ISSN 2042-2695

CENTRE *for* ECONOMIC PERFORMANCE

CEP Discussion Paper No 1706

July 2020

Better Together? Heterogeneous Effects of Tracking on Student Achievement

Sönke Hendrik Matthewes

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UKRI Economic and Social Research Council

Abstract
I study the effects of early between-school ability tracking on student achievement, exploiting institutional differences between German federal states. In all states, about 40% of students transition to separate academic-track schools after comprehensive primary school. Depending on the state, the remaining student body is either directly tracked between two additional school types or taught comprehensively for another two years. Comparing these students before and after tracking in a triple-differences framework, I find evidence for positive effects of prolonged comprehensive schooling on mathematics and reading scores. These are almost entirely driven by low-achievers. Early and rigid forms of tracking can thus impair both equity and efficiency of school systems.

Key words: Tracking, student achievement, school systems, inequality, difference-in-differences, triple-differences, value-added
JEL Codes: I24, I28, J24

This paper was produced as part of the Centre's Education and Skills Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

I am grateful for advice and comments from three anonymous referees, Gilat Levy, Jan Marcus, Guglielmo Ventura, Heike Solga, Katharina Spieß, seminar participants at WZB, DIW and CEP/LSE, as well as participants of the EALE conference 2018, COMPIE conference 2018, DFG SPP 174 workshop 2018, the BeNA labour workshop 2017. I acknowledge financial support by the Jacobs Foundation and German Federal Ministry for Education and Finance (BMBF) through the College for Interdisciplinary Educational Research (CIDER). This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort (SC) Grade 5, doi:10.5157/NEPS:SC3:8.0.1, and SC Kindergarten, doi:10.5157/NEPS:SC2:8.0.0. 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.

Sönke Hendrik Matthewes, WZB Berlin Social Science Center and a visiting PhD student at Centre for Economic Performance, London School of Economics.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

In the face of decreasing employment opportunities for low-skilled workers, the pressure on education systems to equip students with the necessary skills to succeed in modern labour markets is growing (European Commission, 2014). Wössmann (2016) demonstrates that national school systems differ markedly in how well they live up to this task. This raises the question how the optimal school system should be organised. While the (positive) effect of some institutional features of school systems on student achievement is relatively well-established by now (e.g. central exit exams), others remain fiercely debated. One of the most controversial issues in this regard is the practice of ability tracking. Tracking means grouping students by ability into vertically ordered school tracks. Countries differ widely on the degree to which they track students, and the age at which students begin to be tracked (Betts, 2011). Some countries, like Finland, eschew tracking altogether, relying only on comprehensive compulsory schooling. Others, like Germany, separate students into one of three ranked schools types at an age as early as 10. Between these two extremes lie countries like the US, which stream students into different tracks within schools.

The argument behind grouping students by ability is always one of efficiency.[1] Proponents of tracking posit that lower variance classrooms allow for better tailoring of curricula, instruction speed and pedagogy to students' abilities and should, therefore, benefit learning for all students (Duflo et al., 2010). Critics, in contrast, fear that only high track/ability students benefit from tracking, whereas students assigned to lower tracks are condemned to lower achievement compared to a scenario with comprehensive schooling. Indeed, there are many mechanisms that might make the effects of tracking heterogeneous. First, to the extent that high performing peers are beneficial to learning (or low performing ones harmful), tracking mechanically exerts an unequal influence as it deprives lower track students of more able peers (Sacerdote, 2011). Second, there might be motivational consequences of separating students by ability. Lower track students, knowing they are deemed to be of lower aptitude, might feel discouraged and reduce their learning efforts. Third, if (financial) resources differ between tracks, students of certain tracks might be disadvantaged (Betts, 2011). Additionally, even if ability tracking is theoretically Pareto efficient, practical

[1]The debate on tracking being a long-standing one, there is a vast social-scientific literature that discusses its pros and cons. For seminal contributions see e.g. Oakes (1985), Gamoran and Mare (1989) and Slavin (1990).

implementation is likely to be error-prone as ability is not directly observable and proxies like teacher assessments and tests are noisy and can be socio-economically biased (Brunello *et al.*, 2007; van Ewijk, 2011).

Given these opposing mechanisms, the net effect of tracking on student achievement is theoretically ambiguous and ultimately an empirical question. If proponents are right and homogeneous classrooms increase the effectiveness of teaching, tracking should raise average achievement by benefiting students of all ability levels. If the hypothesised negative effects are at work, tracked school systems should depress student achievement at the bottom. In terms of efficiency, the net effect of tracking then depends on scope and relative strength of these effects.[2] In terms of equity, tracking might thus translate small performance differentials at young ages into substantial inequalities in later life. These dynamics should be more pronounced the earlier tracking starts, as divergences can accumulate, and in between-school tracking systems as compared to within-school ones, as the vertical differentiation between tracks is stronger (Betts, 2011).

Indeed, achievement differences between students of different tracks are large and well-documented (e.g. Dustmann, 2004) and countries with more rigid tracking systems tend to exhibit higher levels of educational inequality (Waldinger, 2007). The problem is that such correlational findings, whether at the individual or the country level, are likely to suffer from severe endogeneity. Students are not randomly allocated to school tracks but explicitly selected on ability. Similarly, countries' educational systems are affected by historical factors that also directly influence student outcomes. In the face of these selection problems no clear consensus on the effect of early between-school tracking has emerged in the empirical literature. While, in line with theory, effect estimates for high-ability students seem to vary between positive and null, the evidence on how tracking affects low-ability students is more mixed.

This paper exploits unique within-country between-state variation in tracking practices in Germany to isolate the effect of early between-school tracking on the achievement of students in lower tracks. While in all German states primary school is comprehensive, the grouping of students in secondary school, which commences in fifth grade when students are about ten years of age, differs between states: some states have a three-tiered and others a two-tiered secondary school

[2] It appears that the costs of tracked and untracked school systems are roughly comparable (Hanushek and Wössmann, 2006). Following the literature, I therefore loosely refer to differences in mean outcomes as efficiency differences.

system. In the former, students are tracked between low-, intermediate- and academic-track schools based on their performance in primary school. In states with the two-tiered regime, low- and intermediate-track schools were conflated, so that students are only tracked between academic and non-academic-track schools. Also these combined non-academic-track schools sort students by ability eventually, but in the first two years of secondary school (i.e. in grades 5 and 6) classes are formed disregarding previous performance or ability. Academic-track schools, called *Gymnasium*, do not differ between states and cater to about 40% of students in either regime. Accordingly, between-state differences in tracking are relevant only for the non-academic part of the student body: after comprehensive primary school, non-academic-track students are either directly tracked into low- and intermediate-track schools or taught comprehensively for another two years.[3] Note that these between-state differences pertain to the ability grouping of students only, as curricula are fully general in the first years of secondary school everywhere.

My research design exploits this variation in tracking in a difference-in-differences (DD) framework: I estimate how the achievement of one cohort of non-academic-track students develops differently over the first two years of secondary school depending on whether students are tracked or taught comprehensively. This strategy controls for grade-constant heterogeneity between states and general achievement trends between grades. Because the DD estimate might still be confounded by state-specific achievement trends, additionally, I compare the between-state differences for non-academic-track students to those for academic-track students, for whom there is no difference in tracking between states (who are thus 'untreated' no matter the state). This is implemented via a triple-differences (DDD) estimator. After having thus established the mean effect, I explore the distributional consequences of tracking. First, I provide non-parametric density estimates of the impact of tracking on the overall achievement distribution. Second, I explore how the effect of tracking depends on students' position in the pre-tracking achievement distribution.

The analysis is based on individual-level panel data for mathematical and reading competence from the German National Educational Panel Study (NEPS), which followed one cohort of students over their school career. The NEPS provides measures of student achievement right before and after the first two years of secondary school (i.e. right before and after the grade window with

[3]This refers to all 12 federal states under investigation (out of 16 in total); see section 2.1 and footnote 18.

clear-cut between-state differences in tracking), as well as detailed information on students' family backgrounds and schooling inputs. In addition, I draw on the Institute for Educational Quality Improvement's (IQB) National Assessment Studies to corroborate my findings in larger samples and to assess the persistence of effects through the end of lower secondary schooling.

In sharp contrast with the predictions of tracking advocates, my results suggest that early between-school tracking *decreases* student achievement. The average effect of continued comprehensive schooling in grades 5 and 6 on seventh-grade test scores is estimated to be 0.17 standard deviations (SD) in mathematics and 0.24 SD in reading. Even though these effects are not very precisely estimated they are statistically significant and remarkably stable across specifications that flexibly control for student and school characteristics, as well as the inclusion of academic-track students as an additional control group in the DDD model. Robustness checks, such as comparing achievement trends in primary school and excluding outlier states, lend further credence to the causal interpretation of, at least, the direction of the effect estimates. Finally, the analysis with the IQB data shows that, while there is some fade-out over time, comprehensively taught non-academic-track students are still significantly better off towards the end of ninth grade.

The heterogeneity analysis reveals that these results are driven by the lower tail of the initial achievement distribution: for low-achievers effects are large and persistent, whereas for high-achievers effect estimates are insignificantly different from zero (yet, strictly non-negative). Consequently, comprehensive schooling has an equalising effect on the distribution of test scores. Delaying tracking does not trade off efficiency against equity, but seems to enhance both. I provide a discussion of the channels through which the effect might operate and find empirical support for the importance of peer effects and socio-emotional mechanisms, like improved school-related motivation and educational aspirations.

Note that the treatment effect identified in this paper pertains to a population of students that excludes the group of highest achievers in academic-track schools. Hence, one cannot directly extrapolate from these results to the effects of fully comprehensive school systems. Still, they prove wrong the premise that there is a monotonously positive relationship between classroom homogeneity and student learning. Early between-school tracking appears to impose large costs on low-achieving students. Accordingly, more dispersed achievement distributions in more tracked

systems do not appear to be a mere artefact of selection and the oft-voiced equity concerns in this context seem warranted.

This paper contributes to the literature on the systemic impact of between-school tracking, which, in the face of the endogeneity issues involved, could only produce tentative evidence so far. The most credible results stem from two strands of the literature.[4] The first exploits temporal *within-country* variation in tracking practices induced by de-tracking reforms. The second leverages the large variation in tracking practices *between countries* in different ways.

A number of prominent studies of the first strand analyse de-tracking reforms in the Nordic countries. Similar to my findings, they find reform-induced achievement gains for students from lower socio-economic backgrounds (see Meghir and Palme, 2005, for Sweden; Aakvik *et al.*, 2010, for Norway; and Kerr *et al.*, 2013, for Finland). Given that these reforms simultaneously changed other features of the school system, like the minimum school-leaving age, the effects cannot be unequivocally attributed to tracking, however. Analyses of Britain's de-tracking reform, which all use the fact that implementation was staggered across regions, have generated more mixed results.[5] Pischke and Manning (2006) argue that this is due to unobserved regional heterogeneity that cannot sufficiently be controlled for with existing data sets.

An arguably cleaner natural experiment, yet more narrow in scope, is the experience of Northern Ireland, which maintained its tracking system but increased the share of students admitted to the high track. Interestingly, the findings concerning the top end of the achievement distribution (medium high performers joining high performers) mirror mine for the bottom end (low performers joining medium performers): weaker students' gains from entering higher track environments are large, whereas losses for the stronger students are small or absent (Guyon *et al.*, 2012). Similarly, Piopiunik (2014) finds that a reform-induced increase in tracking in the German state of Bavaria led to achievement losses at the bottom.[6] A potential explanation for these results (and mine) is offered by Garlick (2018) who shows that low-achieving students are more sensitive to peer group composition than high-achievers, explaining the negative net effect of a residential tracking policy

[4]This brief literature review focuses on papers analysing the *systemic* effects of *between-school* tracking. The discussion of a large related literature on the effects of within-school streaming is deferred to the conclusion.

[5]See Kerckhoff *et al.* (1996); Galindo-Rueda and Vignoles (2004); Pischke and Manning (2006).

[6]The Bavarian pre-post differences analysed by Piopiunik (2014) closely resemble the (contemporaneous) differences in tracking analysed in this paper. Reassuringly, his findings based on a single state's reform are confirmed in this study for the whole of Germany.

in South Africa.

Studies of the second strand have employed different strategies to circumvent the potentially severe endogeneity problems that come with between-country comparisons. One rather descriptive strategy limits attention to inequality, comparing only family background effects between tracked and untracked countries. These studies generally find that early between-school tracking is associated with steeper socio-economic gradients for student achievement (see e.g. Brunello and Checchi, 2007; Schütz *et al.*, 2008).

A second strategy, introduced in a seminal paper by Hanushek and Wössmann (2006), is based on the observation that primary school is comprehensive everywhere, regardless of how the secondary school system looks. These studies use DD to estimate how test scores change differently from primary to secondary school between countries with tracked and comprehensive schooling. Most results indicate that tracking increases inequality in student achievement,[7] though Waldinger (2007) argues that these results are sensitive to the way countries are categorised into tracked and untracked ones. This highlights a major problem of the cross-country literature: when classifying countries as comprehensive or tracked, a range of quite heterogeneous between-school tracking systems are lumped together and compared to an even more diverse group that includes both comprehensive and within-school streaming systems. Hence, the treatment (and the counterfactual) is not clearly defined. Other problems include that also *changes* in outcomes might be related to unobserved differences between tracked and untracked countries (Betts, 2011) and the pooling of incomparable test scores (Contini and Cugnata, 2016).

My study merges the approaches of the within- and the cross-country literatures. I adopt the logic of Hanushek and Wössmann's (2006) DD approach in comparing changes in test scores between elementary and secondary school for the identification of the effect of tracking. Yet, the fact that I exploit within- instead of cross-country differences allows me to improve on a number of important points. First, apart from differences in tracking, school systems are strongly harmonised between German states such that the treatment is clearly defined in my case. Therefore, second, the common trends assumption necessary for DD is much more plausible in my setting than in previous studies. Crucially, I can directly assess its plausibility *ex post* using academic-track students for

[7]Next to Hanushek and Wössmann (2006), see Ammermüller (2013) and Schwerdt and Ruhose (2016).

whom there is no difference in tracking. Third, individual-level panel data allows me go beyond mean effects and estimate how the effect of tracking depends on students' position in the initial achievement distribution. This is key given that the debate on tracking revolves around a perceived efficiency-equity trade-off.

