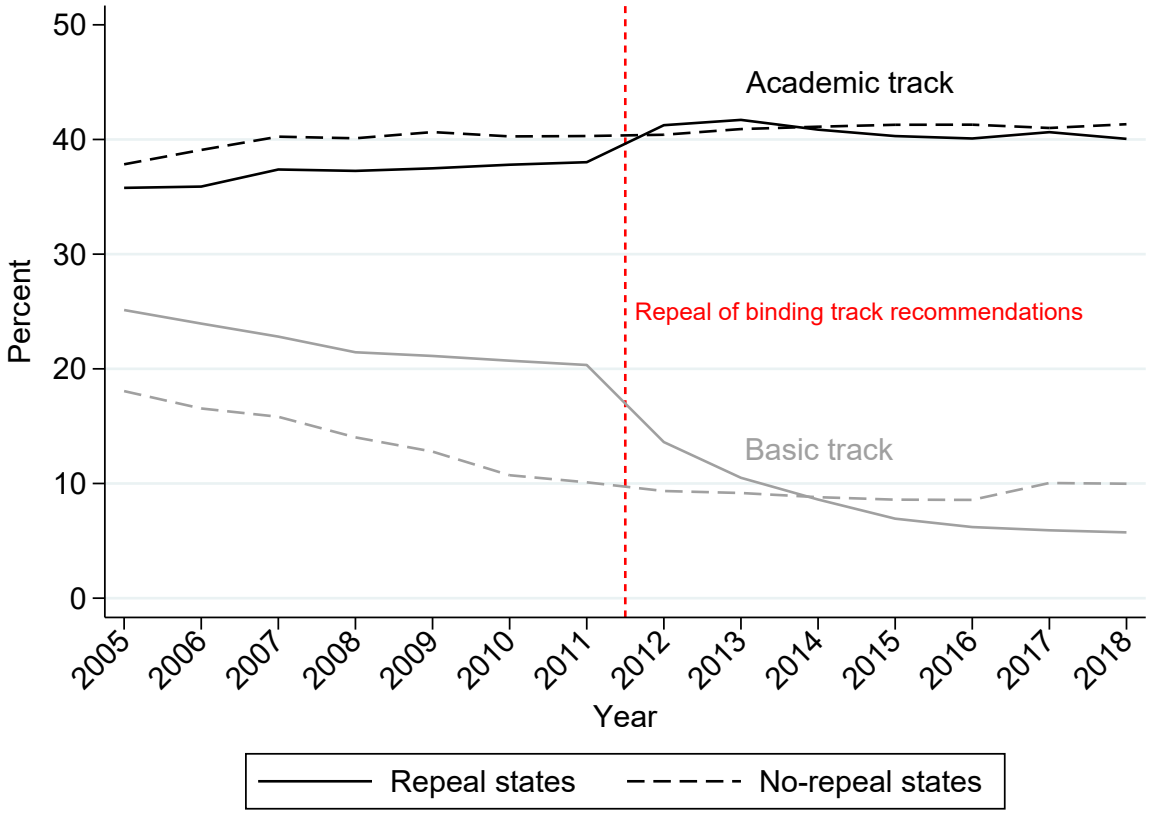
Notes: Figure adapted from Matthewes (2020). Academic track = *Gymnasium*, Intermediate track = *Realschule*, Basic track = *Hauptschule*.

Figure 2: Primary school students' views on school tracks



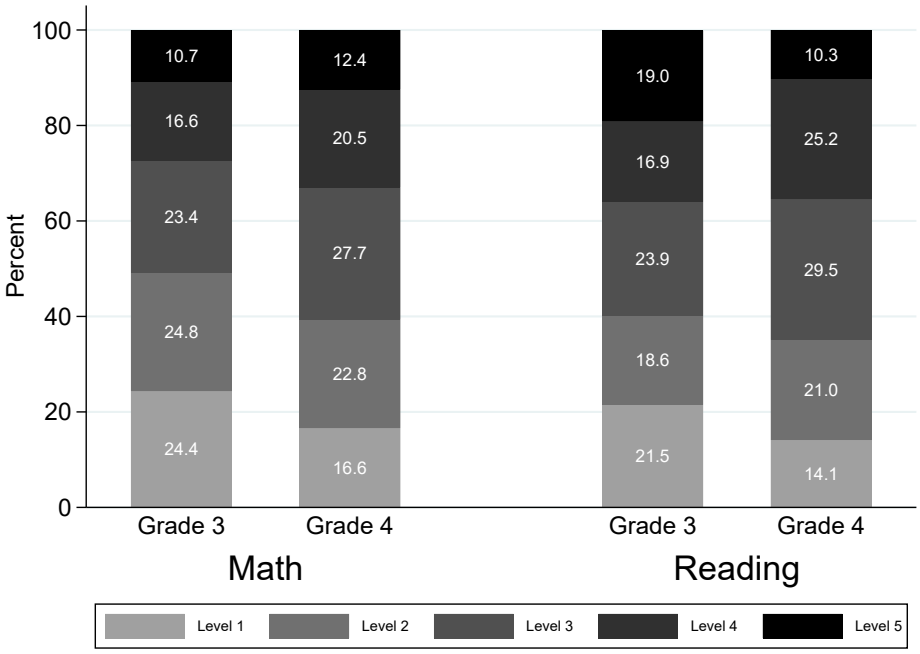
Notes: Panel A shows the fraction of grade 3 student answers to the question: "*What do you think, how good would your prospects of getting a good job be with the following school-leaving qualification?*" Panel B shows the fraction of grade 4 student answers to the question:"*Now matter how good you are in school: Which school degree would you like to have?*" Source: NEPS Wave 5 and 6.

Figure 3: Trends in school track transition rates



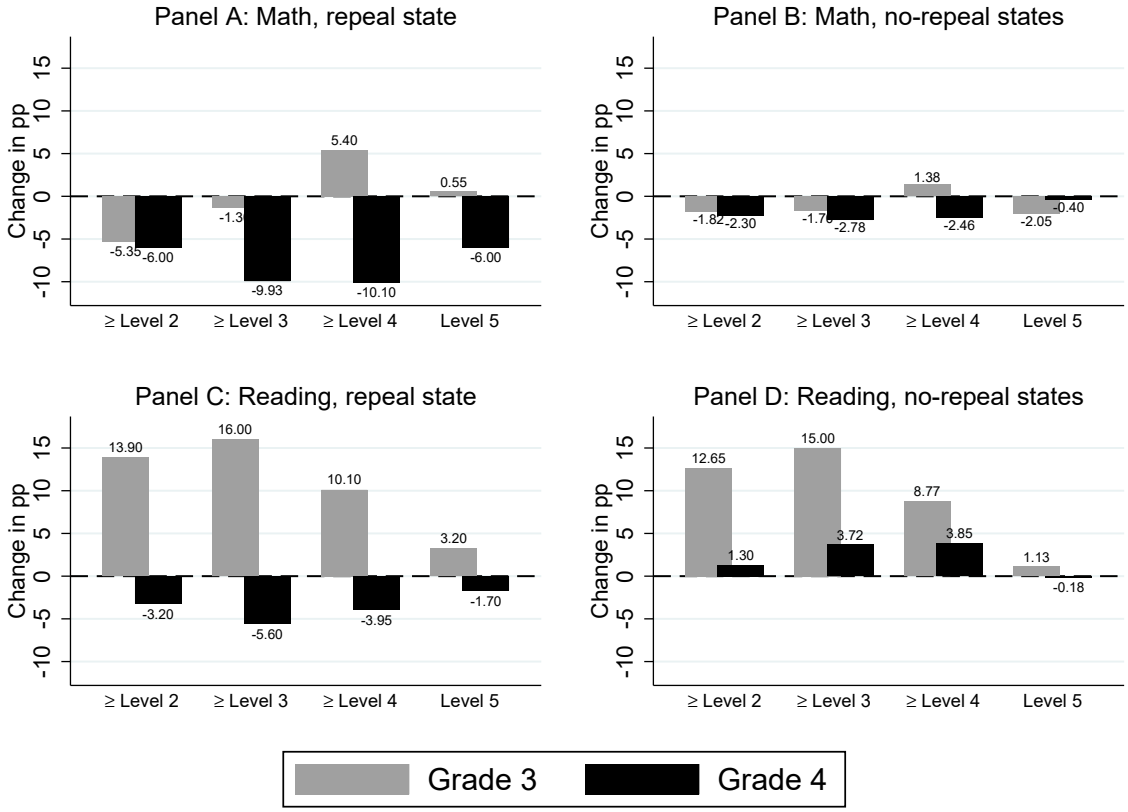
Notes: The graph plots transition rates for the basic and academic track over time. Transition rates for a particular year are calculated as the share of students entering a specific track in grade 5 in that year. Since tracking occurs in grade 7 in Berlin and Brandenburg, transition rates for these two states are calculated as the share of students entering a specific track in grade 7 in year $t+2$. The group of repeal states consists of Baden-Württemberg and Saxony-Anhalt. No-repeal states are Schleswig-Holstein, Hamburg, Lower Saxony, Bremen, Hesse, Rhineland-Palatinate, Bavaria, Saarland, Berlin, Brandenburg, Mecklenburg-Vorpommern, Saxony, and Thuringia. Own illustration based on data from the German Federal Statistical Office: Allgemeinbildende Schulen Fachserie 11, Reihe 1: `https://www.destatis.de/GPStatistik/receive/DESerie_serie_00000110`. (Retrieved: 11/5/2020)

Figure 4: Distribution of proficiency levels



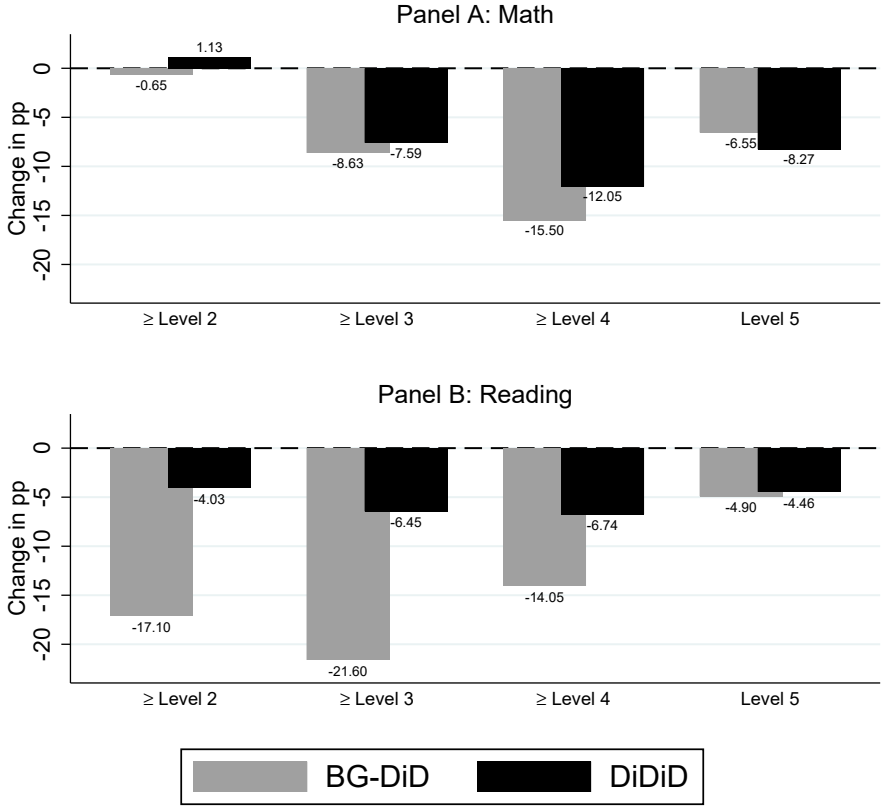
Notes: The graph shows the fraction of students by proficiency levels aggregated over the four states Schleswig-Holstein, Baden-Württemberg, Berlin, and Brandenburg for the school cohorts who are in grade 4 in the school years 2010/2011 and 2015/2016. See Table A.2 for details on the data sources.