Finally, it is important to highlight that I estimate a systemic (state-level) effect of tracking, which contrasts with a literature on marginal effects. For example, Dustmann *et al.* (2017) also study the German context but, using an individual-level instrumental variables strategy, find no effect of track placement on educational attainment or earnings for students *at the margin between two tracks*.[8] Though important for understanding the consequences of (mis)allocation of hard-to-assign students to tracks given an early between-school tracking system, their estimate tells us little about whether tracking is desirable in the first place. My results suggest that the separation of students into different schools at an age as early as 10 depresses achievement for a sizeable group of (non-marginal) low-achievers, thus putting them at a double disadvantage.

The paper is structured as follows: section 2 lays out the institutional background and the identification framework. Section 3 describes my data sources and presents descriptive findings. Section 4 present the estimation results, including robustness checks and a discussion of potential mechanisms. Finally, section 5 discusses implications and concludes.

2 Institutional Background and Research Design

2.1 The German School System and Heterogeneity Therein

In Germany, sovereignty over education policy lies with the state governments. In order to ensure the comparability of educational standards and degrees, however, the federal Standing Conference of the Ministers of Education and Cultural Affairs of the States (*Kultusministerkonferenz*) harmonises education policies between states considerably (Kultusministerkonferenz, 2014). Within this unique situation of educational federalism, a school system has developed that is fairly homogeneous across Germany in terms of basic structure, teaching methods and curricula, but exhibits fine differences within some areas of schooling policy – especially, school structure and, thus, tracking practices.

[8] This is in the spirit of a larger literature on the benefits of entering selective schools (e.g. Abdulkadiroğlu *et al.*, 2014).

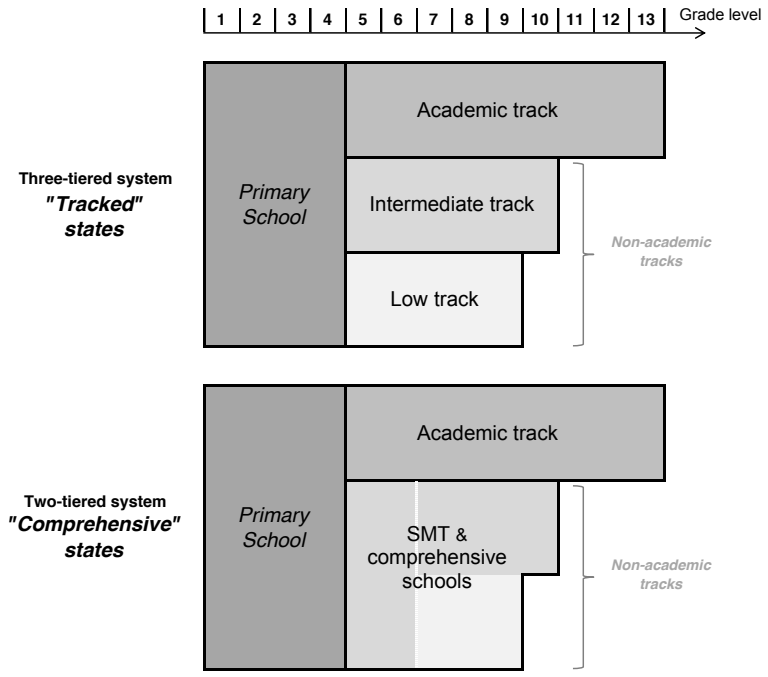

Figure 1. *Schematic overview of the two tracking regimes in Germany.*

Notes: For illustrative purposes the figure abstracts from the fact that in some of the three-tiered Tracked states there are some comprehensive schools (see text and Table 1). Academic track = *Gymnasium*, Intermediate track = *Realschule*, Low track = *Hauptschule*, School with multiple tracks (SMT) = *Schule mit mehreren Bildungsgängen*, Comprehensive schools = *Integrierte Gesamtschule*.

It is this heterogeneity within a context of general comparability that I exploit to shed light on the impact of tracking on student achievement.

Throughout Germany, compulsory schooling starts at the age of 6 with primary school, which covers the first four grade levels and is taught comprehensively with no ability grouping of students within or between schools.[9] Differences between states emerge after the end of comprehensive primary school. They are summarised schematically in Figure 1.

The traditional (West) German secondary school system is three-tiered: upon leaving primary school after fourth grade, i.e. around the age of 10, students are tracked into one of three vertically ordered school types – *Hauptschule*, *Realschule* and *Gymnasium*, representing low, intermediate and academic track – based on their previous performance.[10] These tracks lead to different school-leaving certificates and differ substantially in terms of years of schooling, curriculum, teacher certification and peer composition. The academic track (i.e. *Gymnasium*) has the most demanding

[9]In the two states of *Berlin* and *Brandenburg*, primary school lasts six years. For this reason, they are not part of the analysis.

[10]In all states students receive a track recommendation by their teacher based on their performance in primary school. Whether it is binding depends on the state. All results are fully robust to the inclusion of an indicator variable for binding teacher recommendations (and that indicator variable always turns out to be insignificant itself; see Appendix Table B2). Therefore, all that follows abstracts from this difference between states.

curriculum, lasts eight or nine years and is the only track leading directly to a school-leaving certificate that entitles to entry into university. This makes for a clear divide between the academic and the non-academic segments of the school system (also in reputation). The intermediate track (i.e. *Realschule*) provides general knowledge, lasts six years and is supposed to prepare students for advanced vocational and professional education. If students complete the intermediate track successfully and meet state-specific requirements they may upgrade to the academic track after grade 10. The low track (i.e. *Hauptschule*) provides a more basic general education, lasts five or six years and prepares students for technical vocational education. Also here, after completion, upgrading to intermediate-track schools is possible under specific conditions.

Currently, five states still have the traditional three-tiered system (see Figure 2).[11] The rest has a two-tiered secondary school structure that distinguishes between academic- and non-academic-track schools only. This group consists of three East German states, which never adopted the three-tiered system, and four West German states that reformed their system.[12]

The East German states had to align their (comprehensive) school system with that of the West after German reunification. This led to a compromise where the East adopted the *three-tiered* differentiation in school-leaving certificates but opted for a *two-tiered* school structure (Edelstein and Nikolai, 2013).[13] Instead of separate low- and intermediate-track schools there is only one non-academic school type, labelled 'School with Multiple Tracks' (*Schule mit mehreren Bildungsgängen*; henceforth SMT). Here, all students not attending an academic-track *Gymnasium* school are taught together. If a student leaves an SMT after five years (without failing the year, of course) she receives the low degree. If she stays on for another year and attains the necessary grades she earns the intermediate degree. Hence, the difference between the two systems is one of tracking only.

In many Western states low-track schools have become stigmatised due to falling student numbers and a lack of prospects for its graduates (Helbig and Nikolai, 2015). This led several states to reform their school system along the lines of the two-tiered system. Like in the East, the three

[11]These are *Bavaria*, *Baden-Württemberg*, *Hesse*, *Lower Saxony* and *North Rhine-Westphalia*.

[12]East German states: *Saxony*, *Saxony-Anhalt* and *Thuringia*. West German states: *Bremen*, *Hamburg*, *Saarland* and *Schleswig-Holstein*.

[13]Except for *Mecklenburg-Vorpommern*, which initially adopted and briefly maintained a three-tiered system. For reasons discussed below, this state is not part of the analysis.

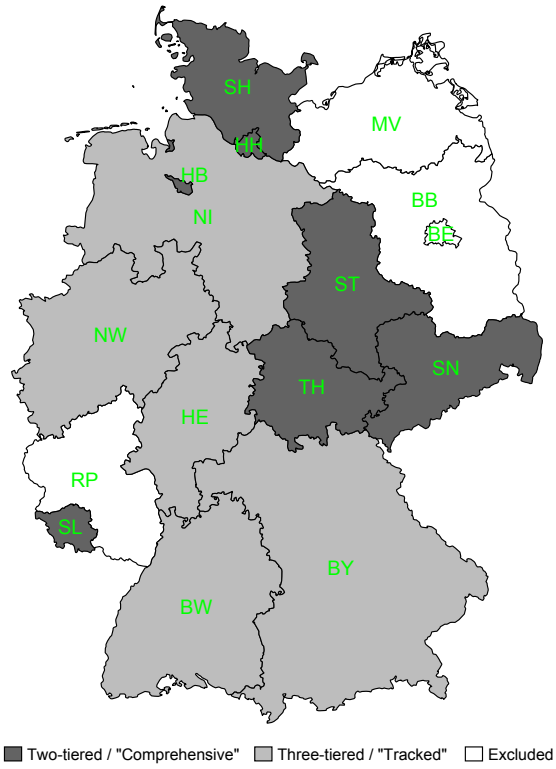

Figure 2. *German federal states coloured by tracking regime.*

Notes: BW = Baden-Württemberg; BY = Bavaria; BE = Berlin; BB = Brandenburg; HB = Bremen; HH = Hamburg; HE = Hesse; MV = Mecklenburg-Vorpommern; NI = Lower Saxony; NW = North Rhine-Westphalia; RP = Rheinland-Pfalz; SL = Saarland; SN = Saxony; ST = Saxony-Anhalt; SH = Schleswig-Holstein; TH = Thuringia.

different school-leaving certificates, as well as a distinct academic track, were retained, but separate low- and intermediate-track schools were abolished and replaced with so-called 'comprehensive schools' (*Gesamtschule*). Thus, just like SMTs in the East, these schools comprise all non-academic-track students.[14]

While both SMTs and comprehensive schools track students internally in higher grades, they are prohibited from doing so in the first two years of secondary schooling (i.e. in grades 5 and 6) (Leschinsky, 2008). Instead, in these two grades classes continue to be formed disregarding ability or previous performance. Only from grade 7 onwards these schools may track students by forming track-specific classes (i.e. separate low- and intermediate-track classes) or by applying subject-specific ability sorting.[15] In most states, it is up to schools to decide if and how to group students starting in grade 7 and, unfortunately, this information is not centrally collected. Accordingly,

[14]The difference between the East German SMTs and the West German comprehensive schools is that in the former only the basic and intermediate degrees can be obtained, while in the latter, mostly, all three degrees can be earned (Helbig and Nikolai, 2015). In practice, this difference is only relevant in much later grades than those studied here.

[15]Schools that use degree-specific within-school streaming from grade 7 onwards are labelled 'cooperative' while those that generally continue to teach comprehensively, except for streaming in particular subjects, are called 'integrative'.

between-state (and school) differences in higher grades are blurry. The first two years of secondary school, however, provide a time window where institutional differences regarding the tracking of students are clear-cut.

Comparability is further bolstered by the fact that the first two years in non-academic secondary schools are strongly harmonised. Official information of the Kultusministerkonferenz (2014) shows that curricula and learning goals for grades 5 and 6 in the non-academic tracks focus on the acquisition of a standard set of basic general knowledge that is virtually indistinguishable between states. By way of example, average weekly instruction hours in mathematics do not differ between to the two-tiered (4.3 hours/week) and three-tiered states (4.4 hours/week).[16] In most three-tiered states, curricula for grades 5 and 6 do not even differ between low- and intermediate-track schools, though the level of detail in which the material is treated might be higher in intermediate-track schools (due to students' higher ability levels) (Bald, 2011). States ensure the compatibility of curricula during the first two years of secondary school because they formally allow for the possibility to switch between tracks (Bellenberg, 2005). In practice, this happens quite rarely (only about 5% of students switch according to Bellenberg, 2012).

The analysis below focuses on the cohort of students that entered fifth grade/secondary school in 2010. The following summarises the previous discussion's key points: In the four years prior to the transition, all of these students attended comprehensive primary school. Moreover, in all states the highest achieving students transition to separate academic-track *Gymnasium* schools. The remaining non-academic-track students, however, are either further *tracked* between two different school types (in states with the three-tiered regime) or taught *comprehensively* for another two years (in states with the two-tiered regime).[17] For ease of exposition, I will refer to the five states with a three-tiered system as the 'Tracked' states and to the seven states with a two-tiered system as the 'Comprehensive' states (see Figure 2). Four states with school systems that do not fit this classification had to be excluded from the analysis.[18]

[16]These numbers are based on official time-table regulations for grades 5 and 6 in non-academic-track students reported in Pant *et al.* (2013).

[17]For completeness, it should be mentioned that in some of the Tracked states municipalities are allowed to offer comprehensive schools, where all three degrees can be earned, next to the ordinary schools of three-tiered system. For the purposes of this paper this can be thought of as non-compliance with regards to the treatment of comprehensive schooling (see section 3.4).

[18]*Berlin*, *Brandenburg* and *Mecklenburg-Vorpommern* are excluded because the tracking decision is made after grade 6 instead of after grade 4. *Rheinland-Palatinate* is excluded because the state was transitioning from a three-tiered to

2.2 Identification Strategy

The idea of this paper is to use the institutional differences in the non-academic segment of the school system between Tracked and Comprehensive to learn about the effects of early between-school ability tracking. The main challenge for this endeavour is that states' tracking policies might correlate with a whole range of other, potentially unobserved, factors determining student achievement, like e.g. student body composition or early childhood education policies.[19]

To account for such unobserved differences between states, my identification strategy uses test scores taken at two points in the educational career of students. The first achievement test is administered right after primary school, at the beginning of grade 5, and the second two years later, at the beginning of grade 7. All students are taught comprehensively in primary school. Hence, grade 5 scores should be unaffected by tracking and capture achievement differences between states unrelated to the tracking system. Grade 7 scores measure achievement right after exposure to either treatment condition and thus reflect both causal effects from tracking and permanent between-state differences. To purge the seventh-grade comparison between Comprehensive and Tracked states of grade-constant confounders I thus propose the following difference-in-differences (DD) design:

$$Y_{isg} = \delta_0 + \delta_1 \, Compr_s + \delta_2 \, Grade7_g + \beta_{DD} \, (Compr \times Grade7)_{sg} + \psi \, \mathbf{X}_{isg} + \theta_s + u_{isg}, \qquad (1)$$

where Y_{isg} is the test score of non-academic-track student i in state s and grade level $g \in \{5, 7\}$, $Compr_s$ and $Grade7_g$ are indicator variables for the Comprehensive states and grade 7 scores, respectively, $\mathbf{X}_{isg}$ is a row vector of predetermined student and school covariates, discussed in further detail below, and θ_s are state fixed effects.

In equation (1), δ_1 absorbs level achievement differences between the two state groups at the

a two-tiered system during the period under investigation and, hence, its treatment status is ambiguous. While *de jure* all separate low- and intermediate-track schools should have been closed by 2010, both the official statistics and the current data set show some students entering such schools in 2010, indicating that *de facto* the fade-out took longer. It seems that these schools were closed in the following years and students re-assigned. Administrative records show that the cohort's share of students in a low- or intermediate-track school declined from 8% in 2010 to 6% in 2011 to 3% in 2012 (Statistisches Bundesamt, 2011, 2012, 2013). A robustness check where *Rheinland-Palatinate* is assigned the Tracked states (as initially there was some tracking) leaves all results unchanged (see Appendix Table B2).

[19]Formally, let Y^1_{isg} denote the (potential) achievement of student i in state s in grade g under tracking and Y^0_{isg} (potential) achievement under comprehensive schooling. The identification challenge is that the average treatment effect is not generally equal to the observed mean difference between Tracked ($Compr_s = 0$) and Comprehensive ($Compr_s = 1$) states: $\tau_{ATE} = \mathbb{E}[Y^1_{is7} - Y^0_{is7}] \neq \mathbb{E}[Y_{is7}|Compr_s = 0] - \mathbb{E}[Y_{is7}|Compr_s = 1]$.

end of primary school, while δ_2 absorbs general achievement trends between grades. Accordingly, the DD estimate β_{DD} captures the differential development of non-academic-track students in the two-tiered system compared to the three-tiered system. To interpret β_{DD} as the causal effect of comprehensive *versus* tracked schooling we require two assumptions: first, that primary school achievement is indeed unaffected by the structure of the secondary school system and, second, that in the absence of differences in tracking, non-academic-track achievement would have developed in parallel between Comprehensive and Tracked states.[20]

A threat to the first assumption are incentive effects: knowing that they will be placed in different tracks depending on their performance in primary school, students might increase their study efforts already prior to the start of tracking in more tracked regimes (Eisenkopf, 2007). Below I explore the importance of this mechanism directly by comparing the two state groups' achievement trends in primary school. For now, however, note that the test scores used in this paper are not used for students' track placement, ruling out immediate incentive effects related to the tests. Further, note that the presence of the academic track creates strong performance incentives for ambitious students (and their parents) in both regimes, dramatically limiting the importance of this mechanism compared to previous applications.

With respect to the second, 'common trends' assumption, a standard concern in DD designs is sample compositions changing differentially between treatment and control groups between periods. Given that I am comparing one cohort across grade levels this is unlikely to play an important role: students would need to strategically move to another state or from academic to non-academic tracks (or vice versa) *after having started secondary school*. I will confirm that this is not the case by means of balance tests on an array of observed predetermined student covariates and, additionally, condition on these covariates, $\mathbf{X}_{isg}$, in the DD regression.

[20]Continuing the potential outcomes notation from above, formally, we require the following two assumptions:

$$\begin{cases} Y_{is7} = (1 - Compr_s) * Y^1_{is7} + Compr_s * Y^0_{is7} \\ Y_{is5} = Y^0_{is5} \end{cases} \quad \text{(Observation rule)}$$

$$\mathbb{E}[Y^0_{is7}|Compr_s = 0] - \mathbb{E}[Y^0_{is5}|Compr_s = 0] = \mathbb{E}[Y^0_{is7}|Compr_s = 1] - \mathbb{E}[Y^0_{is5}|Compr_s = 1] \quad \text{(Common trends)}$$

Then, DD identifies the average treatment effect on the treated, i.e. the effect of tracked *versus* comprehensive schooling for (non-academic-track) students from the Tracked states: $\tau_{ATT} = \mathbb{E}[Y^1_{is7} - Y^0_{is7}|Compr_s = 0] = \{\mathbb{E}[Y_{is7}|Compr_s = 0] - \mathbb{E}[Y_{is5}|Compr_s = 0]\} - \{\mathbb{E}[Y_{is7}|Compr_s = 1] - \mathbb{E}[Y_{is5}|Compr_s = 1]\} = -\beta_{DD}$. (I define comprehensive schooling as the treatment and tracking as the control condition in the regression formulation because in Germany it is more intuitive to think of the newer two-tiered system as the treatment. This is without loss of generality.)

In the current setting, the more serious threats to the common trends assumption come from two sources: systematic differences between the two state groups in *student composition* and in *schooling inputs*. Regarding the former, there are, indeed, non-negligible differences in student bodies between Comprehensive and Tracked states (see Table 1). If, by the end of primary school, the achievement of different types of students not only differs in levels but also continues to develop differently, DD is biased because it merely removes grade-constant achievement differences. To address this concern, I increase the flexibility of the (conditional) DD model by adding interactions between predetermined student characteristics, $\mathbf{X}_{isg}$, and grade level. This allows for different development trajectories for different types of students. The sensitivity of the DD estimate to this exercise is informative of the extent to which such confounding might play a role.

Turning to the latter, note that different schooling inputs in lower secondary school can be considered 'co-treatments': factors other than tracking that change differently between states between primary and secondary school. In that case, the DD estimate would no longer represent the sole effect of tracking but include the effect of other features of the school environment. To see if differently equipped secondary schools between Comprehensive and Tracked states play an important confounding role, we can proceed as before and inspect the sensitivity of β_{DD} to the addition of school input measures to the control set.

Even if these exercises leave the DD estimate unchanged, concerns about unobserved grade-specific differences between states that violate the common trends assumption might remain. Fortunately, the current setting offers the unique opportunity for an additional test. As explained in the previous section, the distinction between the Comprehensive and Tracked States is only meaningful for students in the non-academic tracks. For academic-track students, there is no difference between the two regimes as they enter *Gymnasium* schools after fourth grade everywhere. Under the assumption that selection into the academic track does not differ between the two state groups, they can, therefore, be used as a control group to test for potential regime-specific trends in achievement that the DD model does not pick up.[21]

This additional control group comparison is easily implemented by the following difference-in-difference-in-differences (DDD), or triple-differences, model, which is estimated over all students

[21]The crucial assumption that selection into the academic track is identical between the two state groups is discussed in further detail and tested below.

and hence adds the subscript $t \in \{academic,\ non\text{-}academic\}$ for track:

$$Y_{istg} = \lambda_{sg} + \phi_{tg} + \mu_{st} + \beta_{DDD}\,(Compr \times Grade7 \times NonAcad)_{stg} + \psi\,\mathbf{X}_{istg} + e_{istg}, \qquad (2)$$

where $NonAcad_t$ is an indicator variable for non-academic-track students, λ_{sg}, ϕ_{tg} and μ_{st} are state-grade, track-grade and state-track fixed effects, respectively, and the remaining variables are defined as before. The triple interaction takes value one for grade 7 observations of non-academic-track students in the Comprehensive states. Accordingly, the DDD estimate β_{DDD} measures how comprehensively taught non-academic-track students progress differently in the first two years of secondary school net of state-specific achievement trends as approximated by academic-track students.

If the estimates for β_{DDD} and β_{DD} are roughly identical this indicates that achievement trends in the academic track are roughly identical in Tracked and Comprehensive states. This should increase one's confidence that there are no state-specific trends confounding the DD estimate from above and that the assumptions for it to be interpreted causally hold. If the two estimates differ, then there are divergent trends in the academic track. Causal interpretation of DDD then hinges on the assumption that academic-track students provide an good approximation of non-academic-track students' counterfactual achievement trends.

In terms of inference, the group-level treatment variable means that I need to account for clustering at the state level when estimating the above regression models (Bertrand *et al.*, 2004; Abadie *et al.*, 2017). As in the current setting there are only twelve states, the large sample assumptions necessary for a conventional cluster robust variance estimator are unlikely to hold (Mackinnon and Webb, 2017). Therefore, throughout this paper, inference is based on a wild cluster bootstrap (Cameron *et al.*, 2008), which has been shown to perform well in settings with few clusters (e.g. Mackinnon and Webb, 2017).[22]

[22]The wild cluster bootstrap permutes the outcome variable based on 'restricted' residuals (i.e. those stemming from coefficient estimates that impose the null hypothesis to be tested) and weights from a Rademacher distribution. Webb (2014) shows that with 12 or less clusters, a specific six-point distribution is preferable over the Rademacher distribution. Hence, I implement the latter. However, results do not substantially differ between the standard (Cameron *et al.*, 2008), an unrestricted (Mackinnon and Webb, 2017) or a schools-as-'sub-clusters'-of-states (MacKinnon and Webb, 2018) version of the bootstrap.

3 Data, Descriptive Statistics and Preliminaries

3.1 Data Sources and Analysis Samples

In this section, I present a brief overview of the data used in this study. For a more detailed discussion of the data sets used, the construction of my samples, as well as sample diagnostics the interested reader is referred to Appendix A.

3.1.1 National Educational Panel Study (NEPS)

The main empirical analysis is based on data from Starting Cohort 3 (SC3) of the multi-cohort German National Educational Panel Study (NEPS) (Blossfeld *et al.*, 2011). The NEPS-SC3 survey randomly sampled from the population of newly minted fifth-graders in the school year 2010/11 and, thereafter, followed this cohort over time as it progressed through grade levels of the German school system.

Student achievement is measured using the NEPS-SC3's competence tests in mathematics and reading.[23] The first round of tests was administered in autumn of 2010, two to four months into students' first year of secondary school. Hence, the grade 5 scores should not yet be severely affected by students' secondary school environment and can be conceptualised as the pre-tracking measure of achievement required by the DD design. Note that the DD estimate is attenuated towards zero to the extent that grade 5 scores are already affected by tracking.[24] Accordingly, my estimates should be conservative. Restricting the sample to students on regular schools in one of the twelve states under investigation, the NEPS-SC3 grade 5 cross-section comprises 4,448 students with non-missing test scores, of whom 2,303 are in the non-academic tracks, of whom 330 are from the Comprehensive states.[25]

[23] While, in principle, the NEPS also assesses competencies in other domains, only in maths and reading testing commenced in the first wave of the survey. As the DD design requires pre-treatment outcomes, my analysis restricts attention to maths and reading achievement.

[24] This is because any differences in achievement between states caused by the first couple of months of tracking are absorbed in the baseline and thus cancelled out in the calculation of the double difference. Note that this unless the effect of tracking reverses within the first couple of months of exposure: if the very short-term effects of tracking (i.e. the effects on achievement after 3 months of exposure) are *opposite* to the longer-term effects of tracking I am trying to estimate (i.e. the effects on achievement after 2 years of exposure), the DD estimate could theoretically by biased upwards. As the effects reproduce in the IQB data, which measures pre-tracking achievement at the end of fourth grade, this does not seem to be the case.

[25] Note that smaller number of observations in Comprehensive states simply reflects that these states are smaller and less populous than the Tracked states (also see the map in Figure 2).

Students were tested again two years later, at the beginning of the 2012/13 school year, when the cohort in question had just entered seventh grade. The NEPS-SC3 grade 7 cross-section comprises 5,316 students with non-missing test scores, of whom 2,771 are in the non-academic tracks, of whom 552 are from the Comprehensive states. It consists of students who were already part of the survey in fifth grade and a large randomly drawn refreshment sample to counteract attrition, explaining the larger sample sizes.

As repeated cross-sections suffice for estimating the main DD and DDD models, for these regressions I simply pool the NEPS-SC3 grade 5 and grade 7 cross-sections. The resulting NEPS DD sample includes only non-academic-track students and comprises 5,074 student×grade observations (882 of which are from the Comprehensive states). The NEPS DDD sample adds academic-track students as an additional control group for, in total, 9,764 student×grade observations (1,711 of which are from the Comprehensive states). I standardise the maths and reading test scores to have mean zero and standard deviation one in the group of Tracked states' non-academic-track students (i.e. the 'control group'), separately by grade level. Accordingly, all treatment effects in this paper can be interpreted in standard deviations of test scores.[26]

In contrast to estimation based only on the panel sample, my use of repeated cross-sections retains students who drop out between waves and includes the refreshment sample. This has two advantages: First, it maximises sample sizes and, hence, precision in the estimation of my main models – a key concern given the NEPS' humble sample sizes in the Comprehensive states. Second, the use of the refreshment sample reduces sample selection bias due to attrition. This is because panel non-response is negatively related to achievement and, hence, attrition is substantially higher in the non-academic tracks (29% compared to 13% in the academic track). The majority of observations in the refreshment sample are from the non-academic tracks (63%), thus restoring the representativeness of my sample. Note that there are no significant differences in student-level attrition between the Tracked and Comprehensive states.[27]

[26]Note that the NEPS competence tests are designed to measure the progress of students *on one scale* across grade levels. This 'linking' of scales is achieved through the recurrence of certain anchor items in each wave of the test (see Fischer *et al.*, 2016, for details). For simplicity and in contrast to an earlier version of this paper (Matthewes, 2018), I nonetheless standardise scores within each grade level. As the DD design identifies a seventh-grade-specific treatment effect it is more intuitive to interpret the estimates in seventh-grade standard deviations. Results are virtually unchanged when using a cross-grade standardisation scheme, however.

[27]See Appendix A for a detailed analysis of panel attrition in the NEPS-SC3.

Nevertheless, for a number of heterogeneity analyses I leverage the panel structure of the NEPS and use the sub-sample of students for whom I observe both grade 5 and grade 7 test scores. This NEPS 5-to-7 panel sample comprises 3,521 students, of whom 1,646 are in the non-academic tracks, of whom 269 are from the Comprehensive states.

Finally, a third round of achievement tests was administered in the 2014/15 school year, when the cohort in question was in ninth grade. The grade 9 tests are used to assess effect persistence.

To probe the robustness of the DD estimates to the inclusion of controls for student characteristics, I draw on detailed information from the NEPS' student and parent questionnaires. In particular, I construct the following student-level control variables: age, sex, migration background, single parent household, foreign language spoken at home, highest level of parental education measured in four categories, monthly household income, receipt of unemployment benefits and a standardised index for home possessions.

Additionally, I use information from the principal and teacher questionnaires to construct the following school-level controls, aimed to capture school quality independent of tracking: average teacher age, days of further training received by teachers over the past year, school size measured by the number of students per cohort, student-teacher ratio and four composite indices for schools' facilities; extracurricular programmes; educational support offers and quality control measures.[28] Note that pre-treatment achievement (i.e. grade 5 scores) is a function of primary school inputs, whereas post-treatment achievement (i.e. grade 7 scores) is a function of secondary school inputs. However, only the secondary school environment is observed in the NEPS-SC3 as it commenced in fifth grade. Thus, to impute the missing primary school inputs in the DD sample, I use data from the NEPS' primary school cohort, Starting Cohort 2 (NEPS-SC2). In particular, I calculate state-level averages for all school-level controls in the primary school data and assign each grade 5 observations in the DD sample its state-level average.