Figure 5: Post-pre repeal differences in the share of students scoring at or above specific proficiency levels



Notes: The graph shows changes in the share of students at or above a proficiency level between post- and pre-repeal cohorts by proficiency levels for grade-levels 3 (grey) and 4 (black). Pre- and post-repeal cohorts refer to school cohorts in grade 4 in the school years 2010/2011 and 2015/2016, respectively. Panel A and C show changes for the repeal state (Baden-Württemberg) for math and reading, respectively. Panel B and D show changes for the no-repeal states (Schleswig-Holstein, Berlin, and Brandenburg) for math and reading, respectively. See Table A.2 for details on the data sources.

Figure 6: Estimated effects on the the share of students scoring at or above specific proficiency levels: Between-grade difference-in-differences and difference-in-difference-in-differences



Notes: The graph shows BG-DiD (grey) and DiDiD (black) estimates of the effect of free track choice on the share of students scoring at or above a proficiency level in math (Panel A) and reading (Panel B) based on equations (3) and (4), respectively. See Table A.2 for details on the data sources.

Table 1: Binding school track assignment at the end of primary school

School year	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17
Schleswig-Holstein (SH)							
Hamburg (HH)							
Lower Saxony (NI)							
Bremen (HB)							
North Rhine-Westphalia (NW)	✓						
Hesse (HE)							
Rhineland-Palatinate (RP)							
Baden-Württemberg (BW)	✓	✓					
Bavaria (BY)	✓	✓	✓	✓	✓	✓	✓
Saarland (SL)							
Berlin (BE)							
Brandenburg (BB)	✓	✓	✓	✓	✓	✓	✓
Mecklenburg-Vorpommern (MV)							
Saxony (SN)	✓	✓	✓	✓	✓	✓	✓
Saxony-Anhalt (ST)	✓	✓					
Thuringia (TH)	✓	✓	✓	✓	✓	✓	✓

Notes: The checkmark (✓) indicates that the school's recommendation for a secondary school track was binding in the respective school year. Source: State-specific school laws and www.kmk.org.

Table 2: Descriptive statistics: NAS

	Mean (1)	SD (2)	N (3)
Panel A: Student characteristics			
Repeal state	0.11	0.32	48,080
Male	0.47	0.50	48,080
Age at test in grade 4	10.43	0.49	48,047
Years in public childcare	3.64	0.77	37,438
Grade repeater	0.08	0.27	48,080
Grade skipper	0.01	0.11	48,080
Any learning disability	0.08	0.27	47,944
Mother's years of education	14.45	2.26	27,653
Fathers's years of education	14.83	2.66	25,393
Parents' highest ISEI	51.53	18.44	35,298
Books at home (in 5 categories)	3.43	1.19	43,700
Mothers's employment status			
Full-time employed	0.30	0.46	35,666
Part-time employed	0.49	0.50	35,666
Not employed	0.07	0.25	35,666
Other employment status	0.15	0.26	35,666
Father's employment status			
Full-time employed	0.86	0.35	33,275
Part-time employed	0.04	0.20	33,275
Not employed	0.02	0.15	33,275
Other employment status	0.07	0.26	33,275
Migration background	0.26	0.44	44,804
Non-native German speaker	0.18	0.39	45,078
Born in Germany	0.96	0.20	44,047
Panel B: School characteristics			
Private school	0.05	0.21	45,078
Municipality size (in 6 categories)	3.64	1.84	44,412
Total enrolment	273.19	149.31	44,684
Full-day status			
No full-day	0.55	0.50	48,080
Binding full-day	0.07	0.25	48,080
Partly binding full-day	0.06	0.23	48,080
Open binding full-day	0.33	0.47	48,080
Panel C: Class characteristics			
Weekly instruction hours in German	5.14	0.86	44,170
Weekly instruction hours in Math	6.03	0.46	44,397
Math teacher's working experience (in years)	21.23	12.42	40,157
German teacher's working experience (in years)	20.27	12.27	41,206
Share special-needs students	0.04	0.06	47,782

Notes: The sample consists of all students without special-needs status with at least one non-missing test score. Means and standard deviations are conditional on non-missing survey responses. Source: National Assessment Study.

Table 3: Descriptive statistics: NEPS

	Mean (1)	SD (2)	N (3)
Panel A: Student characteristics			
Track choice state	0.751	0.431	4,915
Male	0.480	0.500	4,915
Age at test in grade 4	9.747	0.379	4,915
Early school enrolment	0.048	0.214	4,197
Late school enrolment	0.052	0.221	4,197
Grade repeater	0.014	0.116	4,518
Premature birth	0.087	0.281	4,915
Any learning disability	0.165	0.371	4,915
Dyslexia	0.064	0.245	4,518
Dyscalculia	0.017	0.130	4,518
Mother's years of education	14.278	2.364	4,395
Fathers's years of education	14.329	2.435	3,967
Mother's ISEI	52.496	19.423	4,306
Father's ISEI	52.838	21.980	3,864
Books at home (in 6 categories)	4.158	1.264	4,403
Number of siblings	1.370	1.047	4,026
Migration background			
1. Generation migrant	0.021	0.144	4,516
2. Generation migrant	0.196	0.397	4,516
3. Generation migrant	0.113	0.317	4,516
Parents' marital status			
Married	0.825	0.380	4,517
Married but living apart	0.033	0.178	4,517
Divorced	0.048	0.215	4,517
Widowed	0.004	0.066	4,517
Single	0.087	0.282	4,517
Registered civil partnership	0.002	0.049	4,517
Panel B: School characteristics			
Private school	0.047	0.211	3,901
Total enrolment	278.684	139.342	4,290
Schools within 10km radius	8.081	8.429	4,155
Student-teacher ratio	0.071	0.022	4,182
Share teaching staff under 35 years	0.195	0.140	4,121
Share teaching staff 35 to under 45 years	0.278	0.150	4,136
Share teaching staff 45 to under 55 years	0.271	0.153	4,113
Share teaching staff 55 to under 65 years	0.280	0.174	4,083
Share teaching staff 65 years and older	0.007	0.043	3,474
Share special-needs students	0.029	0.033	4,100
Share low SES students	0.247	0.172	3,607
Share high SES students	0.202	0.142	3,566
Share migrant students	0.241	0.215	4,125
Panel C: Class characteristics			
Class size	21.875	3.527	4,629
Weekly instruction hours in German	5.923	0.790	3,389
Weekly instruction hours in Math	5.092	0.448	3,391
Two teaching staff per class 25% of the teaching time	0.363	0.481	4,670
Two teaching staff per class 50% of the teaching time	0.084	0.277	4,670
Two teaching staff per class 75% of the teaching time	0.031	0.173	4,670
Two teaching staff per class 100% of the teaching time	0.010	0.098	4,670

Notes: The sample consists of all students with non-missing grade 2 test scores and at least one non-missing grade 4 test score. Means and standard deviations are conditional on non-missing survey responses. Source: German National Educational Panel Study Starting Cohort 2.

Table 4: Estimated effects on 4th grade test scores: difference-in-differences

	(1)	(2)	(3)	(4)
Panel A: Math				
Track choice	-0.166***	-0.113***	-0.103***	-0.105***
	(0.048)	(0.036)	(0.036)	(0.035)
N	47,039	47,039	47,039	47,039
Adj. R^2	0.044	0.135	0.136	0.289
Panel B: Reading				
Track choice	-0.122***	-0.078**	-0.072**	-0.071**
	(0.044)	(0.035)	(0.035)	(0.035)
N	46,075	46,075	46,075	46,075
Adj. R^2	0.022	0.092	0.092	0.211
Panel C: Listening				
Track choice	-0.142***	-0.102***	-0.099***	-0.099***
	(0.046)	(0.036)	(0.036)	(0.036)
N	45,500	45,500	45,500	45,500
Adj. R^2	0.012	0.106	0.107	0.204
State & year FE	✓	✓	✓	✓
School composition controls		✓	✓	✓
School input controls			✓	✓
Student-level controls				✓

Notes: Each cell of the table reports results from a separate regression of the respective test score on an indicator for free track choice. All regressions include state, year, and test booklet fixed effects. School compositions controls include a categorical variable for the size of the municipality in which the school is located (in 6 categories); class-level averages of parents' years of education, highest ISEI, books at home; percent of students in class with a any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater. School input controls include school enrolment, a private school indicator, controls for the type of full-day offer in 4 categories (no full-day program, binding full-day, partly binding full-day, open full-day program), grade 4 instruction hours in math and German; years of experience of the German and math teacher. Student-level controls include indicators for any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater; years spent in public childcare, linear and quadratic age at test; mother's and father's highest years of education, highest ISES, country of birth in 5 categories (Germany, Poland, Russia, Turkey, other), work status in 4 categories (full-time, part-time, not employed, other), EGP class in 11 categories. Standard errors in parentheses allow for clustering at the school level. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 5: Specification check: lagged value-added

	Grade 1		Grade 2		
	(1)	(2)	(3)	(4)	(5)
Panel A: Average composite score					
Track choice	-0.200***	0.006	-0.122***	-0.015	-0.002
	(0.0441)	(0.040)	(0.031)	(0.037)	(0.030)
N	5,966	5,966	5,139	5,139	5,139
Adj. R^2	0.010	0.345	0.005	0.223	0.496
Panel B: Math					
Track choice	-0.155***	0.045	-0.158***	-0.008	0.002
	(0.045)	(0.046)	(0.045)	(0.047)	(0.039)
N	6,210	6,210	5,333	5,333	5,333
Adj. R^2	0.004	0.238	0.005	0.233	0.543
Panel C: Language (vocabulary and grammar)					
Track choice	-0.251***	-0.014			
	(0.053)	(0.052)			
N	6,203	6,203			
Adj. R^2	0.014	0.350			
Panel D: Reading					
Track choice			-0.134***	-0.054	-0.034
			(0.047)	(0.051)	(0.048)
N			5,283	5,283	5,283
adj. R^2			0.003	0.170	0.280
Panel E: Science					
Track choice	-0.183***	0.015			
	(0.050)	(0.046)			
N	6,221	6,221			
Adj. R^2	0.006	0.249			
Panel F: Cognition (non-verbal)					
Track choice			-0.062	0.017	0.027
			(0.038)	(0.049)	(0.045)
N			5,275	5,275	5,275
Adj. R^2			0.001	0.097	0.194
School & class controls		✓		✓	✓
Student controls		✓		✓	✓
Grade 1 test score controls					✓

Notes: Each cell of the table reports results from a separate regression of the respective test score on an indicator for free track choice. School and class controls include community size (in 7 categories), settlement structure (in 11 categories), school enrolment, number of schools within 10 km radius, a private school indicator, number of teachers at school, share of full-time teachers at school, share of teachers at the school in each of the following age categories: below 35, between 35 and below 45, between 45 and 55, between 55 and below 65, and 65 or older; class size, grade 2-3 instruction hours in math and German; the first principal component of the class teacher's qualifications grades, indicators for whether lessons are taught by more than one teacher, whether there is additional socio-educational staff or special educational needs staff for the class; the class level share of low SES students, high SES students, special needs students, and students with migration background. Student controls include sex, indicators for early or late schol enrolment, premature birth, Dyslexia, migration background, grade repetition, mother and father's years of education and ISEI, reported books at home, number of siblings, and marital status (in 5 categories). Grade 1 test score controls include cubic functions of math, language, and, science test scores. Standard errors in parentheses allow for clustering at the school level. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 6: Estimated effects on 4th grade test scores: lagged value-added