Moreover, I use the NEPS' primary school cohort to investigate pre-tracking achievement trends. For this, I apply the DD model of equation (1) to grade 2 and grade 4 mathematics test scores from the NEPS-SC2.[29] The NEPS-SC2 follows a later cohort than the NEPS-SC3 but, given that there were no major changes to primary education in Germany in this time period, their trends

[28]For more information on these indices see Appendix A.

[29]Unfortunately, reading scores are not available for these grades.

should be similar.

3.1.2 IQB National Assessment Studies

Next to the panel structure that can be exploited for heterogeneity analyses, the main advantage of the NEPS is that it measures student achievement in seventh grade, right after exposure to either treatment condition and when between-state differences in tracking and other school policies are still clear-cut. Its main downside is the modest number of observations in the Comprehensive states, raising concerns about sampling variation. As a robustness check, I therefore double check my results using two large cross-sectional student assessments carried out by the Institute for Educational Quality Improvement (IQB).

The IQB studies do not randomly sample from the population of all students in a particular grade level like the NEPS but, instead, draw separate random samples of roughly similar sizes within each state. Hence, in the IQB data I achieve much larger samples with rough parity in the number of observations between Tracked and Comprehensive states. Due to this sampling design all analyses with the IQB data use student sampling weights to obtain estimates representative of Germany. The IQB data's main downside is that post-treatment outcomes are measured in ninth grade, meaning that the estimates represent a mixture of effect persistence and effects from continued (but somewhat unclear) differences in tracking and other schooling inputs.

The IQB National Assessment Study 2011 (IQB11) tested fourth-graders in maths, reading and listening at the end of the 2010/11 school year, when students were at the end of their primary school time (see Stanat *et al.*, 2012, for details). This is one cohort later than that of the main analysis (see Appendix Figure A1), so that my analysis with the IQB data operates under the assumption that these two consecutive cohorts' primary school experiences match. Fourth-grade students are not yet assigned to academic- or non-academic tracks, but testing happened late enough in the school year for students' secondary school and, hence, track to be determined already. This allows me to classify students as non-academic or academic using information provided by parents and teachers.[30] The IQB11 grade 4 cross-section comprises 18,904 students on regular schools with non-missing test scores, of whom 11,158 are assigned the non-academic tracks, of whom 6,573 are

[30]The details of the classification procedure are described in Appendix A.

from the Comprehensive states.

The IQB National Assessment Study 2015 (IQB15) tested ninth-graders in reading and listening at the end of the 2014/15 school year, which is the same cohort as in the main analysis (see Stanat *et al.*, 2016, for details). All analyses with the IQB15 data restrict attention to students on regular non-academic-track schools with non-missing test scores. The non-academic-track IQB15 grade 9 cross-section comprises 13,742 students, of whom 7,009 are from the Comprehensive states. Analogously to above, I pool the non-academic parts of the IQB11 grade 4 and the IQB15 grade 9 cross-sections to construct the IQB DD sample, which comprises 24,900 student×grade observations (13,582 of which are from the Comprehensive states).

3.2 Descriptives and Balance Tests

The first two columns of Table 1 compare Comprehensive and Tracked states in terms of the distribution of students over tracks (panel A), pre-tracking achievement (panel B) and student characteristics (panel C) in the non-academic-track NEPS DD sample. For reference, column 5 describes academic-track students.

I discuss panel A below. First, note that pre-tracking achievement in panel B is extremely well balanced between the two state groups, as indicated by small and insignificant differences in test scores at the beginning of secondary school. Though pre-treatment balance in outcomes (or covariates) is not technically required by the DD design, this should raise one's confidence that student achievement is generally comparable between Comprehensive and Tracked states. Stark differences in mean scores between non-academic- and academic-track students of (more than) one standard deviation indicate that track assignment is very much a function of achievement despite the absence of strict cut-off rules. Consequently, the treatment variation in the non-academic segment of the school system analysed in this paper concerns a negatively selected group of students. Still, test score distributions of academic- and non-academic-track students overlap substantially (see next section).

In terms of student characteristics, panel C of Table 1 shows moderate compositional differences between the two state groups, highlighting that simple cross-sectional comparisons between states might well be confounded. On the one hand, the Comprehensive states, composed mostly of the

Table 1. *Descriptive statistics and balance tests in the NEPS data.*

	Non-academic tracks				Academic track
	Compr. states	Tracked states	*p*-value (1)=(2)	*p*-value DD=0	(both)
	(1)	(2)	(3)	(4)	(5)
Panel A: Distribution over tracks/school types					
Share non-academic track	0.52	0.52	(0.96)	(0.44)	–
Of those:					
Low-track school	0.00	0.32	–	–	–
Intermediate-track school	0.00	0.55	–	–	–
Comprehensive/SMT school	1.00	0.12	–	–	–
Panel B: Pre-treatment outcomes					
Grade 5 maths score	0.06	0.00	(0.66)	–	1.28
Grade 5 reading score	0.01	-0.00	(0.93)	–	0.99
Panel C: Student characteristics					
Female (binary)	0.47	0.50	(0.02)	(0.92)	0.51
Age in fifth grade (in years)	11.00	10.96	(0.46)	(0.10)	10.67
Single parent household (binary)	0.20	0.13	(0.07)	(0.56)	0.08
Migration background (binary)	0.18	0.32	(0.05)	(0.72)	0.23
Foreign language at home (binary)	0.09	0.14	(0.17)	(0.90)	0.10
Highest parental education level:					
None, low, intermediate w/o appr.	0.16	0.28	(0.07)	(0.84)	0.06
Intermediate w/ apprenticeship	0.48	0.38	(0.31)	(0.59)	0.24
Academic track, some college	0.27	0.25	(0.77)	(0.62)	0.36
University degree	0.09	0.09	(0.91)	(0.13)	0.34
Monthly household income (in Euros)	2781	3072	(0.12)	(0.73)	4070
Unemployment benefits (binary)	0.20	0.11	(0.07)	(0.79)	0.03
Home possessions (index)	-0.31	-0.26	(0.39)	(0.55)	0.24
Observations	882	4192			4690

Notes: Table 1 reports the distribution of students over tracks, as well as variable means for pre-treatment test scores and student covariates in the pooled grade 5 and grade 7 NEPS data. The first two columns describe the NEPS DD sample of non-academic-track students, separately by state group. Academic-track students, who are added as an additional control group in the DDD model, are described in column 5 (for brevity not split by state group). Corresponding to the later regressions, the shares in panel A and grade 5 test scores in panel B are unweighted. The remainder of the table uses student sampling weights to reflect the underlying populations as accurately as possible. Column 3 reports *p*-values from testing whether covariate means are equal in the Comprehensive and Tracked states. Column 4 reports *p*-values from testing for zero double differences (i.e. the second difference between the Comprehensive and Tracked states between grade 5 and grade 7). This tests whether the parallel trends assumption holds for the respective covariate. All tests are based on 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights.

poorer East German states and city states, score slightly worse on socio-economic variables like household income and unemployment. On the other hand, they have lower shares of students with migration background, mainly reflecting the different migration histories of West and East Germany. Importantly, however, column 4 shows that there is no significant double difference in any of the covariates, indicating that these differences in sample composition stay roughly constant between grades. Appendix Table B1 provides a detailed comparison of school characteristics in primary and secondary school in both the NEPS and the IQB data, which has considerably more power for inference at the school level, to show that the same holds true for these. As, especially among the school-level covariates, some of the level differences are not small the analysis below will pay close attention to the sensitivity of the DD estimates to the inclusion of these controls,

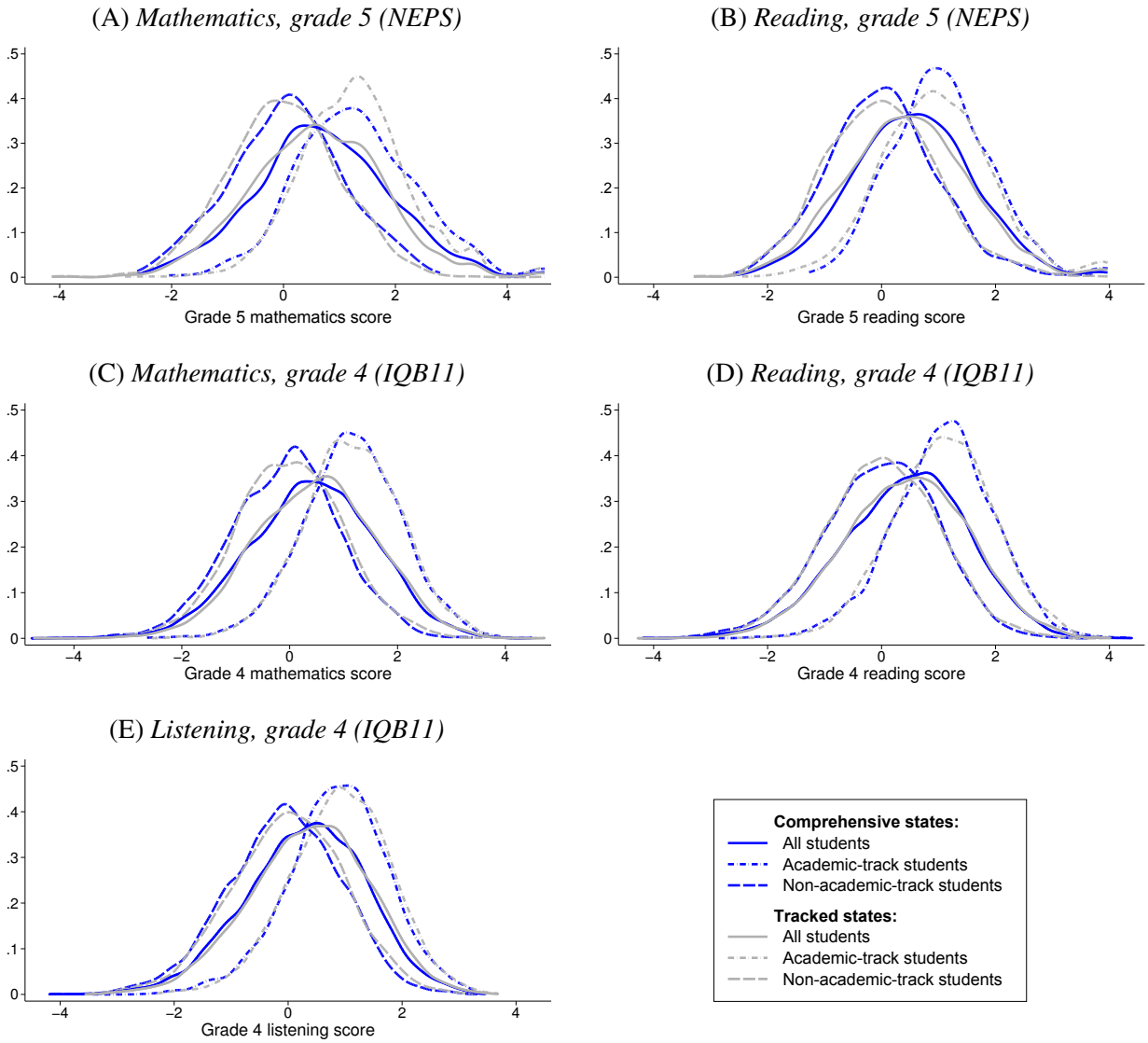

Figure 3. *Pre-tracking test score distributions by track and tracking regime.*

Notes: Figure 3 shows kernel density estimates of different test score distributions for all, only non-academic-track and only academic-track students, separately for Comprehensive and Tracked states. All density estimates use a Gaussian kernel with optimal bandwidth for normally distributed variables (Silverman, 1986). Panels A and B are based on the NEPS-SC3 grade 5 cross-section. Panels C, D and E are based on the IQB11 grade 4 cross-section, using the first plausible value of each test score and student sampling weights.

despite the insignificance of the presented balance tests.

3.3 Selection into the Academic Track

The treatment a student receives depends on her state of residence (Tracked or Comprehensive) and whether she is assigned the academic track or not. My identification strategy requires that selection into the academic track does not differ between the two state groups. Otherwise, neither the academic-track nor the non-academic-track student bodies would be comparable. Panel A of Table 1 shows that the non-academic-track sample shares are 52% in both state groups. Academic-

track students appear to be slightly overrepresented in the NEPS, as according to administrative records the true non-academic shares for the cohort in question are 60% and 57% in Tracked and Comprehensive states, respectively (Statistisches Bundesamt, 2011). Reassuringly, the shares are very similar both in the population and in my sample.

Equal shares leave open the possibility of compositional differences, however. For example, it is conceivable that competition for the academic track is stronger when there are only two tracks, because the alternative school type necessarily comprises all low-achievers. This might amplify average ability differences between academic and non-academic tracks in two-tiered *versus* three-tiered systems. To test for the presence of such differences in selection, Figure 3 plots pre-tracking test score distributions by state group, both overall and for academic and non-academic-track students separately. Panels A and B refer to the (beginning of) grade 5 maths and reading scores from the NEPS and panels C through D refer to the (end of) grade 4 maths, reading and listening scores from the IQB11 data. Across achievement domains and data sets, the distributions look very similar in Tracked and Comprehensive states; in particular, the gaps between the academic- and non-academic-track distributions do not seem to differ between states. Correspondingly, the mean gap between non-academic- and academic-track students does not significantly differ between the two states groups for any of the five scores.[31] Therefore, I conclude that selection into the academic track does not meaningfully differ between the two state groups.[32]

3.4 Distribution over Non-Academic Tracks

Panel A of Table 1 reports the distribution of non-academic-track students over different school types by state group. In the Tracked states one third of students attend low-track schools and about half attend intermediate-track schools. In the counterfactual scenario of a two-tiered school system these two groups would be taught together instead of being separated into different tracks. Note that a small percentage of students in the Tracked states (12%) attends comprehensive schools where

[31] The wild cluster bootstrapped p-values for these 'differences in differences' are 0.82 for maths and 0.60 for reading in the NEPS and 0.58 for maths, 0.55 for reading and 0.18 for listening in the IQB data.

[32] It might seem puzzling that alternative choice options are largely irrelevant for selection into the academic track. It likely is explained by the special status of the academic-track *Gymnasium* in Germany: virtually all ambitious and high-SES students will aspire to the academic track regardless of what other school forms are present because of its reputation and academic focus (Paulus and Blossfeld, 2007). For example, in my sample 78% of students with college-educated parents attend the academic track.

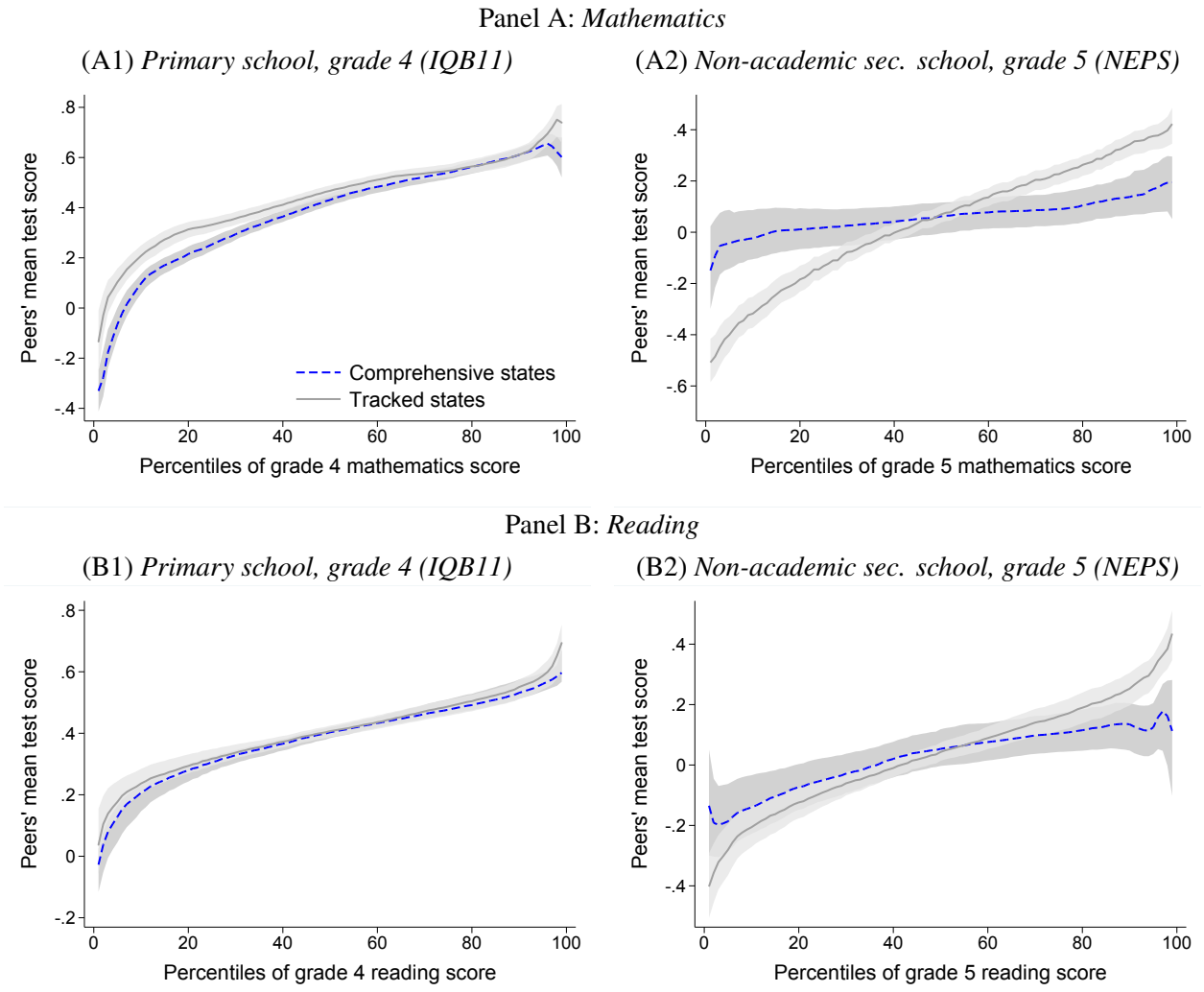

Figure 4. *Effect of tracking on peer group composition.*

Notes: All curves show fitted values from student-level local constant regressions of mean classroom test scores in maths (Panel A) and reading (Panel B) on students' own test score, separately for the Tracked and Comprehensive states. The fitted values are evaluated at each percentile of the respective test score distributions. The left-hand side figures are based on the IQB11 grade 4 cross-section (restricted to students classified as non-academic), thus describing the relation between own and classroom peers' performance at the end of primary school, right before tracking commences. The right-hand side figures are based on the NEPS-SC3 grade 5 cross-section (restricted to non-academic-track students), thus describing the same relation a couple of months later, when students have been tracked according to the state-specific rules. The shaded areas show pointwise 95% confidence intervals from 999 iterations of a percentile bootstrap, clustering at the classroom level.

all three degrees can be obtained and there might or might not be within-school streaming.[33] In the language of the treatment effects literature, these students can be thought of as 'always-takers', slightly attenuating my 'intent-to-treat' effect estimates towards zero.

In the Comprehensive states, there are no low- and intermediate-track schools. All non-academic-track students in these states attend SMT or comprehensive schools. As explained above, there is no within-school streaming in grades 5 and 6 in these schools. It is the effect of this comprehensive schooling for non-academic track students in the two-tiered regime, as compared

[33]These sample shares are very close to the true population shares: 32% low-track, 51% intermediate-track and 17% comprehensive schools (Statistisches Bundesamt, 2011).

to the between-school tracking in the three tiered-regime, that I aim to estimate.

3.5 Peer Group Composition

To get a better idea of what these differences in tracking mean from the perspective of students, Figure 4 explores how tracking affects students' peer group composition. Using the IQB11 (end of) grade 4 test scores in maths (panel A) and reading (panel B), the left-hand panels compare the relationship between students' own achievement and that of their classroom peers between Comprehensive and Tracked states in the final year of primary school, right before tracking commences. Using the NEPS (beginning of) grade 5 scores, the right-hand panels depict the same relationship a couple of months later,[34] when students have been tracked according to the state-specific rules.

Despite the absence of tracking, the relationship between students' own achievement and that of their peers is clearly positive already in primary school, representing residential sorting.[35] Importantly, however, this relation looks very similar in both state groups. If anything, the gradient for maths is slightly steeper in the Comprehensive states. With the transition to secondary school, this relationship changes quite dramatically between the two state groups. Now, the gradient is much steeper in the Tracked states, where non-academic-track students are assigned to different schools based on their previous performance, than in the Comprehensive states, where classes continue to be formed disregarding previous performance.[36] These differences in classroom heterogeneity form the core of the (composite) treatment of comprehensive *versus* tracked schooling.

4 Results

4.1 Level Effects of Comprehensive *versus* Tracked Schooling

This section presents my findings on the average effect of comprehensive, as compared to tracked, schooling in the first two years of lower secondary school for non-academic-track students. For

[34]Note that this is under the assumption that this relation stayed constant between the two consecutive cohorts that are tested in IQB11 and NEPS-SC3.

[35]There are much more primary than secondary schools in Germany, such that students generally attend schools closer to home and residential sorting plays a larger role.

[36]Note that one cannot meaningfully compare the slopes between primary and secondary school without very strong assumptions, as IQB and NEPS differ in their test design and thus do not measure achievement on the same scale. Therefore, I restrict attention to between-state comparisons within (and not across) data sets.

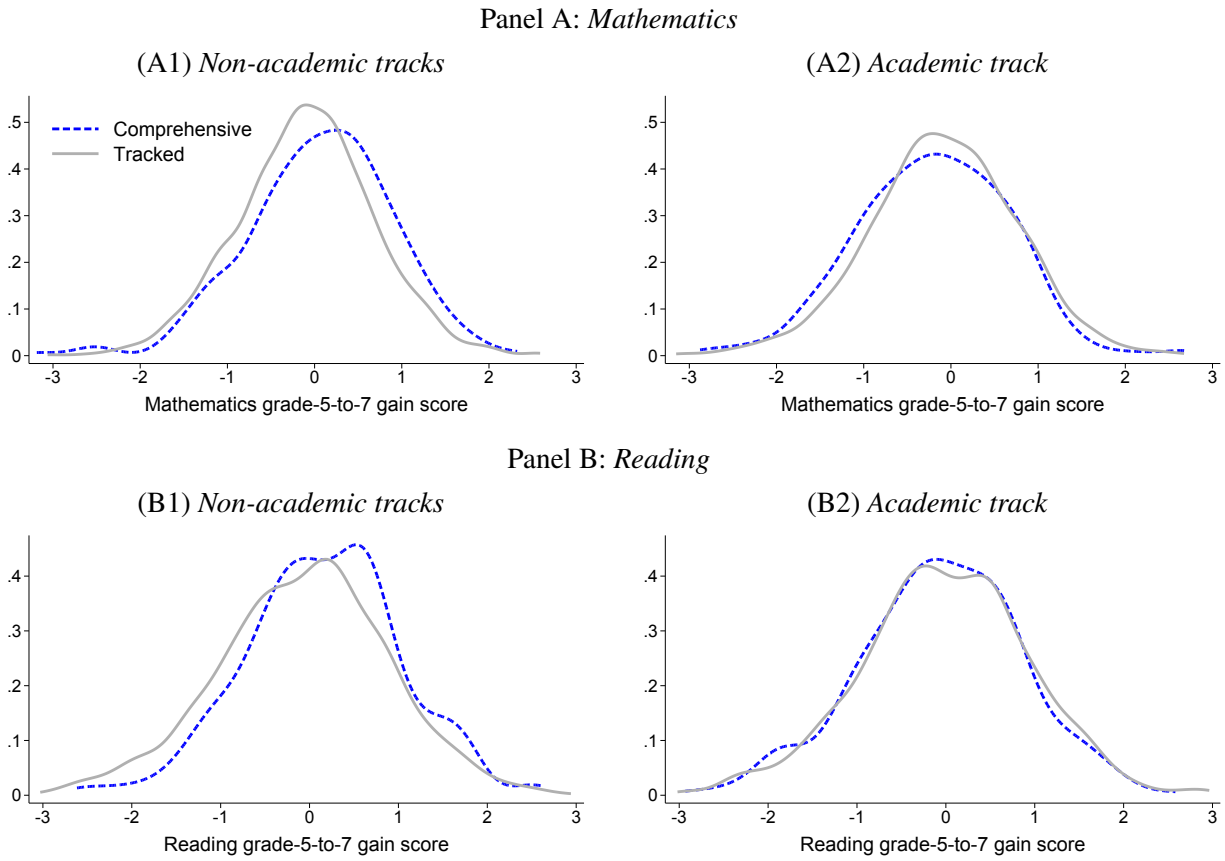

Figure 5. *Distribution of grade-5-to-7 gain-scores by track and tracking regime.*

Notes: Figure 5 shows kernel density estimates of the grade-5-to-7 gain score distribution in maths (Panel A) and reading (Panel B), separately for Comprehensive and Tracked states, for non-academic-track (left) and academic-track students (right). All density estimates use a Gaussian kernel with optimal bandwidth for normally distributed variables (Silverman, 1986). Estimates are based on the NEPS 5-to-7 panel sample ($N = 1,646$).

illustrative purposes, I begin by comparing students' progress between Tracked and Comprehensive states graphically. Using the NEPS 5-to-7 panel sample, Figure 5 plots kernel density estimates of gain-score ($\Delta_7 Y_{is} = Y_{is7} - Y_{is5}$) distributions in maths and reading for academic- and non-academic-track students in both tracking regimes. A striking picture emerges: whereas academic-track students' progress is very similar between regimes in both domains – if anything, those in the Tracked states progress slightly more in maths – for non-academic-track students the Comprehensive states' distribution of gains appears to stochastically dominate that of the Tracked states. These graphs provide strong initial evidence for the existence of efficiency gains from *comprehensive* schooling. In the following, I assess the significance and robustness of this descriptive finding more formally by estimating the DD and DDD models.

Column 1 of Table 2 displays the regression results for the simple (unsaturated) DD model, corresponding to equation (1) without control variables and state fixed effects, estimated using the

non-academic NEPS DD sample. The results for maths are presented in panel A and those for reading in panel B. Next to point estimates, in parentheses I present p-values and in brackets 95% studentised bootstrap confidence intervals from 999 wild cluster bootstrap iterations, clustering at the state level.[37] As indicated by small and insignificant coefficients on the Comprehensive states indicator, there seem to be no substantial level achievement differences between the two state groups prior to tracking. The DD coefficients, equal to the (double) difference between Comprehensive and Tracked states' achievement changes between grades 5 and 7, indicate that comprehensively taught non-academic-track students progress about 0.18 standard deviations (SD) more in maths and 0.26 SD more in reading, confirming the graphical finding from above.

The next columns probe the robustness of this result, following the steps outlined in section 2: Column 2 presents the saturated DD model, which replaces the Comprehensive states indicator with state fixed effects for increased flexibility and precision. Column 3 adds the complete set of student covariates, described in Table 1, to correct for potential compositional changes in the sample. Column 4 interacts the vector of student covariates with the grade 7 indicator to allow for different development trajectories for different types of students. Finally, column 5 adds the complete set of school covariates to control for potential differences in schooling inputs.

Considering the overall level of imprecision in the estimates due to the moderate number of observations in the Comprehensive states, the DD estimate stays remarkably stable across all specifications. As the student control variables in the NEPS are very detailed, substantially increasing the model's explanatory power as evidenced by the sharp increase in R^2 between columns 2 and 3, it is very unlikely that between-state differences in student body composition explain the advantage for comprehensively taught students. Despite my controls for schooling inputs being somewhat less detailed, the fact that the DD estimate *increases* upon their inclusion provides strong evidence against the (null) hypothesis that the effects are driven by differences in schooling inputs. Appendix Table B2 further corroborates the robustness of the DD results to alternative model specifications and the inclusion of further potential confounders at the school (e.g. private schools and class size) and state level (e.g. binding track recommendations and school

[37]Note that in my case inference based on the wild cluster bootstrap is strictly more conservative than inference based on conventional cluster-robust standard errors, clustered at the state level (which in turn is more conservative than clustering at the state-track level, which in turn is more conservative than clustering at the school level). For details on the wild cluster bootstrap implementation used in this paper see Roodman *et al.* (2019).