	(1)	(2)	(3)
Panel A: Math			
Track choice	-0.138***	-0.136***	-0.140***
	(0.043)	(0.031)	(0.034)
N	4,800	4,800	4,379
Adj. R^2	0.202	0.515	0.542
Panel B: Reading			
Track choice	-0.066	-0.062*	-0.071**
	(0.042)	(0.034)	(0.035)
N	4,798	4,798	4,378
Adj. R^2	0.195	0.429	0.475
Panel C: Orthography			
Track choice	-0.211***	-0.205***	-0.199***
	(0.044)	(0.037)	(0.038)
N	4,533	4,533	4,130
Adj. R^2	0.242	0.567	0.565
School & class controls	✓	✓	✓
Individual controls	✓	✓	✓
Grade 2 test score controls		✓	✓
Grade 1 test score controls			✓

Notes: Each cell of the table reports results from a separate regression of the respective test score on an indicator for free track choice. School and class controls include community size (in 7 categories), settlement structure (in 11 categories), school enrolment, number of schools within 10 km radius, a private school indicator, number of teachers at school, share of full-time teachers at school, share of teachers at the school in each of the following age categories: below 35, between 35 and below 45, between 45 and 55, between 55 and below 65, and 65 or older; class size, grade 2-3 instruction hours in math and German; the first principal component of the class teacher's qualifications grades, indicators for whether lessons are taught by more than one teacher, whether there is additional socio-educational or special educational needs staff in the class; the class level share of low SES students, high SES students, special needs students, and students with migration background. Student controls include sex, indicators for early or late schol enrolment, premature birth, Dyslexia, migration background, grade repetition, mother and father's years of education and ISEI, reported books at home, number of siblings, and marital status (in 5 categories). Grade 2 test score controls include cubic functions of math, reading, and cognition test scores, and teacher assessments of students' social skills, persistence and ability to concentrate, language skills in German, written language skills, science knowledge, and mathematical skills (each measured in 5 categories). Grade 1 test score controls include cubic functions of math, language, and, science test scores. Standard errors in parentheses allow for clustering at the school level. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 7: Estimated effects on well-being and intrinsic motivation: lagged value-added

	Satisfaction w/		Anxiety about		
	life	health	grades	own future	Intrinsic motivation
	(1)	(2)	(3)	(4)	(5)
Panel A: Grade 3					
Track choice	-0.029	-0.003			0.008
	(0.044)	(0.042)			(0.056)
N	4,724	4,716			3,237
Adj. R^2	0.013	0.018			0.117
Panel B: Grade 4					
Track choice	0.146***	0.033	-0.114**	-0.166***	0.129**
	(0.044)	(0.042)	(0.048)	(0.045)	(0.063)
N	4,698	4,707	3,116	4,663	4,663
Adj. R^2	0.017	0.020	0.129	0.127	0.022
Panel C: Grade 5					
Track choice	0.114**	0.049			0.133**
	(0.051)	(0.051)			(0.059)
N	2,955	2,976			2,291
Adj. R^2	0.023	0.010			0.086
School & class controls	✓	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓	✓
Grade 2 test score controls	✓	✓	✓	✓	✓

Notes: Each cell of the table reports results from a separate regression of the variable in the column header on an indicator for free track choice. School and class controls include community size (in 7 categories), settlement structure (in 11 categories), school enrolment, number of schools within 10 km radius, a private school indicator, number of teachers at school, share of full-time teachers at school, share of teachers at the school in each of the following age categories: below 35, between 35 and below 45, between 45 and 55, between 55 and below 65, and 65 or older; class size, grade 2-3 instruction hours in math and German; the first principal component of the class teacher's qualifications grades, indicators for whether lessons are taught by more than one teacher, whether there is additional socio-educational or special educational needs staff in the class; ; the class level share of low SES students, high SES students, special needs students, and students with migration background. Student controls include sex, indicators for early or late schol enrolment, premature birth, Dyslexia, migration background, grade repetition, mother and father's years of education and ISEI, reported books at home, number of siblings, and marital status (in 5 categories). Grade 2 test score controls include cubic functions of math, reading, and cognition test scores, and teacher assessments of students' social skills, persistence and ability to concentrate, language skills in German, written language skills, science knowledge, and mathematical skills (each measured in 5 categories). Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 8: Estimated effects on self-study and parental behavior: lagged valued-added

			Frequency of parental help w/		
	Daily self-study (in min)	Stricter parenting style	homework	other school exercises	Private tutoring
	(1)	(2)	(3)	(4)	(5)
Track choice	-12.773***	0.139***	0.043	-0.036	0.004
	(1.664)	(0.048)	(0.052)	(0.045)	(0.011)
School & class controls	✓	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓	✓
Grade 2 test score controls	✓	✓	✓	✓	✓
Mean dependent variable	50.785				
N	3,459	4,089	3,256	2,162	3,475
Adj. R^2	0.214	0.054	0.122	0.118	0.119

Notes: Each cell of the table reports results from a separate regression of the respective variable in the column header on an indicator for free track choice. School and class controls include community size (in 7 categories), settlement structure (in 11 categories), school enrolment, number of schools within 10 km radius, a private school indicator, number of teachers at school, share of full-time teachers at school, share of teachers at the school in each of the following age categories: below 35, between 35 and below 45, between 45 and 55, between 55 and below 65, and 65 or older; class size, grade 2-3 instruction hours in math and German; the first principal component of the class teacher's qualifications grades, indicators for whether lessons are taught by more than one teacher, whether there is additional socio-educational or special educational needs staff in the class; ; the class level share of low SES students, high SES students, special needs students, and students with migration background. Student controls include sex, indicators for early or late schol enrolment, premature birth, Dyslexia, migration background, grade repetition, mother and father's years of education and ISEI, reported books at home, number of siblings, and marital status (in 5 categories). Grade 2 test score controls include cubic functions of math, reading, and cognition test scores, and teacher assessments of students' social skills, persistence and ability to concentrate, language skills in German, written language skills, science knowledge, and mathematical skills (each measured in 5 categories). Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 9: Effect heterogeneity: lagged valued-added

		4th grade test scores		
	Daily self-study (in min)	Math	Reading	Orthography
	(1)	(2)	(3)	(4)
Panel A: By gender				
Track choice * female	-14.887***	-0.170***	-0.122***	-0.211***
	(2.074)	(0.040)	(0.040)	(0.042)
Track choice * male	-10.469***	-0.101***	0.000	-0.200***
	(1.939)	(0.036)	(0.043)	(0.043)
N	3,459	4,800	4,798	4,533
Adj. R^2	0.215	0.515	0.430	0.567
Panel B: By lagged grade 2 test scores				
Track choice * bottom quartile lagged test scores	-18.359***	-0.148**	-0.108*	-0.305***
	(3.241)	(0.058)	(0.058)	(0.055)
Track choice * second quartile lagged test scores	-11.890***	-0.137**	-0.075	-0.142***
	(2.576)	(0.054)	(0.057)	(0.052)
Track choice * third quartile lagged test scores	-14.503***	-0.130***	-0.083	-0.166***
	(2.732)	(0.048)	(0.053)	(0.049)
Track choice * top quartile lagged test scores	-8.966***	-0.132***	0.001	-0.245***
	(2.125)	(0.044)	(0.051)	(0.054)
N	3,459	4,800	4,798	4,533
Adj. R^2	0.215	0.514	0.429	0.568
Panel C: By highest school degree of parents				
Track choice * academic degree	-12.735***	-0.145***	-0.089**	-0.192***
	(1.767)	(0.035)	(0.039)	(0.039)
Track choice * less than academic degree	-12.858***	-0.123***	-0.022	-0.227***
	(2.089)	(0.038)	(0.041)	(0.044)
N	3,459	4,800	4,798	4,533
Adj. R^2	0.213	0.515	0.430	0.567
School & class controls	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓
Grade 2 test score controls	✓	✓	✓	✓

Notes: Each column within a panel reports results from a separate regression of the outcome in the column header on the variables in the leftmost column. School and class controls include community size (in 7 categories), settlement structure (in 11 categories), school enrolment, number of schools within 10 km radius, a private school indicator, number of teachers at school, share of full-time teachers at school, share of teachers at the school in each of the following age categories: below 35, between 35 and below 45, between 45 and 55, between 55 and below 65, and 65 or older; class size, grade 2-3 instruction hours in math and German; the first principal component of the class teacher's qualifications grades, indicators for whether lessons are taught by more than one teacher, whether there is additional socio-educational or special educational needs staff in the class; ; the class level share of low SES students, high SES students, special needs students, and students with migration background. Student controls include sex, indicators for early or late school enrolment, premature birth, Dyslexia, migration background, grade repetition, mother and father's years of education and ISEI, reported books at home, number of siblings, and marital status (in 5 categories). Grade 2 test score controls include cubic functions of math, reading, and cognition test scores, and teacher assessments of students' social skills, persistence and ability to concentrate, language skills in German, written language skills, science knowledge, and mathematical skills (each measured in 5 categories).