Table 2. *Level effect of comprehensive schooling on seventh-grade achievement.*

Model specification:	Double differences					Triple differences			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Mathematics									
Comprehensive schooling	0.176***	0.207***	0.167***	0.165***	0.259**	0.257**	0.214*	0.199*	0.187*
	(p = 0.01)	(0.01)	(0.00)	(0.01)	(0.04)	(0.03)	(0.07)	(0.10)	(0.10)
	[0.07, 0.40]	[0.10, 0.49]	[0.10, 0.29]	[0.09, 0.28]	[0.02, 0.60]	[0.03, 0.54]	[-0.01, 0.51]	[-0.04, 0.51]	[-0.03, 0.46]
Indicator Compr. states	0.061								
	(0.65)								
	[-0.33, 0.31]								
R^2	0.005	0.027	0.172	0.177	0.199	0.296	0.378	0.380	0.394
Panel B: Reading									
Comprehensive schooling	0.264**	0.277**	0.241**	0.243**	0.363***	0.293	0.235	0.231	0.286**
	(0.02)	(0.02)	(0.04)	(0.02)	(0.00)	(0.12)	(0.14)	(0.14)	(0.03)
	[0.04, 0.59]	[0.05, 0.64]	[0.02, 0.53]	[0.04, 0.48]	[0.20, 0.60]	[-0.06, 0.73]	[-0.07, 0.57]	[-0.06, 0.54]	[0.02, 0.59]
Indicator Compr. states	0.012								
	(0.94)								
	[-0.34, 0.32]								
R^2	0.007	0.021	0.108	0.113	0.131	0.220	0.273	0.276	0.285
Individual controls			✓				✓		
Grade × Ind. controls				✓	✓			✓	✓
School controls					✓				✓
State FE		✓	✓	✓	✓	✓	✓	✓	✓
State×grade FE						✓	✓	✓	✓
State×track FE						✓	✓	✓	✓
Track×grade FE						✓	✓	✓	✓
N state clusters	12	12	12	12	12	12	12	12	12
N Compr. state students	882	882	882	882	882	1711	1711	1711	1711
N Tracked state students	4192	4192	4192	4192	4192	8053	8053	8053	8053

Notes: Table 2 reports OLS regression results for the double- (DD) and triple-differences (DDD) models for fifth- and seventh-grade maths and reading test scores. The DD models in columns 1–5 are estimated using the NEPS DD sample of non-academic-track students. For the DDD models in columns 6–9 academic-track students are added to the sample. Column 1 reports results for the unsaturated DD model, i.e. from regressing test scores on an intercept, an indicator for the Comprehensive states, an indicator for grade 7 observations and their interaction. Column 2 report results for the saturated DD model, which replaces the Comprehensive state indicator with state fixed effects. Column 3 adds student covariates: sex, age, age squared, migration background, foreign language spoken at home, single parent household, household income, parental unemployment, parental education and an index for home possessions (incl. missing data indicators). Column 4 adds interactions between the grade 7 indicator and all student covariates. Column 5 adds school covariates: average teacher age, average days of further training received by teachers, school size, student-teacher ratio and indices for schools' equipment, extracurricular programmes, educational support offers and quality control measures (incl. missing data indicators). Column 6 reports the results for the saturated DDD model, i.e. from regressing all students' test scores on an indicator for non-academic-track grade 7 observations from the Comprehensive states and state-grade, state-track and grade-track fixed effects. Columns 7–9 add covariates to the DDD model analogously to before. p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

funding).[38]

Despite the stability of the DD estimate across different control sets of varying flexibility, concerns about non-parallel (counterfactual) achievement trends might remain. Hence, as a more design- and less covariate-based test for the common trends assumption, column 6 presents the uncontrolled DDD model that uses academic-track students, for whom there are no differences in tracking between states, as an additional control group. Columns 7–9 add control variables analogously to before. As to be expected the DDD estimates are less precise than the ones from DD, but, reassuringly, they are very similar in magnitude – if anything, slightly larger. The similarity between the double- and triple-difference estimates implies that there are no divergent achievement trends between the Comprehensive and Tracked states in the academic track, which suggests that the different trends in the non-academic tracks are indeed due to differences in tracking.

Given the small number of states it is important to check that the results are not driven by any particular outlier state whose performance diverged extremely from the others. To this end I perform a simple leave-one-out analysis. Figure 6 plots coefficient estimates against p-values from repeatedly re-estimating the DD and DDD models eat time leaving out one state. While the precision is slightly affected when some (larger) states are dropped, the results appear robust to the exclusion of any particular state.

Given that differences in tracking between states only emerge with students' transition to secondary school, a natural requirement for interpreting my results causally is that Comprehensive and Tracked states exhibit parallel achievement trends in primary school, prior to tracking. Inspecting such pre-trends also allows to test for the presence of the discussed anticipation effects of tracking

[38] In particular, I first add all school controls separately to ensure that coefficient movements are largely homogeneous across variables (columns 2–9). Then, I show robustness to applying student sampling weights (column 11) and using the *unsaturated* DD specification (column 12). Next, I interact the school controls with grade level just like the student controls (column 13). This drastically increases imprecision, as the data lacks power to allow this degree of flexibility at the school level – especially since in grade 5 these variables only vary at the state level. Still, point estimates are rather similar. Finally, in columns 14–19 I show that results are fully robust to controlling for a school-level indicator for private schools (which is excluded in the main regressions because there are so few private schools in the NEPS that the imputed primary school state averages are highly unrepresentative); a school-level measure of average class size (which is excluded because it is arguably a 'bad control': class sizes are likely to be a function of how students are sorted – e.g. low-track schools tend to have smaller class sizes); the time students are back in school since the end of the summer break at the day of testing; a state-level measure of per pupil public expenditure for schools (which is excluded because its comparability across states is somewhat doubtful, mainly because both living expenses and teacher salary-unrelated expenditures vary greatly between states); a state-level indicator for binding teacher recommendations (which is excluded due to its irrelevance – see section 2.1); and adding *Rheinland-Palatinate* to the sample as a Tracked state.

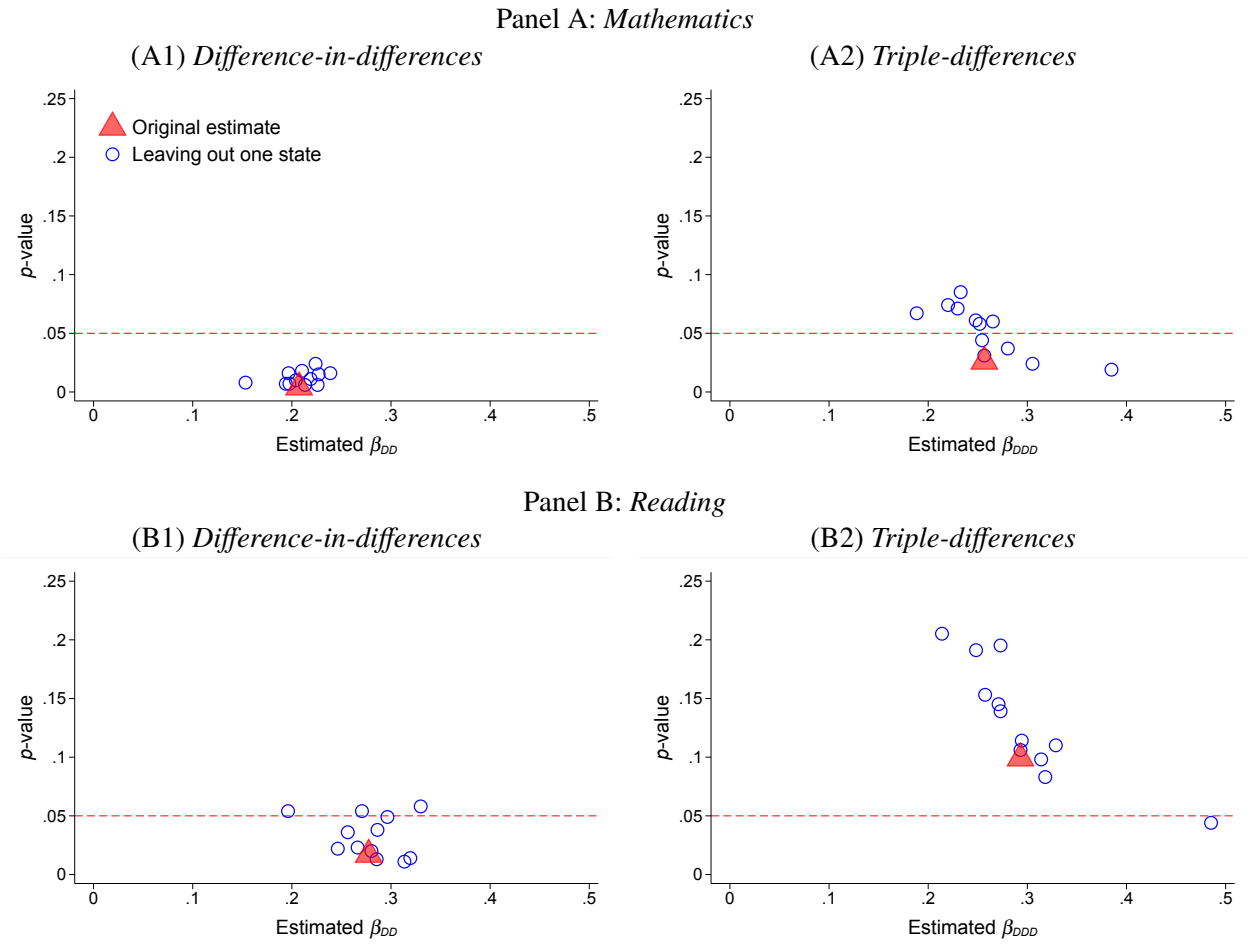

Figure 6. *Leave-one-state-out DD and DDD estimates.*

Notes: Figure 6 compares the DD and DDD estimates for the effect of comprehensive schooling on seventh-grade maths (Panel A) and reading (Panel B) scores in the original sample with those obtained when excluding single states. Point estimates on the horizontal axis are plotted against *p*-values testing the null of no effect on the vertical axis, obtained from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Each panel has 13 data points: one triangle showing the original estimate and 12 circles showing the estimates when excluding each state. State names cannot be revealed for data confidentiality reasons.

in secondary school on pre-tracking achievement. To this end, I first apply the unsaturated DD model from equation (1) to grade 2 and 4 maths scores in the NEPS primary school cohort and find no differences in second-grade achievement ($\delta_1 = 0.043$; $p = 0.77$) or achievement growth in the two years prior to tracking ($\beta_{DD} = -0.015$; $p = 0.77$).[39]

This result concerns all and not only non-academic-track students: as students are not yet assigned to tracks in primary school I cannot restrict the sample accordingly. To investigate pre-trends specifically for lower achieving students, who are more likely to attend non-academic-track schools later on, second, I leverage the NEPS' panel structure and investigate achievement growth *by previous achievement*. In particular, using the NEPS 2-to-4 panel sample comprising all students

[39]See Appendix Table B3 for the complete regression results. Note that I cannot inspect pre-trends for reading, as reading scores are only available for grade 4.

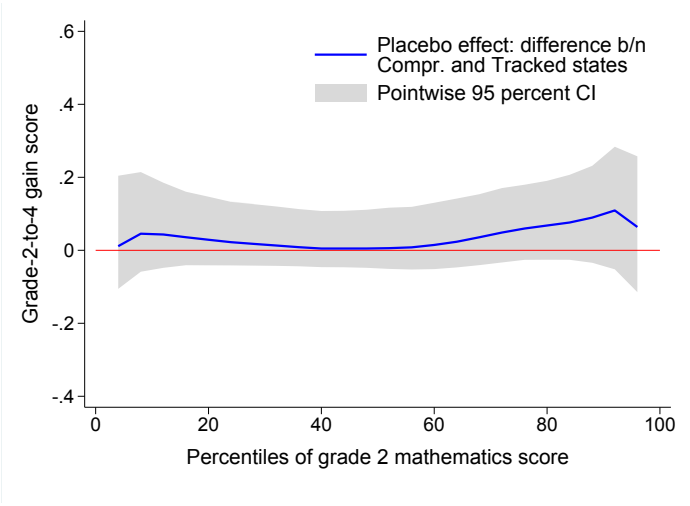

Figure 7. *Pre-tracking achievement trends by previous performance.*

Notes: Figure 7 shows the difference in average second to fourth grade achievement growth between Comprehensive and Tracked states across the grade 2 maths score distribution. Estimation is based on the primary school NEPS-SC2 2-to-4 panel sample that includes all students for whom both grade 2 and grade 4 maths scores are observed ($N = 4,676$). The curve is constructed as follows: First, separately for Comprehensive and Tracked states, I estimate a student-level local constant regression of grade-2-to-4 gain scores on grade 2 test scores. Second, I calculate the difference between Comprehensive and Tracked states' fitted values at every fourth percentile of the grade 2 distribution. Third, I construct pointwise 95% confidence intervals from a percentile bootstrap with 999 iterations, clustering at the state level, stratifying by tracking regime and holding the bandwidth constant across bootstrap iteration (Hall and Kang, 2001).

for whom I observe both second and fourth grade maths scores, I non-parametrically estimate the difference in grade 2-to-4 gain scores between Comprehensive and Tracked states at different percentile of the grade 2 test score distribution.[40] The results, presented in Figure 7, indicate that achievement growth in the two years prior to tracking is indeed roughly parallel across the entire achievement distribution.

In summary, this section showed that achievement levels in the non-academic tracks diverge between Comprehensive and Tracked states during the first two years of secondary school. The presented evidence suggests that this divergence is caused by differences in ability grouping: comprehensive, instead of between-school tracked, schooling at the ages 10 through 13 appears to boost achievement for non-academic-track students – a group comprised of low and, as is visible from the pre-tracking achievement distributions displayed in Figure 3, also a considerable share of medium to high achievers. My preferred specification is the DD model in column 4 of Table 2, which flexibly controls for student characteristics but omits school-level controls that, despite being chosen carefully to only include inputs that are independent of a state's tracking policy, might raise concerns of controlling for mechanisms. Here, 95% confidence sets for the effect of comprehensive

[40]Formally, I estimate the two conditional expectations $\mathbb{E}[\Delta_4 Y_{is} | Y_{is2}, Compr_s = k]$, for $k \in \{0,1\}$, non-parametrically and evaluate their difference at every fourth percentile of Y_{is2}.

schooling are [0.09, 0.28] in maths, with a point estimate of 0.17 SD, and [0.04, 0.48] in reading, with a point estimate of 0.24 SD. Whilst the small sample size prohibits a more precise estimation of the average effect, these results strongly reject tracking proponents' claim that comprehensive schooling impedes achievement. As comprehensive systems reduce the homogeneity of classrooms in terms of ability, these findings are at odds with the notion that there is a monotonously positive relation between classroom homogeneity and performance.