Appendix

A Additional figures and tables

Figure A.1: Distribution of test dates by grade and subject in the NEPS

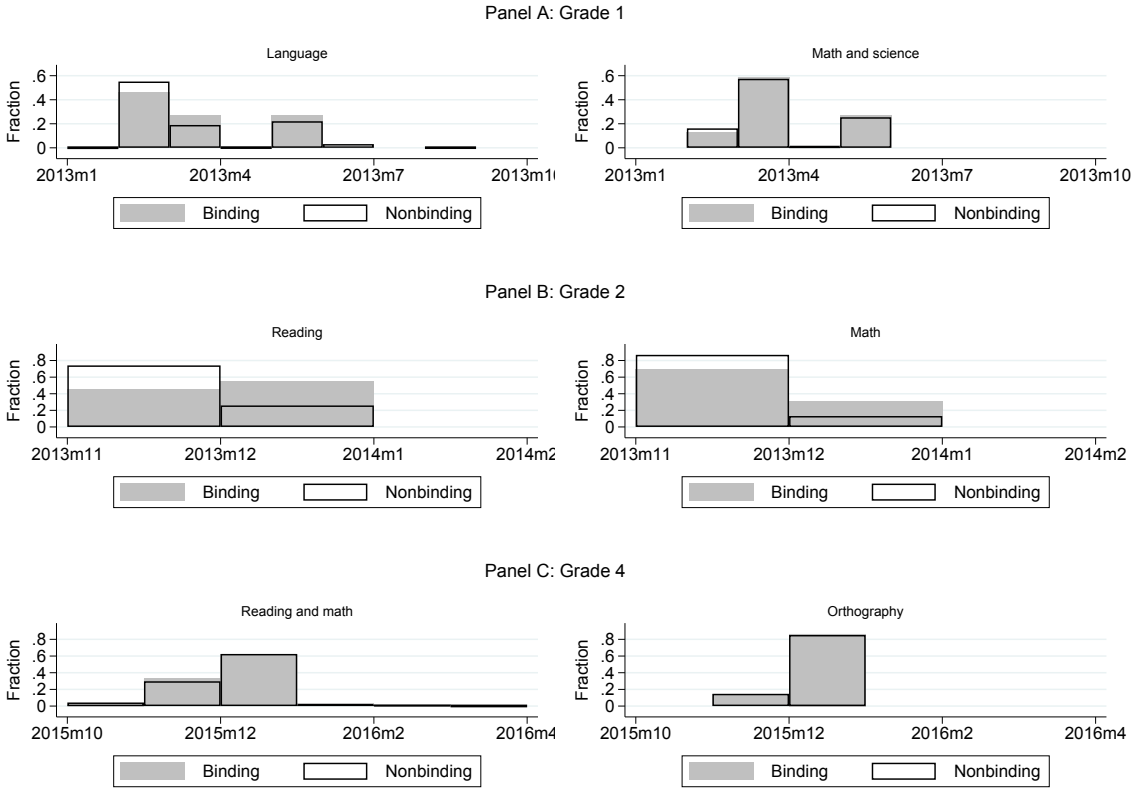
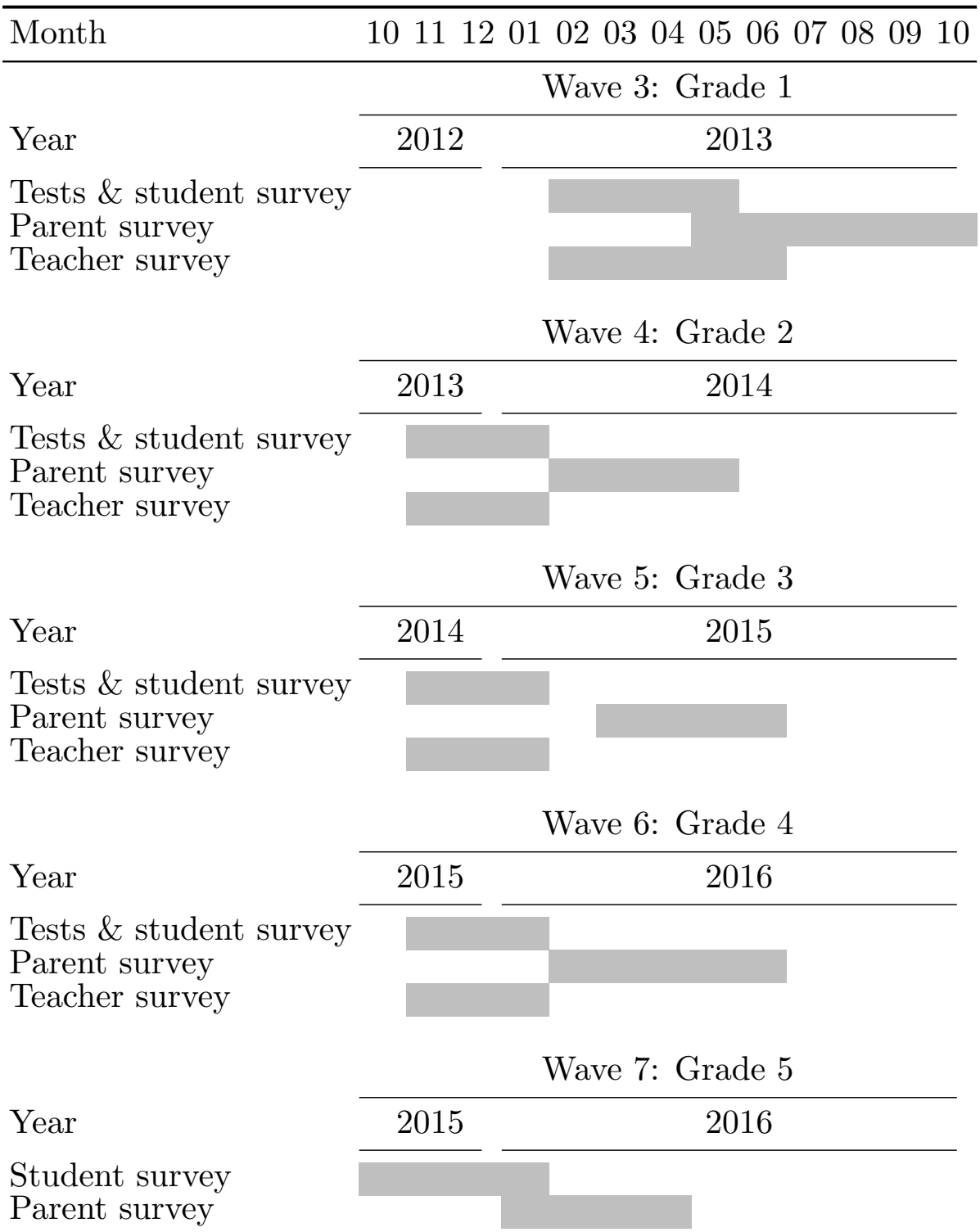


Table A.1: Structure of NEPS data



Notes: They grey cells indicate the months during which the respective information was collected. Own illustration based on: `https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/8-0-0/SC2_Studien_W1-8.pdf`.

Table A.2: State-level data sources

Data	Year	Source
Vera 3 (proficiency levels in grade 3)		
Schleswig-Holstein	2010	IQSH (received via mail from IQSH on 28.10.2020)
Schleswig-Holstein	2015	IQSH https://www.schleswig-holstein.de (last accessed 20.10.2020)
Baden Württemberg	2010	LS BW www.ls-bw.de (last accessed 10.10.2020)
Baden Württemberg	2015	LS BW www.ls-bw.de (last accessed 10.10.2020)
Berlin & Brandenburg	2010	ISQ-BB www.isq-bb.de (last accessed 10.10.2020)
Berlin & Brandenburg	2015	ISQ-BB www.isq-bb.de (last accessed 10.10.2020)
IQB (proficiency levels in grade 4)		
Schleswig-Holstein, Baden Württemberg, Berlin & Brandenburg	2011	IQB Berlin www.iqb.hu-berlin.de/ (last accessed 05.4.2020)
Schleswig-Holstein, Baden Württemberg, Berlin & Brandenburg	2016	IQB Berlin www.iqb.hu-berlin.de/ (last accessed 05.4.2020)

This table shows the data source for the state-level analyses which are published online.

Table A.3: School year starting dates by year

State	2012	2013	2015
Schleswig-Holstein	Aug. 04	Aug. 03	Aug. 29.
Hamburg	Aug. 01.	Jul. 31	Aug. 26.
Lower Saxony	Aug. 31.	Aug. 07	Sep. 02.
Bremen	Aug. 31.	Aug. 07	Sep. 02.
North Rhine-Westphalia	Aug. 31.	Sep. 03	Aug. 11.
Hesse	Aug. 10.	Aug. 16	Sep. 04.
Rhineland-Palatinate	Aug. 10.	Aug. 16	Sep. 04.
Baden-Württemberg	Sep. 08.	Sep. 07	Sep. 12.
Bavaria	Sep. 12.	Sep. 11	Sep. 14.
Saarland	Aug. 14.	Aug. 16	Sep. 05.
Berlin	Aug. 13.	Aug. 02	Aug. 28.
Brandenburg	Aug. 03.	Aug. 03	Aug. 28.
Mecklenburg-Vorpommern	Aug. 04.	Aug. 03	Aug. 29.
Saxony	Aug. 31.	Aug. 28	Aug. 21.
Saxony-Anhalt	Sep. 05.	Aug. 28	Aug. 26.
Thuringia	Aug. 31.	Aug. 23	Aug. 21.

Source: `https://www.schulferien.org/`(last accessed 15.01.2020)

Table A.4: Estimated effects on academic track recommendation: difference-in-differences

	Academic track recommendation				
	(1)	(2)	(3)	(4)	(5)
Track choice	-0.054**	-0.031*	-0.034*	-0.035**	-0.010
	(0.022)	(0.018)	(0.018)	(0.018)	(0.018)
State & year FE	✓	✓	✓	✓	✓
School composition controls		✓	✓	✓	✓
School input controls			✓	✓	✓
Student-level controls				✓	✓
Test score controls					✓
N	37,188	37,188	37,188	37,188	36,033
Adj. R^2	0.012	0.068	0.069	0.233	0.394

Notes: Each cell of the table reports results from a separate regression of an indicator for a recommendation for the academic track on an indicator for free track choice. All regressions include state, year, and test booklet fixed effects. School compositions controls include a categorical variable for the size of the municipality in which the school is located (in 6 categories); class-level averages of parents' years of education, highest ISEI, books at home; percent of students in class with a any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater. School input controls include school enrolment, a private school indicator, controls for the type of full-day offer in 4 categories (no full-day program, binding full-day, partly binding full-day, open full-day program), grade 4 instruction hours in math and German; years of experience of the German and math teacher. Student-level controls include indicators for any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater; years spent in public childcare, linear and quadratic age at test; mother's and father's highest years of education, highest ISES, country of birth in 5 categories (Germany, Poland, Russia, Turkey, other), work status in 4 categories (full-time, part-time, not employed, other), EGP class in 11 categories. Test score controls include test scores in math, reading, and listening. Standard errors in parentheses allow for clustering at the school level. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.5: Robustness check of estimated effects on well-being and intrinsic motivation in grade 5: lagged value-added

	Satisfaction w/					
	life		health		Intrinsic motivation	
	(1)	(2)	(3)	(4)	(5)	(6)
Track choice	0.114**	0.105*	0.049	0.026	0.133**	0.135**
	(0.051)	(0.057)	(0.051)	(0.060)	(0.059)	(0.062)
School & class controls	✓	✓	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓	✓	✓
Grade 2 test score controls	✓	✓	✓	✓	✓	✓
School track controls		✓		✓		✓
Parental track preference controls		✓		✓		✓
N	2,955	2,674	2,976	2,691	2,291	2,186
Adj. R^2	0.023	0.013	0.010	0.003	0.086	0.094

Notes: Each cell of the table reports results from a separate regression of the respective test score on an indicator for free track choice. School and class controls include community size (in 7 categories), settlement structure (in 11 categories), school enrolment, number of schools within 10 km radius, a private school indicator, number of teachers at school, share of full-time teachers at school, share of teachers at the school in each of the following age categories: below 35, between 35 and below 45, between 45 and 55, between 55 and below 65, and 65 or older; class size, grade 2-3 instruction hours in math and German; the first principal component of the class teacher's qualifications grades, indicators for whether lessons are taught by more than one teacher, whether there is additional socio-educational or special educational needs staff in the class; the class level share of low SES students, high SES students, special needs students, and students with migration background. Student controls include sex, indicators for early or late schol enrolment, premature birth, Dyslexia, migration background, grade repetition, mother and father's years of education and ISEI, reported books at home, number of siblings, and marital status (in 5 categories). Grade 2 test score controls include cubic functions of math, reading, and cognition test scores, and teacher assessments of students' social skills, persistence and ability to concentrate, language skills in German, written language skills, science knowledge, and mathematical skills (each measured in 5 categories). School track controls include separate indicators the attended secondary school track. Parental track preference controls include a control for the stated track preference of the parents in grade 2. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.6: Estimated effects on private tutoring: difference-in-differences

	Private tutoring in grade 4			
	(1)	(2)	(3)	(4)
Track choice	0.039***	0.029***	0.030***	0.028***
	(0.011)	(0.010)	(0.010)	(0.010)
State & year FE	✓	✓	✓	✓
School composition controls		✓	✓	✓
School input controls			✓	✓
Student level controls				✓
N	36,054	36,054	36,054	36,054
adj. R^2	0.024	0.035	0.036	0.099

Notes: Each column within a panel reports results from a separate regression of an indicator for private tutoring on an indicator for free track choice. All regressions include state, year, and test booklet fixed effects. School compositions controls include a categorical variable for the size of the municipality in which the school is located (in 6 categories); class-level averages of parents' years of education, highest ISEI, books at home; percent of students in class with a any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater. School input controls include school enrolment, a private school indicator, controls for the type of full-day offer in 4 categories (no full-day program, binding full-day, partly binding full-day, open full-day program), grade 4 instruction hours in math and German; years of experience of the German and math teacher. Student-level controls include indicators for any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater; years spent in public childcare, linear and quadratic age at test; mother's and father's highest years of education, highest ISES, country of birth in 5 categories (Germany, Poland, Russia, Turkey, other), work status in 4 categories (full-time, part-time, not employed, other), EGP class in 11 categories.

Table A.7: Effect heterogeneity: difference-in-differences

	4th grade test scores		
	Math	Reading	Listening
	(1)	(2)	(3)
Panel A: By gender			
Track choice * female	-0.114***	-0.076**	-0.101***
	(0.036)	(0.037)	(0.037)
Track choice * male	-0.106***	-0.072*	-0.092**
	(0.037)	(0.038)	(0.038)
N	47,039	46,075	45,500
adj. R^2	0.282	0.205	0.201
Panel B: By highest school degree of parents			
Track choice * academic degree	-0.069	-0.062	-0.105**
	(0.045)	(0.044)	(0.047)
Track choice * less than academic degree	-0.074*	-0.061	-0.102**
	(0.042)	(0.041)	(0.044)
N	27,084	26,723	26,280
adj. R^2	0.274	0.190	0.191
State & year FE	✓	✓	✓
School composition controls	✓	✓	✓
School input controls	✓	✓	✓
Student level controls	✓	✓	✓

Notes: Each column within a panel reports results from a separate regression of the outcome in the column header on the variables in the leftmost column.All regressions include state, year, and test booklet fixed effects. School compositions controls include a categorical variable for the size of the municipality in which the school is located (in 6 categories); class-level averages of parents' years of education, highest ISEI, books at home; percent of students in class with a any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater. School input controls include school enrolment, a private school indicator, controls for the type of full-day offer in 4 categories (no full-day program, binding full-day, partly binding full-day, open full-day program), grade 4 instruction hours in math and German; years of experience of the German and math teacher. Student-level controls include indicators for any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater; years spent in public childcare, linear and quadratic age at test; mother's and father's highest years of education, highest ISES, country of birth in 5 categories (Germany, Poland, Russia, Turkey, other), work status in 4 categories (full-time, part-time, not employed, other), EGP class in 11 categories.

B Robustness checks

B.1 Controlling for compositional differences in the difference-in-differences

Here we check whether time varying compositional differences within states over time bias our DiD results. This could happen, for example, through non-linearities in the relationship between the class composition and student achievement. Table B.1 therefore shows results for specifications with different ways to control for compositional differences. The estimates in column 1 are taken from our preferred specification, that is the model reported in column 4 of Table 4. The first variation on this specification in column 2 adds quadratic terms for all demographic class-level controls. The next specification in column 3 allows the class-level controls to differ by year and state. As a final check, we report in column 4 results from a propensity score matching DiD estimator, following Heckman et al. (1997) and Blundell et al. (2004).[55] Our results are insensitive to these robustness checks, suggesting that our parametric linear difference-in-differences estimator is effective in eliminating bias from compositional differences.

B.2 Inference

In our student-level analyses, we estimate standard errors by clustering at the school level. This is done to combat the false precision that arises if outcomes are correlated for children within the same school. For example, due to the proverbial dog barking outside the class room on the day of testing. To be more conservative, Table B.2 provides p-values for our main test score results in Tables 4 and 6 based on standard errors clustered at state-year level (column 3), the federal state level (column 4), and, since the number states is with 15 low, p-values based on the wild-cluster bootstrap procedure with varying weights (in columns 4-9), testing under the null (columns 4, 6, and 8) and under the alternative hypothesis (in columns 5, 7, and 9).[56] As can be seen, our estimates remain statistically significant using all procedures, with the exception of the reading results which sometimes turn insignificant.

[55]Essentially, we extend propensity score matching by estimating two propensity scores—one for binding track assignment and one for time period. Propensity scores are estimated using the same class-level compositional controls as in Table 4. We then create a matched sample based on the two propensity scores. That is, we match to the group of students in repeal states in the post-repeal period three separate control groups (repeal states before the repeal and no-repeal states before and after the repeal). These four groups are then used in a DiD design to control for time-invariant state fixed effects. This ensures that the class-composition distribution is the same in the four cells defined by binding track assignment and time. We use the Stata-program *diff* by Villa (2016). In addition to the variables specified in the estimation of the propensity score, we also include the same variables as those in our preferred specification in column 4 of Table 4.

[56]We use the Stata-program *BOOTTEST* by Roodman et al. (2018) for all wild-cluster bootstrap estimations with 999 replication.

B.3 Jackknife leave-one-state-out results

Another concern is that our estimated effects are driven by conditions specific to one particular state. To assuage this concern, we present in Figure B.1 estimated effects on grade 4 test scores based on the lagged value added specification with full controls dropping one state at a time. In all cases, the results are very similar or even more pronounced compared to when using the full sample, and for math and orthography all estimated coefficients are significant at the 1% level. We take the robustness of the negative track choice coefficient across samples to be compelling evidence that our estimated impacts are not driven by any particular state, but rather reflect a general pattern. Unfortunately, we can not perform the same exercise with the DiD design as the group of repeal states only consists of two states and we are legally prohibited from comparing groups of states that consist of only one state.

B.4 Specifications with school factors

To assuage concerns that our lagged valued added results are confounded by differences in the schooling environment between choice and no-choice states, Table B.3 presents results with extensive school environment controls. Since the number of variables describing the schooling environment in the NEPS is vast, we reduce the dimensionality of this information by forming factors based on principal components factor analysis and only include factors for each battery of items with eigenvalues greater than one.[57] Column 2 in Table B.3 adds to our preferred specification controls for factors describing teachers' behavior (e.g., time devoted to certain classroom activities or teaching styles). In column 3, we control for factors describing the school (e.g., the quality of school facilities) and in column 4 we control for a factors describing parents' involvement with the school. In Column 5, we control for all factors simultaneously. We consider these models to be "over controlling" since many of these factors are potentially endogenous to students' study effort, but view the similarity of these estimates to that of our preferred models as strong evidence that our main effects are not driven by other differences in the schooling environment besides binding track assignment.

B.5 Cross-sectional estimates based on the NAS wave 2016

The NEPS and the second wave of the NAS tested the same cohort of students—those who made the transition to a secondary school track in 2016. Moreover, the lagged value-added results in Table 6 suggest that adding prior test score controls does not substantially affect point estimates in a cross-sectional comparison of states with and without track choice when one regression adjusts for socio-demographic differences. Hence, to check the robustness of the NEPS results, we repeat the same analysis with the second wave of

[57] The factors and the items used to obtain them are reported in Sections D.4.1-D.4.3.

the NAS for which prior test scores are not available. The results are shown in columns 1-4 in Table B.4. In the specification with the full set of controls in column 4, free track choice is predicted to reduce math and reading test scores by 0.13 and 0.06 standard deviations, respectively.[58] Both effects are significant at the 1% significance level and similar to those based on the NEPS data. This limits concerns that the NEPS results are driven by selective panel attrition in grade 4, which is not a problem in the NAS since test participation was mandatory for all 4th grade students in the sampled public schools. However, we do not find an effect on listening test scores in the specification with full controls, which was not tested in the NEPS.

[58] We use the same control variables as for the DiD approach in Table 4, but exclude state and year fixed effects.

Table B.1: Alternative specifications: difference-in-differences

	DiD			Matching DiD
	(1)	(2)	(3)	(4)
Panel A: Math				
Track choice	-0.105***	-0.108***	-0.094**	-0.112***
	(0.035)	(0.035)	(0.044)	(0.016)
N	47,039	47,039	47,039	47,035
Adj. R^2	0.289	0.290	0.297	
Panel B: Reading				
Track choice	-0.071**	-0.077**	-0.111***	-0.060***
	(0.035)	(0.035)	(0.041)	(0.017)
N	46,075	46,075	46,075	46,073
Adj. R^2	0.211	0.211	0.216	
Panel C: Listening				
Listening Track choice	-0.099***	-0.103***	-0.090**	-0.100***
	(0.036)	(0.036)	(0.043)	(0.017)
N	45,500	45,500	45,500	45,498
Adj. R^2	0.211	0.211	0.216	
State & year FE	✓	✓	✓	✓
School composition controls	✓	✓	✓	✓
Quadratic composition controls		✓		
Interacted school composition controls			✓	
School input controls	✓	✓	✓	✓
Student-level controls	✓	✓	✓	✓

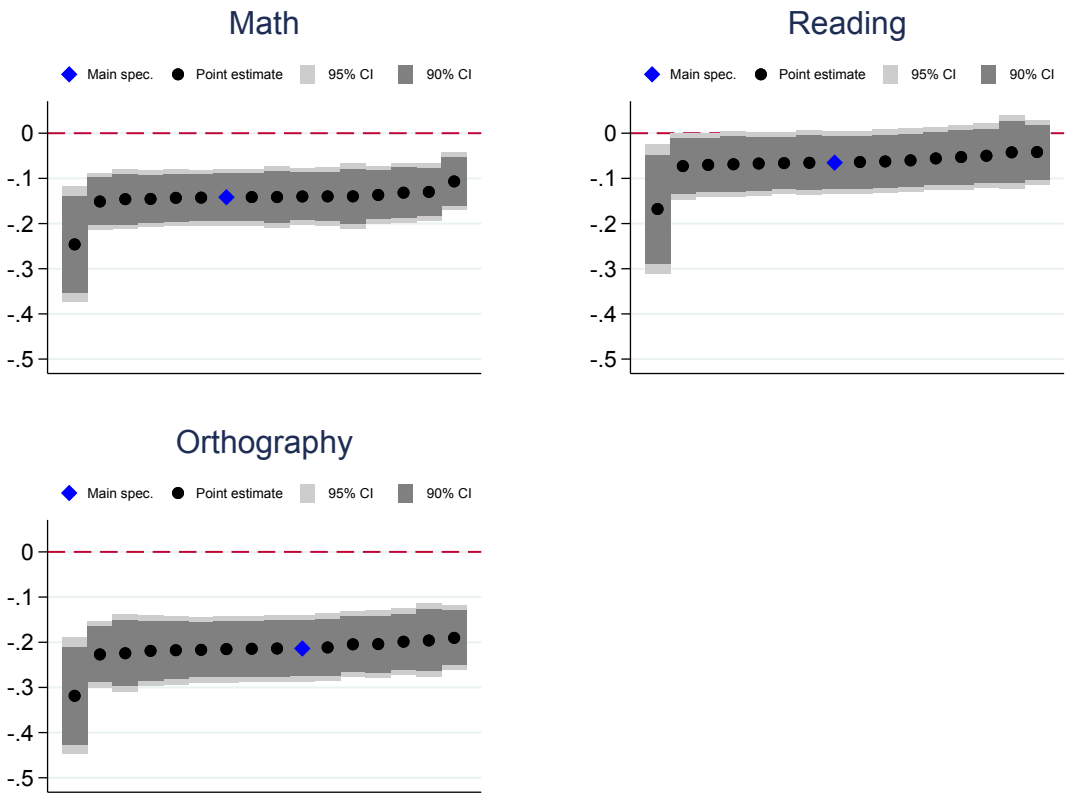
Notes: All regressions include state, year, and test booklet fixed effects. School compositions controls include a categorical variable for the size of the municipality in which the school is located (in 6 categories); class-level averages of parents' years of education, highest ISEI, books at home; percent of students in class with a any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater. School input controls include school enrolment, a private school indicator, controls for the type of full-day offer in 4 categories (no full-day program, binding full-day, partly binding full-day, open full-day program), grade 4 instruction hours in math and German; years of experience of the German and math teacher. Student-level controls include indicators for any learning disability, special-needs status, migration background, male, non-native German speakers, grade repeater; years spent in public childcare, linear and quadratic age at test; mother's and father's highest years of education, highest ISES, country of birth in 5 categories (Germany, Poland, Russia, Turkey, other), work status in 4 categories (full-time, part-time, not employed, other), EGP class in 11 categories. Column 2 additional includes quadratic of all school composition controls. Column 3 includes interactions of all school composition controls with the year and state. Column 4 presents results for a propensity score matching DiD estimator, where all school composition controls are used to estimate the propensity score. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table B.2: Alternative ways of statistical inference