How large are these estimates? The point estimate of 0.17 SD in maths is roughly half the female-male achievement gap in maths (0.35), roughly one-third of the migration-native gap (0.50) and roughly one-fifth of the gap between children of parents from the lowest and the highest education category (0.94). The point estimate of 0.24 SD in reading is roughly double the male-female achievement gap in reading (0.11), roughly half the migration-native gap (0.42) and roughly one-fourth of the parental education gap (0.92).[41] Note that the effect sizes are measured in non-academic-track standard deviations. They are marginally smaller when measured in terms of the overall student population at 0.15 SD in maths and 0.21 in reading. Still, they are larger than the zero effect found by Hanushek and Wössmann (2006) at the age of 15. However, most of the tracked countries in their sample start tracking students at much later ages than considered here, when effects can generally be expected to be smaller. Importantly, they find that comprehensive schooling decreases the dispersion of test scores, indicating that weaker students benefit, which can reconcile these findings. Somewhat similarly, Kerr *et al.* (2013) find very small average effects of a Finnish comprehensive schooling reform and larger positive effects for disadvantaged students. My estimates are similar in magnitude to the one found by Garlick (2018) on South-African college students' GPA who are either (ability) tracked or randomly assigned to student dormitories. Also those are driven by low-achievers. Hence, in the following I will explore the heterogeneity of these average effects.

4.2 Effect Heterogeneity

As a first step to go beyond average effects, I extend the logic of the DD estimator and, instead of limiting attention to the mean, inspect how the whole achievement distribution changes differently

[41] These figures refer to NEPS grade 5 test score gaps for non-academic-track students.

between Comprehensive and Tracked states from grade 5 to 7.[42] Let $f_g^C(\cdot)$ be the density of non-academic-track students' grade g test scores for the Comprehensive states. The difference $f_7^C(y) - f_5^C(y)$ measures the change in the density at level $Y_{isg} = y$ between grades 5 and 7 for this group. $f_7^T(y) - f_5^T(y)$ is the equivalent change for the Tracked states. Comparing these two quantities across the support of the test score distribution allows me to map out the distributional consequences of comprehensive schooling:

$$\{f_7^C(y) - f_5^C(y)\} - \{f_7^T(y) - f_5^T(y)\} \tag{3}$$

Figure 8 presents the results from this exercise. Panels A and B plot density estimates for grade 5 and grade 7 maths scores respectively for Comprehensive and Tracked states. In the former the distribution appears to tighten slightly between grades, whereas in the latter it stays relatively constant. As the differences between the densities are small relative to the scale, panel C plots the vertical distances between the grade 5 and 7 densities by state group. Thus, these lines describe how the shape of the test score distributions changes between grades. Finally, panel D plots the vertical distance between these two lines, corresponding to the expression in equation (3). It appears that Comprehensive schooling shifts probability mass from the bottom end of the distribution (approximately from the range [-2.5, -0.5]) to the middle part (approximately to the range [-0.5, 1.5]). The picture for reading scores is very similar (see Appendix Figure B1). This means that, next to a positive average effect, comprehensive schooling has an equalising effect on the achievement distribution.

Figure 4 revealed that a state's tracking regime strongly affects peer group composition – but differently for students at different positions in the previous achievement distribution: in the Tracked states, high-achieving students study together with higher achieving peers and low-achieving students study together lower achieving peers, whereas in the Comprehensive states a student's peer group depends much less on her own performance. If peer quality matters, the effect of comprehensive schooling is therefore likely to vary with students' previous achievement. In fact, it might well be that certain groups of students lose out from being taught comprehensively, but

[42] Neumark *et al.* (2004) proposed this method to estimate the effect of minimum wages on the distribution of family income.

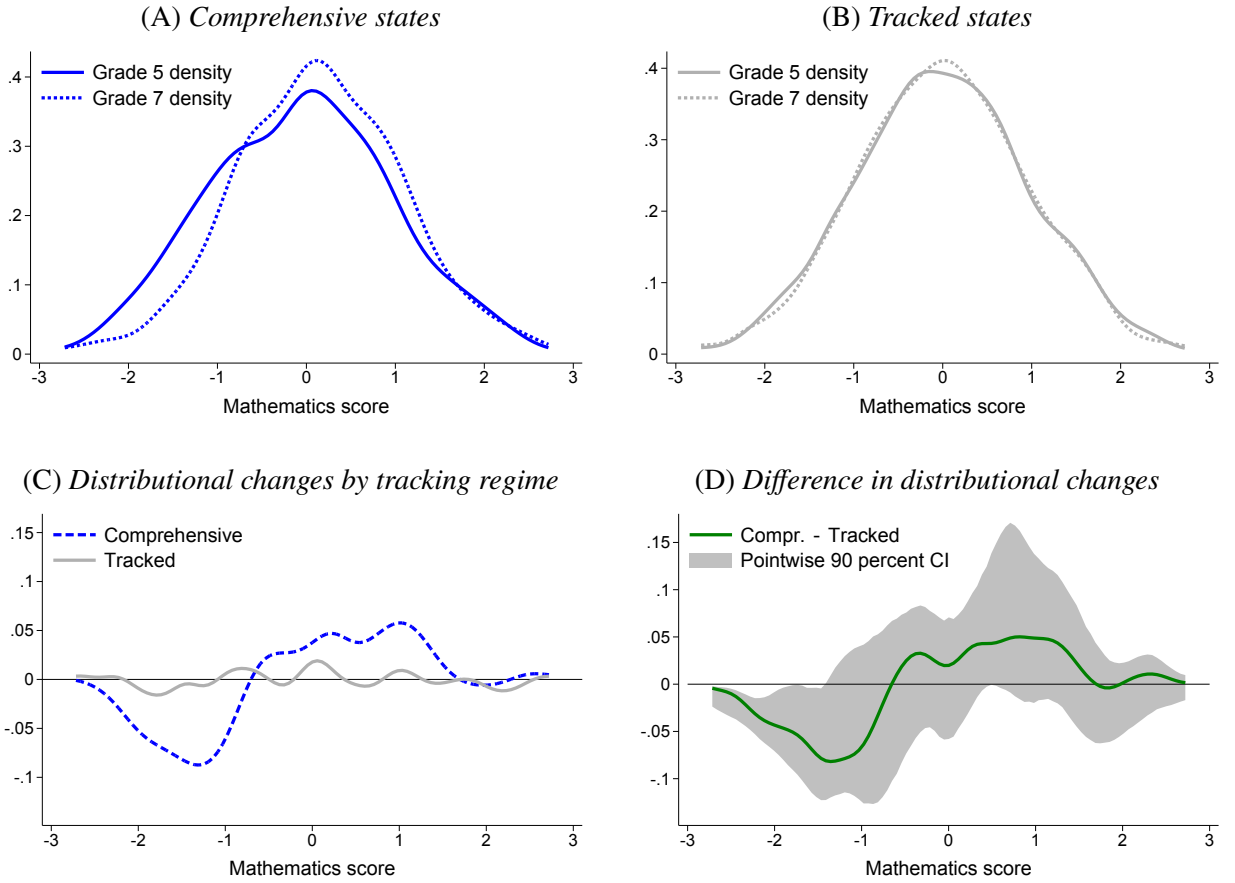

Figure 8. *Maths score distributions before and after treatment exposure.*

Notes: Figure 8 describes how the test score distribution in maths changes differently between fifth and seventh grade depending on the tracking regime. Panels A and B, respectively for Comprehensive and Tracked states, display kernel density estimates for non-academic-track students' grade 5 and grade 7 scores at 100 equally spaced points between the 0.5th and 99.5th percentiles of the (cross-grade) maths score distribution. Estimation is based on the NEPS DD sample and a Gaussian kernel with optimal bandwidth for normally distributed variables (Silverman, 1986). Panel C plots the between-grade differences in the estimated densities at each point, separately for Comprehensive and Tracked states. Panel D plots the difference between Comprehensive and Tracked states in the between-grade density differences, including pointwise 90% confidence intervals from a percentile bootstrap with 999 iterations, clustering at the state level and stratifying by tracking regime. The bandwidth of the kernel estimator is held constant at its optimal level for the original sample in each bootstrap iteration (Hall and Kang, 2001).

that these losses are compensated by the gains of other groups, resulting in a positive net effect.

To explore this possibility I use the NEPS 5-to-7 panel sample of non-academic-track students, which allows me to match students on previous achievement. Note that in the panel sample the average effects are marginally smaller: a simple 'value-added model' that regresses grade-5-to-7 gain scores on grade 5 scores and an indicator for the Comprehensive states gives effect estimates of 0.15 SD for maths ($p = 0.04$) and 0.23 SD for reading ($p < 0.01$).[43] If it is low-achieving

[43]The DD model estimated on the NEPS 5-to-7 panel sample gives virtually identical results to the value-added (VA) model. Appendix Table B4 presents both the VA model (in column 6) and the DD model (in column 5). The latter is labelled 'first-differenced' (FD) model because with individual-level panel data the DD model can be rewritten in FD form: $\Delta_7 Y_{is} = Y_{is7} - Y_{is5} = \delta_2 + \beta_{DD}\, Compr_s + \Delta u_{is}$. Note that the DD/FD model and the VA model are non-nested: The former controls for grade-constant heterogeneity that correlates with both treatment and outcomes, whereas the latter controls for selection into treatment based on (previous) outcomes (Angrist and Pischke, 2009). As my goal is to control for unobserved differences *between states*, instead of controlling for selection at the *individual level* the former seems more appropriate for the context at hand and is used in the main models. The VA model is presented here to

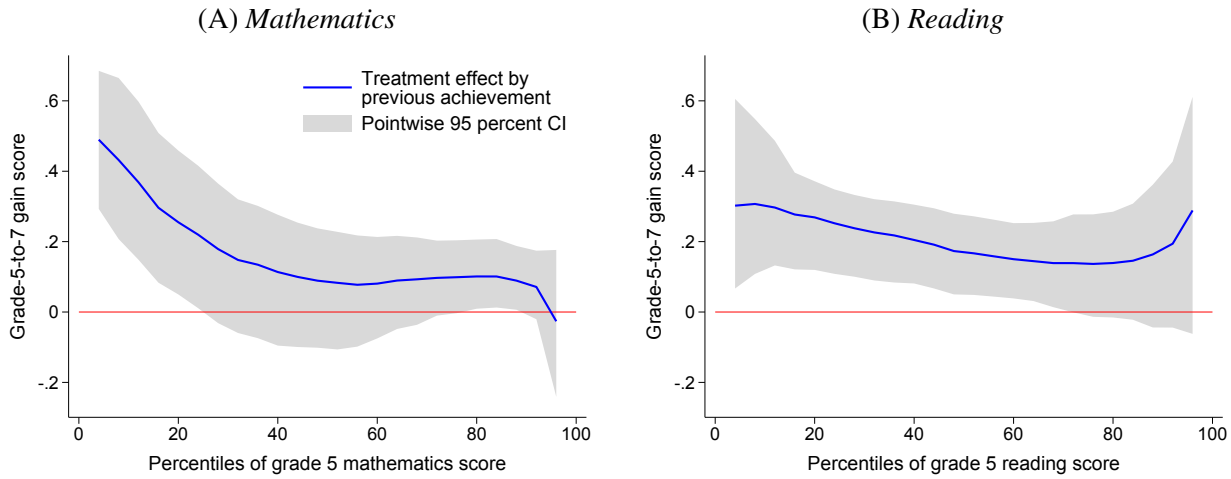

Figure 9. *Effect heterogeneity by previous achievement.*

Notes: Figure 9 shows the difference in average fifth to seventh grade achievement growth between Comprehensive and Tracked states across the grade 5 test score distribution in maths (Panel A) and reading (Panel B). Estimation is based on the non-academic-track NEPS 5-to-7 panel sample ($N = 1,646$). The curves are constructed as follows: First, separately for Comprehensive and Tracked states, I estimate a student-level local constant regression of students' grade-5-to-7 gain scores on grade 5 test scores. Second, I calculate the difference between Comprehensive and Tracked states' fitted values at every fourth percentile of the grade 5 test score distribution. Third, I construct pointwise 95% confidence intervals from a percentile bootstrap with 999 iterations, clustering at the state level, stratifying by tracking regime and holding the bandwidth constant across bootstrap iteration (Hall and Kang, 2001).

students who benefit from comprehensive schooling, then the slightly smaller estimates for the average effect might be explained by the fact that low-achievers are more likely to drop out between waves and thus are slightly under-represented in the panel sample. To assess effect heterogeneity by previous achievement explicitly, analogously to above, I estimate the two conditional expectations $\mathbb{E}[\Delta_7 Y_{is} | Y_{is5}, Compr_s = k]$, for $k \in \{0,1\}$, non-parametrically and evaluate their difference at different percentiles of Y_{is5}. This identifies the effects of comprehensive schooling throughout the pre-tracking achievement distribution.

The results for maths in panel A of Figure 9 reveal that the effect exhibits a steep gradient with respect to previous achievement: effects appear to be monotonically decreasing in grade 5 test scores in the first half of the distribution before flattening out, with large and significant effects from 0.5 to 0.2 SD in the bottom quartile, smaller and insignificant effects from 0.2 to 0 SD in the second quartile and roughly zero effects for all remaining students. In the results for reading in panel B the gradient by previous achievement is also visible but less pronounced: effects are significant and positive from the first through the 65th percentile of the grade 5 distribution, monotonically decreasing from about 0.3 to 0.1 SD, and larger but very imprecisely estimated and insignificantly

show the robustness of the results to this alternative modelling of the selection process and to provide an average effect benchmark for the VA-style heterogeneity analysis below.

different from zero thereafter.

These results imply that it is low achievers – and, to the extent that grade 5 achievement measures ability, low-ability students – who drive the positive level effects found before. They seem to benefit immensely from studying together with their higher achieving peers in a more demanding scholastic environment for another two years, especially in maths. Importantly, I do not find a negative effect at any point of the achievement distribution, meaning that higher achievers do not seem to lose out from learning together with their lower achieving peers. Remember, while non-academic-track students are a negatively selected group with substantially lower test scores than academic-track students on average, Figure 3 shows that the distributions of these groups overlap substantially. The top 25% non-academic-track students would be above-median students even in the academic track.

To investigate effect heterogeneity along other dimensions, in Appendix Table B3 I present results from fully interacting the DD model with indicators for female, low socio-economic status (SES) and migration background students. All treatment-covariate interactions are insignificant and without clear directional pattern across maths and reading scores but, clearly, the analysis is underpowered to detect smaller effect heterogeneities. Regardless, the striking pattern found above suggests that previous achievement is the most important dimension for effect heterogeneity in the current context.[44] Several studies found that especially low-SES students benefit from comprehensive schooling (e.g. Kerr *et al.*, 2013) but in the selected group of non-academic-track students investigated here SES differences are not very pronounced to begin with. For instance, in my sample, only 22% of students with college-educated parents even attend a non-academic-track school. There is a much more salient socio-economic divide between academic- and non-academic tracks than between different school types in the non-academic segment. Accordingly, the first-order effect of the treatment of comprehensive schooling in my setting is the mingling of students of different abilities rather than of different socio-economic backgrounds.