	Cluster			**Wild cluster bootstrap**					
	School	State × year	State	Webb		Mammen		Rademacher	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Difference-in-differences									
Math	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105	-0.105
p-value	0.002	0.000	0.003	0.084	0.003	0.070	0.001	0.081	0.005
Reading	-0.075	-0.075	-0.075	-0.075	-0.075	-0.075	-0.075	-0.075	-0.075
p-value	0.029	0.041	0.158	0.131	0.241	0.069	0.182	0.148	0.224
Listening	-0.101	-0.101	-0.101	-0.101	-0.101	-0.101	-0.101	-0.101	-0.101
p-value	0.005	0.000	0.002	0.102	0.020	0.041	0.002	0.072	0.016
Panel B: Lagged valued-added									
Math	-0.141	-0.141	-0.141	-0.141	-0.141	-0.141	-0.141	-0.141	-0.141
p-value	0.000		0.012	0.018	0.067	0.007	0.064	0.019	0.052
Reading	-0.071	-0.071	-0.071	-0.071	-0.071	-0.071	-0.071	-0.071	-0.071
p-value	0.044		0.121	0.078	0.169	0.074	0.154	0.061	0.166
Orthography	-0.199	-0.199	-0.199	-0.199	-0.199	-0.199	-0.199	-0.199	-0.199
p-value	0.000		0.000	0.005	0.000	0.007	0.000	0.001	0.000

Notes: The table displays free track choice coefficients and their p-values based on alternative procedures for inference for the main DiD and lagged valued-added specifications. Panel A is based on the specification in column 4 in Table 4. Panel B is based on the specification in column 2 in Table 6. Columns 1-3 implement clustered standard errors, where the level of clustering is the school, state-year, and state. The p-values in columns 4-9 are based on wild cluster bootstrap procedures with states as clusters and varying weights. Testing is one-sided under the null hypothesis in columns 4, 6, and 8. Testing is one-sided under the alternative hypothesis in columns 5, 7, and 9. All estimations are performed using the user-written Stata-program BOOTTEST (Roodman et al., 2018) with 999 bootstrap iterations.

Figure B.1: Jackknife leave-one-state-out estimates for 4th grade test scores: lagged valued-added



Notes: This figures plots estimates, and 90% and 95% confidence intervals for the track choice coefficient in equation (2) for the specification in column 2 in Table 6 if one drops any state. The estimates are ordered by coefficient size. The diamonds correspond to estimates for the full sample with no state excluded. Source: NEPS.

Table B.3: Robustness check with additional controls: Main results lagged value-added

	(1)	(2)	(3)	(4)	(5)
Panel A: Math					
Track choice	-0.136***	-0.137***	-0.127***	-0.136***	-0.133***
	(0.031)	(0.034)	(0.033)	(0.033)	(0.035)
N	4,800	4,800	4,800	4,800	4,800
Adj. R^2	0.515	0.514	0.515	0.514	0.514
Panel B: Reading					
Track choice	-0.062*	-0.065*	-0.069*	-0.065*	-0.079**
	(0.035)	(0.037)	(0.036)	(0.036)	(0.037)
N	4,798	4,798	4,798	4,798	4,798
Adj. R^2	0.429	0.431	0.430	0.429	0.432
Panel C: Orthography					
Track choice	-0.205***	-0.186***	-0.200***	-0.205***	-0.192***
	(0.037)	(0.039)	(0.037)	(0.037)	(0.038)
N	4,533	4,533	4,533	4,533	4,533
Adj. R^2	0.567	0.569	0.568	0.567	0.570
School & class controls	✓	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓	✓
Grade 2 test score controls	✓	✓	✓	✓	✓
Teacher PCA controls		✓			✓
School environment PCA controls			✓		✓
Parent PCA controls				✓	✓

Notes: Each cell of the table reports results from a separate regression of the respective test score on an indicator for free track choice. School and class controls include community size (in 7 categories), settlement structure (in 11 categories), school enrolment, number of schools within 10 km radius, a private school indicator, number of teachers at school, share of full-time teachers at school, share of teachers at the school in each of the following age categories: below 35, between 35 and below 45, between 45 and 55, between 55 and below 65, and 65 or older; class size, grade 2-3 instruction hours in math and German; the first principal component of the class teacher's qualifications grades, indicators for whether lessons are taught by more than one teacher, whether there is additional socio-educational or special educational needs staff in the class. Student controls include sex, indicators for early or late schol enrolment, premature birth, Dyslexia, migration background, grade repetition, mother and father's years of education and ISEI, reported books at home, number of siblings, and marital status (in 5 categories). Grade 2 test score controls include cubic functions of math, reading, and cognition test scores, and teacher assessments of students' social skills, persistence and ability to concentrate, language skills in German, written language skills, science knowledge, and mathematical skills (each measured in 5 categories). The PCA controls are all factors with eigenvalue greater than one reported in Appendix D.4.1-D.4.3. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table B.4: Robustness: Cross-state comparison based on NAS wave 2016

	(1)	(2)	(3)	(4)
Panel A: Math				
Track choice	-0.264***	-0.125***	-0.138***	-0.127***
	(0.028)	(0.023)	(0.025)	(0.025)
N	23,807	23,807	23,807	23,807
Panel B: Reading				
Track choice	-0.153***	-0.052**	-0.068***	-0.061***
	(0.026)	(0.022)	(0.023)	(0.023)
N	23,310	23,310	23,310	23,310
Panel C: Listening				
Track choice	-0.106***	-0.017	-0.033	-0.027
	(0.029)	(0.022)	(0.023)	(0.023)
N	22,710	22,710	22,710	22,710
School composition controls		✓	✓	✓
School input controls			✓	✓
Student-level controls				✓

Notes: Each cell of the table reports results from a separate regression of the respective test score on an indicator for free track choice. All regressions include test booklet fixed effects. School compositions controls include a categorical variable for the size of the municipality in which the school is located (in 6 categories); class level averages of parents' years of education, highest ISEI, books at home; percent of students in class with a disability, special-needs status, migration background; percent of male students; percent of non-native German speakers; percent of students who speak mostly German at home; percent of students who have repeated a grade. School input controls include school enrolment, a private school indicator, controls for the type of full-day offer in 4 categories (no full-day program, binding full-day, partly binding full-day, open full-day program), grade 4 instruction hours in math and German; years of experience of the German and math teacher. Student-level controls include indicators for male, German spoken at home, non-native German speaker, migrant background, born in Germany, grade repeater, grade skipper, any disability; years spent in public childcare, linear and quadratic age at test; mother's and father's highest years of education, highest ISES, country of birth in 5 categories (Germany, Poland, Russia, Turkey, other), work status in 4 categories (full-time, part-time, not employed, other), EGP class in 11 categories. Standard errors in parentheses allow for clustering at the school level. Column 2 additional includes quadratic of all school composition controls. Column 3 includes interactions of all school composition controls with the year and state. Column 4 presents results for a propensity score matching DiD estimator, where all school composition controls are used to estimate the propensity score. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

C School policies

In this section, we briefly describe three major education policy reforms that were implemented in the years we investigate in Germany and that could have affected early achievement in primary school.

C.1 Early childcare expansion

Germany offers public child care at two levels. Early child care is available for children aged 0–2, and Kindergarten is available for children aged 3–6. Since 1996, every child has been legally entitled to a place in Kindergarten from age 3 until primary school and Kindergarten attendance rates for children aged 3-6 have been constant around 90% since 2005. Early child care, on the other hand, saw a rapid expansion in Germany beginning in 2005. However, Panel A in Figure D.1 shows that the expansion of early child care was roughly in parallel in repeal and no-repeal states.

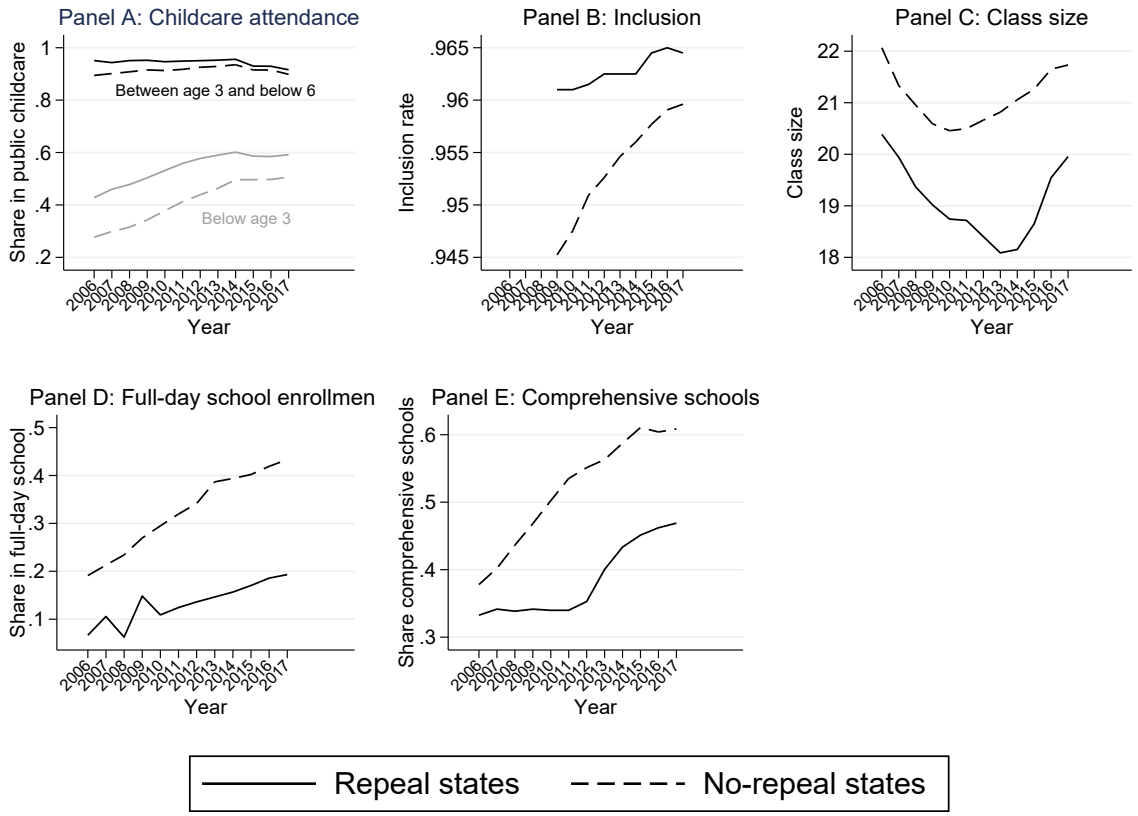
C.2 Inclusion of special needs students

In 2009, Germany enacted the United Nations Convention on the Rights of Persons with Disabilities which states that persons with disabilities should be guaranteed the right to inclusive education at all levels. Panel B in Figure D.1 shows that since then the gap in the share of special needs students in regular schools between repeal and no-repeal states has narrowed from 2 to .5 percentage points in 2017. This convergence is due to a stronger increase in the inclusion rate in no-repeal states which started out with a lower inclusion rates in 2009 compared to repeal states. However, these differences are small in magnitude and hence unlikely to drive our main effects. Nevertheless, we include controls for the share of special needs students in class in our DiD approach.

C.3 Full-day school reform

The German central government and federal states have invested massively in the expansion of full-day schools since 2001 with the declared aim of reducing educational inequality and improving the reconciliation of family and work life. Panel D of Figure D.1 shows that the share of students in full-day schools has grown faster in no-repeal states compared to repeal states between 2006 and 2017, raising concerns that different trends in full-day school attendance could confound our DiD estimates. However, an important feature of the full-day school expansion was that it happened gradually and differed at the school level as the responsibility of setting up full-day schools lay with the municipality (Shure, 2019). As a result, there is large variation in the expansion of full-day schools even within states which allows us to control for it at the individual school level. Doing so does not change our results (see columns 2 and 3 of Table 4).

Figure D.1: Trends in various characteristics



Notes: The group of repeal states consists of Baden-Württemberg and Saxony-Anhalt. No-repeal states are Schleswig-Holstein, Hamburg, Lower Saxony, Bremen, Hesse, Rhineland-Palatinate, Bavaria, Saarland, Berlin, Brandenburg, Mecklenburg-Vorpommern, Saxony, and Thuringia. Panel A plots the share of children in public childcare by age groups. Panel B plots the inclusion rate which is defined as the share of special-needs students taught in regular schools rather than schools for children with special-needs. Panel C plots class size in primary school. Panel D plots the share of all primary school students enrollend in full-day schools. Panel E plots the share of comprehensive schools among all secondary schools. The data source for Panel A is the German Federal Statistical Office. Specifically, data on the number of children in childcare comes from: `https://www-genesis.destatis.de/genesis//online?operation=table&code=22541-0002&levelindex=0&levelid=1588851385135` (Retrieved: 01/13/2020). Data on the number of children in each age-group comes from: `https://www-genesis.destatis.de/genesis//online?operation=table&code=12411-0012&levelindex=1&levelid=1588853841953` (Retrieved: 01/16/2020). The data for Panel B comes from Knauf and Knauf (2019). The date source for Panels C and E is the German Federal Statistical Office: Allgemeinbildende Schulen Fachserie 11, Reihe 1: `https://www.destatis.de/GPStatistik/receive/DEHeft_heft_00112288` (Retrieved: 01/16/2020). The data source for Panel D is the standing conference of the ministers of education and cultural Affairs (*KMK*): `https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/allgemeinbildende-schulen-in-ganztagsform.html` (Retrieved: 01/16/2020).

D Factor analysis

Here, we perform principal component factor analysis for different sets of questionnaire items to create factors that describe students' intrinsic motivation, parenting styles, and the schooling environment. As a decision rule for the factors to be included in the analysis in Table B.3, we follow Kaiser's criterion and only keep those factors with eigenvalues greater than 1. The results for the eigenvalues and the mapping of questionnaire items to the factors with eigenvalues greater than 1 are shown in the tables below.

D.1 Intrinsic motivation

Table D.1: The mapping of intrinsic motivation items to first principal-component

	Comp1
Eagerness to learn 1: I like attending school.	.3801852
Eagerness to learn 2: I enjoy school	.3657373
Eagerness to learn 3: I enjoy learning at school very much.	.3792075
Parent: Joy of learning 1: <name of target child> likes attending school.	.4480518
Parent: Joy of learning 2: <name of target child> enjoys school.	.4443316
Parent: Joy of learning 3: <name of target child> enjoys learning at school.	.4443526
Feeling at school: bored	-.5347926
Feeling at school: I enjoyed class	.3232581

D.2 Parenting styles

Table D.2: Testing for the number of factors in parenting style measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.782587	1.637993	0.3975	0.3975
Comp2	1.144594	.3537433	0.1635	0.5610
Comp3	.7908505	.0573243	0.1130	0.6740
Comp4	.7335262	.0575553	0.1048	0.7788
Comp5	.6759709	.1935522	0.0966	0.8754
Comp6	.4824187	.0923655	0.0689	0.9443
Comp7	.3900532	.	0.0557	1.0000

Table D.3: The mapping of parenting style measures to parenting style factor

	Comp1	Comp2
My parents pay a great deal of attention to how much time I spend on homework.	.3318548	.3836045
My parent decide for how long I can watch TV.	.4126713	-.3103091
My parents want me to always do homework at the same time.	.2791953	.5654729
My parents insist that I spend a certain amount of time each day reading.	.2586883	.5300178
My parents pay a great deal of attention to how much time I spend watching TV or playing on the computer.	.4482909	-.2877236
My parents pay a great deal of attention to what I do on the computer.	.4326428	-.195061
My parents pay a great deal of attention to what I watch on TV.	.4319266	-.1871997

D.3 Teacher credentials

Table D.4: Testing for the number of factors in teacher credentials

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.755859	.9631339	0.5853	0.5853
Comp2	.7927253	.3413098	0.2642	0.8495
Comp3	.4514155	.	0.1505	1.0000

Table D.5: The mapping of parental items to the teacher credential factor

	Comp1
Grade of university entrance qualification	.4746464
Grade in first state examination	.6397937
Grade in second state examination	.6044624

D.4 Determination of school environment factors for robustness check

D.4.1 Teacher principal-components

Table D.6: Testing for the number of factors in classroom activities

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.848558	.4947116	0.2311	0.2311
Comp2	1.353847	.2605239	0.1692	0.4003
Comp3	1.093323	.1106796	0.1367	0.5370
Comp4	.9826431	.0543096	0.1228	0.6598
Comp5	.9283335	.0365148	0.1160	0.7758
Comp6	.8918187	.1429876	0.1115	0.8873
Comp7	.7488311	.5961849	0.0936	0.9809
Comp8	.1526462	.	0.0191	1.0000

Table D.7: The mapping of classroom activies to factors

	Comp1	Comp2	Comp3
Time spent each week - homework	.388106	.1335385	.4555481
Time spent each week - lecture teacher	.3337196	-.2044993	.4968003
Time spent each week - tasks/exercises with assistance	.2274838	-.5190346	-.1177694
Time spent each week - tasks/exercises without assistance	-.6635105	.0071501	.2629223
Time spent each week - repetitive drills and exercises	.2955223	-.2770449	-.5497798
Time spent each week - tests, quizzes or guessing games	.3123104	.2401403	.1253596
Time spent each week - classroom management	.2420969	.4905609	.0045187
Time spent each week - other student activities	.0502772	.5439009	-.3803608

Table D.8: Testing for the number of factors in teachers' job satisfaction measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.652024	2.492871	0.4565	0.4565
Comp2	1.159154	.3761999	0.1449	0.6014
Comp3	.7829537	.1227434	0.0979	0.6993
Comp4	.6602103	.0936725	0.0825	0.7818
Comp5	.5665378	.1449623	0.0708	0.8526
Comp6	.4215755	.0418397	0.0527	0.9053
Comp7	.3797358	.0019269	0.0475	0.9528
Comp8	.3778089	.	0.0472	1.0000

Table D.9: The mapping of teachers' job satisfaction measures to factors

	Comp1	Comp2
Teacher: Professional pressure: Enjoy job	-.3426858	.5294752
Teacher: Professional pressure: Considered leaving the profession	.3079385	-.4966412
Teacher: Professional pressure: Satisfied with work	-.3819497	.2713637
Teacher: Professional pressure: Professional ideals cannot be realized	.2917555	-.0378445
Teacher: Professional pressure: Constantly overloaded	.3983976	.277262
Teacher: Professional pressure: Cannot switch off from job	.3736666	.3910459
Teacher: Professional pressure: Responsibility puts me under great pressure	.3259451	.1220959
Teacher: Professional pressure: Time pressure too great	.3900588	.391452

Table D.10: Testing for the number of factors in teachers' skill development measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.12846	2.688166	0.7821	0.7821
Comp2	.4402948	.1628571	0.1101	0.8922
Comp3	.2774377	.1236306	0.0694	0.9615
Comp4	.1538071	.	0.0385	1.0000

Table D.11: The mapping of teachers' skill development measures to factors

	Comp1
Teacher: Skills development: Developed competencies in last five years	.4759908
Teacher: Skills development: Teaching significantly improved in last five years	.50548
Teacher: Skills development: Developed didactic knowledge in last five years	.4945882
Teacher: Skills development: Learned a lot in the last five years (lessons)	.522786

Table D.12: Testing for the number of factors in the first set of teaching styles measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.083193	.8579295	0.2976	0.2976
Comp2	1.225264	.2770702	0.1750	0.4726
Comp3	.9481935	.1609745	0.1355	0.6081
Comp4	.7872189	.0588102	0.1125	0.7206
Comp5	.7284087	.0950231	0.1041	0.8246
Comp6	.6333856	.0390491	0.0905	0.9151
Comp7	.5943365	.	0.0849	1.0000

Table D.13: The mapping of the first set of teaching styles measures to factors

	Comp1	Comp2
I demand considerably less from students who are less capable.	.3081349	-.2574456
I form groups of students with similar capabilities.	.2490675	-.6068413
I form groups of students with different capabilities.	.051335	.7106913
I assign homework with varying difficulty to students.	.4836976	.0117447
Faster students continue with their work, slower students still practice.	.4135286	.1284812
If students don't understand something we do dedicated additional exercises.	.4556197	.1110094
I challenge capable students more.	.4773096	.1772423

Table D.14: Testing for the number of factors in the second set of teaching styles measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.511853	1.004858	0.2284	0.2284
Comp2	1.506994	.4666153	0.1370	0.3653
Comp3	1.040379	.0386831	0.0946	0.4599
Comp4	1.001696	.087786	0.0911	0.5510
Comp5	.9139099	.0530825	0.0831	0.6341
Comp6	.8608274	.1380924	0.0783	0.7123
Comp7	.7227351	.0441929	0.0657	0.7780
Comp8	.6785422	.0208215	0.0617	0.8397
Comp9	.6577207	.0726793	0.0598	0.8995
Comp10	.5850414	.0647397	0.0532	0.9527
Comp11	.5203017	.	0.0473	1.0000

Table D.15: The mapping of the second set of teaching styles measures to factors

	Comp1	Comp2	Comp3	Comp4
Teacher: Teaching: differentiated assignments	.17673	.3941009	.5836551	-.0733348
Teacher: Teaching: quickly noticing trouble	.2973317	.2169118	.3134146	-.2226329
Teacher: Teaching: knowing the rules	.3309356	.0681771	.0519867	.3313106
Teacher: Teaching: repeating assignments	.3364888	-.2657489	.2770581	-.0331024
Teacher: Teaching: discuss general topics	.313651	.1097867	-.0820825	-.6370366
Teacher: Teaching: teach proven concepts	.3223802	-.4595981	-.1424452	-.1434558
Teacher: Teaching: summarize material	.339955	-.3887769	.012668	-.1254873
Teacher: Teaching: asking for justifications	.2950808	.2008882	-.5589497	-.0409272
Teacher: Teaching: quiet classes	.2817987	-.2807327	.1486585	.5256419
Teacher: Teaching: identifying mistakes	.3092068	.3284462	-.344522	.1661602
Teacher: Teaching: extra tasks for faster students	.2786576	.34769	-.0352989	.2940548

Table D.16: Testing for the number of factors in the third set of teaching styles measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.949359	.2181744	0.2437	0.2437
Comp2	1.731185	.8349468	0.2164	0.4601
Comp3	.8962383	.0982682	0.1120	0.5721
Comp4	.7979701	.0479607	0.0997	0.6718
Comp5	.7500093	.0615138	0.0938	0.7656
Comp6	.6884956	.0593411	0.0861	0.8517
Comp7	.6291545	.0715669	0.0786	0.9303
Comp8	.5575876	.	0.0697	1.0000

Table D.17: The mapping of the third set of teaching styles measures to factors

	Comp1	Comp2
Teacher: Opinion: Make decisions	-.4256848	.2285789
Teacher: Opinion: Role of teacher with regard to investigating and exploring	.1950975	.4526778
Teacher: Opinion: Learning through independent problem-solving	.4607761	.2934843
Teacher: Opinion: Lessons with clear answers	-.2833438	.4581431
Teacher: Opinion: Teaching of facts	-.3595482	.3765329
Teacher: Opinion: Possibility of independent problem-solving	.4320361	.3340418
Teacher: Opinion: Quiet classroom	-.2568858	.3532776
Teacher: Opinion: Thinking and reasoning processes	.3258996	.2619929

D.4.2 School environment principal-components

Table D.18: Testing for the number of factors in facility quality

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.943625	1.95581	0.4906	0.4906
Comp2	.9878146	.2368979	0.1646	0.6552
Comp3	.7509167	.1354478	0.1252	0.7804
Comp4	.6154688	.1928814	0.1026	0.8830
Comp5	.4225874	.1429996	0.0704	0.9534
Comp6	.2795878	.	0.0466	1.0000

Table D.19: The mapping of facility quality items to the facility quality factor

	Comp1
Class: Facilities: Classroom size (aggregated)	.1460446
Class: Classroom condition, brightness	.4013457
Class: Classroom condition, size	.4383296
Class: Classroom condition, functionality	.4884833
Class: Classroom condition, structural integrity	.454125
Class: Classroom condition, acoustics	.4249874

Table D.20: Testing for the number of factors in school climate measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.73435	2.568829	0.6224	0.6224
Comp2	1.165521	.7233172	0.1943	0.8166
Comp3	.4422042	.1223115	0.0737	0.8903
Comp4	.3198927	.0956492	0.0533	0.9437
Comp5	.2242435	.1104557	0.0374	0.9810
Comp6	.1137878	.	0.0190	1.0000

Table D.21: The mapping of school climate measures to factors

	Comp1	Comp2
Teacher: School climate: Conflicts open and constructive	.3801852	.3519051
Teacher: School climate: Teachers mutual help and practical support	.3657373	.4720067
Teacher: School climate: Atmosphere of trust	.3792075	.4780671
Teacher: School climate: Teachers have faith in skills of students	.4480518	-.2885274
Teacher: School climate: Teachers have faith in willingness to learn of students	.4443316	-.4012073
Teacher: School climate: Have faith in students' willingness to make efforts	.4239293	-.4249811

Table D.22: Testing for the number of factors in school culture measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.765089	2.686881	0.6275	0.6275
Comp2	1.078208	.6289038	0.1797	0.8072
Comp3	.4493044	.1968947	0.0749	0.8821
Comp4	.2524097	.0080756	0.0421	0.9242
Comp5	.2443341	.0336799	0.0407	0.9649
Comp6	.2106542	.	0.0351	1.0000

Table D.23: The mapping of school culture measures to factors

	Comp1	Comp2
Teacher: School culture: Clear pedagogical goals – binding for all	.4177294	-.4117596
Teacher: School culture: Agreement teachers/school management with learning goal	.4211118	-.3953834
Teacher: School culture: Teachers clear about learning goals	.4404828	-.3380583
Teacher: School culture: Faculty place high demands on the students	.3953836	.27562
Teacher: School culture: Teachers attach great importance to efforts of students	.4013163	.4729206
Teacher: School culture: Teachers convey students – effort	.3698094	.5101199

Table D.24: Testing for the number of factors in school management measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.673921	4.983234	0.8106	0.8106
Comp2	.6906868	.4839817	0.0987	0.9092
Comp3	.2067051	.0610539	0.0295	0.9388
Comp4	.1456512	.0205109	0.0208	0.9596
Comp5	.1251403	.0284995	0.0179	0.9774
Comp6	.0966408	.0353856	0.0138	0.9912
Comp7	.0612552	.	0.0088	1.0000

Table D.25: The mapping of school management measures to factors

	Comp1
Teacher: Management: School management manages very well	.3893949
Teacher: School management: Leads the school efficiently and goal-oriented	.3882421
Teacher: Management: School management organizes very well	.3865541
Teacher: Management: School management efficient administration of the school	.385921
Teacher: Information evaluation: Good information flow	.3673509
Teacher: Information evaluation: Relevant information provided in time	.3650379
Teacher: Information evaluation: Sufficient sharing for important decisions	.3620515

Table D.26: Testing for the number of factors in teachers' constraints measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.537873	1.644103	0.5076	0.5076
Comp2	.8937695	.0835862	0.1788	0.6863
Comp3	.8101833	.3244709	0.1620	0.8484
Comp4	.4857124	.2132503	0.0971	0.9455
Comp5	.2724621	.	0.0545	1.0000

Table D.27: The mapping of teachers' constraints measures to factors

	Comp1
Teacher: Decision making constrained by: Ministry of education	.3321538
Teacher: Decision making constrained by: School management	.5070176
Teacher: Decision making constrained by: Teacher/comprehensive conference	.5146473
Teacher: Decision making constrained by: Student body	.4392568
Teacher: Decision making constrained by: Parents	.4180894

Table D.28: Testing for the number of factors in teachers' school involvement measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.737342	1.350368	0.2488	0.2488
Comp2	1.386974	.1139429	0.1261	0.3749
Comp3	1.273031	.3364388	0.1157	0.4907
Comp4	.9365921	.0230548	0.0851	0.5758
Comp5	.9135373	.0959593	0.0830	0.6589
Comp6	.817578	.0912944	0.0743	0.7332
Comp7	.7262836	.1125439	0.0660	0.7992
Comp8	.6137396	.0365463	0.0558	0.8550
Comp9	.5771933	.0212668	0.0525	0.9075
Comp10	.5559266	.0941236	0.0505	0.9580
Comp11	.461803	.	0.0420	1.0000

Table D.29: The mapping of teachers' school involvement measures to factors

	Comp1	Comp2	Comp3
Participation: teacher conferences	.3334724	.2185579	-.4540678
Participation: development of school curriculum	.3079677	.1829208	-.520668
Participation: discussing/decisions on media teaching	.3290743	-.0777018	-.3573447
Participation: exchange of teaching materials	.2451245	-.5088302	.0150429
Participation: team discussions	.3124801	-.4254904	.1772221
Participation: discussion about learning process of individual students	.3239088	-.3524304	.0995537
Participation: team teaching in a class	.264029	.1354085	.3970559
Participation: professional learning activities	.3104576	.2961224	.3061027
Participation: sitting in on classes	.2996432	.3680211	.2721089
Participation: joint activities across different grades	.2442405	.2998474	.1584011
Participation: discussion/coordination of homework	.3278424	-.1311695	.0547798

Table D.30: Testing for the number of factors in the third set of teaching styles measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.744717	1.978405	0.5489	0.5489
Comp2	.7663122	.1357564	0.1533	0.7022
Comp3	.6305558	.1511824	0.1261	0.8283
Comp4	.4793734	.1003318	0.0959	0.9242
Comp5	.3790416	.	0.0758	1.0000

Table D.31: The mapping of the third set of teaching styles measures to factors

	Comp1
Teacher: Cover: Well-organized cover plan	.4269793
Teacher: Cover: Replacement lessons by specialist teachers	.4438474
Teacher: Cover: Contents of replacement lessons discussed	.5008795
Teacher: Cover: Canceled lesson – working material available	.454463
Teacher: Cover: Responsibilities clearly regulated	.4040683

Table D.32: Testing for the number of factors in school decision-making process measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.962291	4.892473	0.6625	0.6625
Comp2	1.069818	.3731815	0.1189	0.7813
Comp3	.6966361	.3373709	0.0774	0.8587
Comp4	.3592653	.108451	0.0399	0.8987
Comp5	.2508143	.0276483	0.0279	0.9265
Comp6	.223166	.0582567	0.0248	0.9513
Comp7	.1649093	.0113614	0.0183	0.9697
Comp8	.1535479	.0339954	0.0171	0.9867
Comp9	.1195525	.	0.0133	1.0000

Table D.33: The mapping of school decision-making process measures to factors

	Comp1	Comp2
Teacher: Decision-making processes: School council fast decisions	.3035316	.5250325
Teacher: Decision-making processes: School council process very efficient	.3373394	.4149928
Teacher: Decision-making processes: School decisions goal-oriented	.3433302	.3009148
Teacher: Decision-making processes: Joint decisions among colleagues	.341928	.0062562
Teacher: Decision-making processes: School decisions rarely criticism	.3034443	.0919795
Teacher: Decision-making processes: Important decisions are accepted by teachers	.3393176	-.1185401
Teacher: Decision-making processes: All teachers involved in important decisions	.3410277	-.3928958
Teacher: Decision-making processes: Opinions of faculty for important decisions	.3354744	-.3929523
Teacher: Decision-making processes: Faculty key role in important decisions	.3510306	-.3608911

Table D.34: Testing for the number of factors in teachers' opinion of colleagues measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.4834	2.771279	0.5806	0.5806
Comp2	.7121203	.1150135	0.1187	0.6993
Comp3	.5971069	.0963864	0.0995	0.7988
Comp4	.5007205	.0897435	0.0835	0.8822
Comp5	.410977	.1153014	0.0685	0.9507
Comp6	.2956756	.	0.0493	1.0000

Table D.35: The mapping of teachers' opinion of colleagues measures to factors

	Comp1
Teacher: Opinion of colleagues: objections to change	-.40529
Teacher: Opinion of colleagues: readiness to evaluate teaching methods	.3851947
Teacher: Opinion of colleagues: openness to new teaching methods	.4379974
Teacher: Opinion of colleagues: lack of readiness to learn new things	-.3600697
Teacher: Opinion of colleagues: effort to define school's pedagogical concept	.4236833
Teacher: Opinion of colleagues: renewal and development	.4317008

D.4.3 Parent principal-components

Table D.36: Testing for the number of factors in cooperation with parent measures

	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.031953	1.827471	0.3790	0.3790
Comp2	1.204483	.2198805	0.1506	0.5296
Comp3	.9846022	.2342561	0.1231	0.6526
Comp4	.7503461	.1197133	0.0938	0.7464
Comp5	.6306327	.100453	0.0788	0.8253
Comp6	.5301797	.0195239	0.0663	0.8915
Comp7	.5106558	.1535082	0.0638	0.9554
Comp8	.3571476	.	0.0446	1.0000

Table D.37: The mapping of cooperation with parent measures to factors

	Comp1	Comp2
Teacher: Working with parents: Fun	.2442994	.4032812
Teacher: Working with parents: Parents as partners	.378373	.2563513
Teacher: Working with parents: Info about school events	.3712089	-.2070268
Teacher: Working with parents: Follow up on complaints	.4215019	.1145481
Teacher: Working with parents: Info about strengths/weaknesses	.4300077	-.3330645
Teacher: Working with parents: Info about learning progress	.347918	-.4995927
Teacher: Working with parents: Appointments	.362355	.069906
Teacher: Working with parents: Speaking outside of school	.210801	.5918236