[44]This conclusion is corroborated by the fact that some sizeable (but insignificant) interaction effects for reading decrease in magnitude when repeating this exercise in the value-added model that controls for previous achievement (see Appendix Table B3).

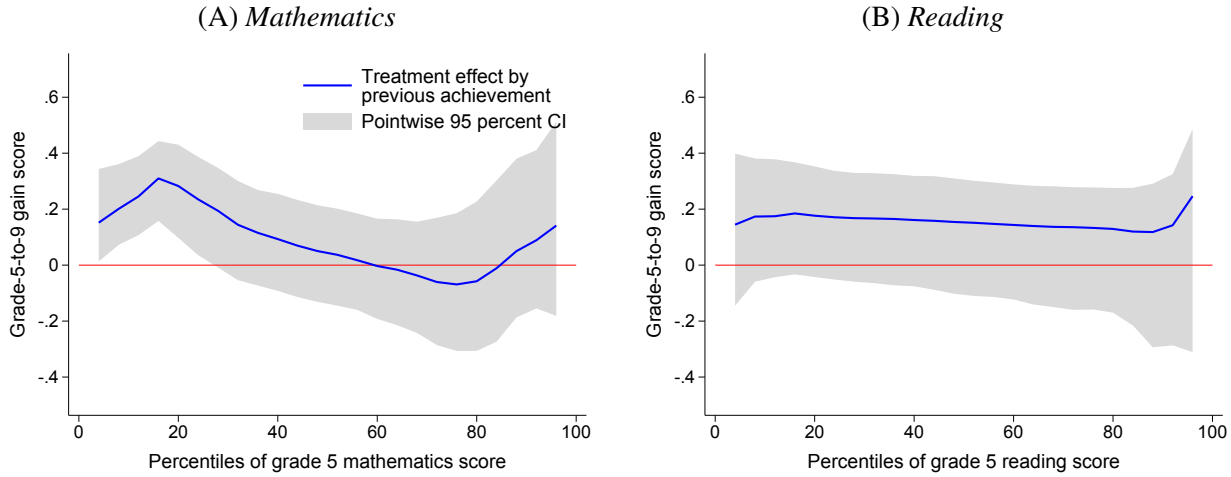

Figure 10. *Effect persistence by previous achievement.*

Notes: Figure 10 shows the difference in average fifth to ninth grade achievement growth between Comprehensive and Tracked states across the grade 5 test score distribution in maths (Panel A) and reading (Panel B). The curves are based on the non-academic-track NEPS 5-to-9 panel sample comprising all students for whom I observe test scores in both grades ($N = 1,286$ for maths and $N = 1,255$ for reading). They are constructed analogously to those in Figure 9.

4.3 Effect Persistence

In this section I present estimation results for ninth-grade outcomes – the grade level after which students can leave school with a low-track degree (conditional on obtaining the required grades). Note that interpretation of these results is complicated by the fact that, from seventh grade onwards, non-academic-track school in the Comprehensive states may sort students by ability but there is no reliable information on the incidence and exact implementation of this within-school streaming. More generally, the harmonisation of schooling policies between states decreases with grade level (Kultusministerkonferenz, 2014). Accordingly, estimates represent a mixture of effect persistence and effects from continued (but somewhat unclear) differences in tracking and other schooling inputs. With these caveats in mind, the purpose of this section is two-fold: to obtain a rough idea of effect persistence and to see whether the patterns found until now replicate in the IQB data.

First, I repeat the above analysis for ninth-grade test scores in the NEPS data. While the DD estimates in Appendix Table B5 continue to show an advantage for students taught comprehensively in grades 5 and 6, they are smaller than before and far from significant as smaller samples and increased interference from other between-state differences seem to take their toll on precision.[45] As average effects might mask persistence for low-achieving students, Figure 10

[45]The differences between grade 7 and grade 9 results are not driven by sample differences. Using the smaller grade 9 sample for the grade 7 regressions reproduces the previous results quite precisely.

Table 3. *DD regressions for ninth-grade achievement in the IQB DD sample.*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Reading							
Comprehensive schooling	0.136	0.155	0.130	0.175	0.066	0.108	0.026
	(p = 0.18)	(0.12)	(0.30)	(0.19)	(0.37)	(0.35)	(0.16)
× Female				-0.036			
				(0.63)			
× Low SES					0.143*		
					(0.10)		
× Migration background						-0.010	
						(0.68)	
Compr. state indicator	-0.018						
	(0.90)						
Classroom peers' mean score							0.844***
							(0.00)
R^2	0.001	0.021	0.164	0.036	0.069	0.059	0.207
Panel B: Listening							
Comprehensive schooling	0.146*	0.155**	0.152**	0.134	0.088	0.111*	0.019
	(0.06)	(0.04)	(0.04)	(0.15)	(0.15)	(0.09)	(0.18)
× Female				0.045			
				(0.53)			
× Low SES					0.101		
					(0.37)		
× Migration background						0.010	
						(0.71)	
Compr. state indicator	-0.088						
	(0.32)						
Classroom peers' mean score							0.887***
							(0.00)
R^2	0.001	0.017	0.199	0.023	0.078	0.098	0.260
(Interacted) controls			✓				
State FE		✓	✓	✓	✓	✓	✓
N state clusters	12	12	12	12	12	12	12
N Compr. state students	13526	13526	13526	13526	13526	11854	13526
N Tracked state students	11276	11276	11276	11276	11276	10186	11276

Notes: Table 7 reports OLS regression results for the DD model for grade 4 and 9 reading (Panel A) and listening (Panel B) test scores of non-academic-track students using the IQB DD sample. Column 1 report results for the unsaturated DD model. Column 2 reports results for the saturated DD model, which replaces the Comprehensive state indicator with state fixed effects. Column 3 adds the following student and school covariates, incl. their interaction with the grade 9 indicator: sex, age, age squared, migration background, foreign language at home, highest parental level of education (HISCED), highest parental occupational status (HISEI), teacher experience, teacher further training, no. days all-day schedule/week, private school, homework support and extracurricular learning offers. Columns 4–6 fully interact the saturated DD model without controls with indicators for female, below-median socio-economic status (SES) and migration background students, respectively. The SES score is the first principal component of the following variables and their missing dummies: student-reported number of books at home, highest parental level of education, highest parental occupational status. Column 7 presents results for the saturated DD model controlling for classroom peers' mean test scores in the respective subject. p-values in parentheses stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Estimations apply student sampling weights and results are pooled across the 15 plausible values per test score domain for each student (see footnote 47 for details). Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

repeats the heterogeneity analysis from above for grade-5-to-9 gain scores using the NEPS 5-to-9 panel sample. Indeed, in maths effects for the lowest quartile of students are both economically and statistically significant, ranging from 0.15 to 0.3 SD, thus indicating persistent benefits from deferring between-school tracking for low-achievers. In reading, point estimates are positive but insignificantly different from zero across the previous achievement distribution.

Second, I re-estimate the DD model using the larger IQB DD sample.[46] Note that I can only

[46]Summary statistics for the IQB DD sample are presented in Appendix Table B6.

report results for reading and listening, as maths was not tested in ninth grade. Table 4 presents the estimation results.[47] Column 1 presents the unsaturated DD model, which estimates the raw double difference between Comprehensive and Tracked states between grades 4 and 9. In line with the NEPS data, I find no significant achievement differences at the end of primary school but an advantage for non-academic-track students from the Comprehensive states in secondary school of 0.14 SD in reading and 0.15 SD in listening. This result is robust to including state fixed effects (column 2) and flexibly controlling for student and school characteristics (column 3), but only statistically significant for listening. In columns 4–6 I fully interact the DD model with indicators for female, low-SES and migration background students to test for effect heterogeneities along observable student characteristics in the larger IQB samples. Of those three, only the SES interaction seems to substantially reduce the main effect, reaching marginal significance in the case of reading. Given that low-SES students are more likely to be low achieving, this is at least qualitatively in line with the heterogeneity by previous achievement found above, which I cannot directly investigate without panel data. Instead, I repeat the distributional analysis, which confirms that comprehensive schooling shifts probability mass from the bottom to the middle of the test score distribution in both reading and listening (see Appendix Figures B2 and B3). In sum, the IQB results confirm those based on the NEPS data.

[47]The IQB results pool across 15 so-called 'plausible values' (PVs): Students answer different subsets of the total pool of IQB assessment questions ('multi-matrix design'). In order to deal with the missing information on questions outside their subset, each student is imputed 15 PVs per test score domain. Standard practice is to run regressions for each and combine point estimates and standard errors according to the rules in Rubin (1987). These state that the variance of a statistic based on m imputations is the sum of the average within-imputation variance and the between-imputation variance: $Var^{total} = m^{-1} \sum_m Var_m^{within} + (1+m^{-1}) Var^{between}$. Problematically, the wild cluster bootstrap does not produce within-variance estimates (i.e. standard errors), but only a distribution of t-statistics from which p-values are computed. Instead of reverting to standard clustered standard errors (which are likely to underestimate the within-imputation variance due to the few-cluster problem) to use Rubin's rule, I decided to ignore the between-imputation variance and simply pool the wild cluster bootstrapped p-values across imputations (pooling means to convert them into t-values, average those across imputations and convert back into a p-value). Differences between PVs are so small that ignoring the between-imputation variance is innocuous in this context. A back-of-the-envelope calculation supports this choice: Appendix Table B7 shows the effect estimates in the saturated DD model for each PV/imputation, from which I calculate the between-imputation variance. For each imputation, I then (under-)estimate the within-imputation variance using a standard cluster-robust variance estimator. The between-variance is only 2.8% of the *underestimated* within-variance for reading and 7.4% for listening. So, applying Rubin's formula, in the case of listening, one would need to scale the within-variance by 1.08 to get the total variance, which would reduce the t-statistic by a factor of $1/\sqrt{1.08} = 0.96$. For reading this factor is even closer to one. Hence, even in this overly conservative calculation (due to underestimated within-variances), ignoring the between-variance is close to irrelevant for the coefficients' significance. I abstain from further assumptions to pool the bootstrapped confidence sets and only report averaged coefficients and pooled p-values.

4.4 Mechanisms

A large part of the preceding analysis has been devoted to understanding whether the estimated effects are indeed due to between-state differences in tracking. However, even if I can rule out confounding from school resources and student body composition, it is unclear what are the precise mechanisms underlying my results. The effect of between-school tracking on student achievement might operate through various channels.

First, and most problematically for my purposes, the effects might be driven by logistical implications of running a two- *versus* a three-tiered school system: in the former states need to maintain two distinct school types and in the latter three. This may impact local school supply, i.e. the size of schools and students' travelling time to school (and thereby time left for homework and other educational investments). For the sake of generalisability, I would like to rule out these channels and isolate the portion of the effect that is solely due to the sorting of students by ability between schools. Therefore, school size has been included in the list of school controls in the main regressions. The results in Table 2 and Appendix Table B2 suggest that it can be ruled out as a relevant channel. Due to lack of data I cannot directly control for students' travelling times. However, in fifth and seventh grade the NEPS questionnaires asked students to report their weekly time spent on homework, allowing me to investigate students' educational time investments directly. Column 1 of Table 5 presents results for the DD model applied to time spent on homework and shows that non-academic-track students from the Comprehensive state spent *less* time on homework and that this difference is constant across grades. As this is the group experiencing higher achievement gains between grades, time constraints are unlikely to play an important role in this context.

Second, the results might be driven by incentives to exert effort and invest for students and their parents. In contrast to the Tracked states, in the Comprehensive states students are not 'locked' into (low and intermediate) tracks in the first two years of secondary school. This might give students the impression that they need to work hard continually to reach their aspired degree; especially since some non-academic-track schools sort students into low- and intermediate-track classes starting in grade 7. The same applies to parents, who might thus be incentivised to invest more in their children's education during these two years. However, these conjectures do not seem to square with the evidence presented in Table 5: as mentioned before, column 1 shows that students in the

Comprehensive states spend less time on homework and, on top of that, columns 2 and 3 show that they are neither more likely to receive help from their parents with with school work, nor to receive private tutoring.[48]

Third, the effects might operate through the taught curriculum: for students who would be assigned the low track in the three-tiered system, comprehensive schooling is likely to increase academic standards, whereas for students who would be assigned the intermediate-track standards might decrease. *A priori* it is unclear how this might affect achievement, as low-achievers could lose out from being held to excessive academic standards or benefit if they grow with the demands. Regardless, given that curricular differences between low- and intermediate-track schools are relatively small in the first two grades of secondary school (Bald, 2011), this is unlikely to be the primary driver behind the results. Of course, the ability composition of the class might influence in what detail the teacher treats the material, but such *peer effects* should not be confused with curricular effects.

Fourth, by attending either a low- or a intermediate-track school, students are labelled and explicitly ranked in the Tracked states, whereas in the Comprehensive states they are not (save for being below the academic track). This social comparison might negatively affect their academic self-concept, educational aspirations, motivation to study and, in turn, achievement (Dumont *et al.*, 2017). Contrary to this, Murphy and Weinhardt (2020) show that classroom rank, which is likely to be higher in lower tracks, positively affects student achievement. However, their study concerns the non-tracked English school system where schools are not explicitly ranked. In the German tracking system, between-school sorting is salient and students are well aware of their track's rank, reducing the significance of favourable within-class comparisons in low-track schools. In line with this conjecture, Dumont *et al.* (2017) find that school-leaving certificates, which correspond to (but are not determined by) tracks, are the primary determinant of students' academic self-concept in German non-academic-track schools.

To investigate such socio-emotional channels the remaining columns of Table 5 present results for students' educational aspirations, their school-related motivation and feelings of helplessness in school. Only aspirations were measured in grades 5 and 7, allowing for implementation of the DD

[48]The IQB surveys included questions about private tutoring, too. In Appendix Table B8 I show that this result reproduces in the IQB data: if anything, there is more private tutoring in the Tracked states.

Table 4. *Effects on seventh-grade behavioural and socio-emotional outcomes.*

Model specification:	Double Differences					Cross-sectional OLS			
Dependent variable:				Aspirations		Helplessness		Motivation	
	Time spent on homework	Help from parents	Private tutoring	> low cert.	> mid cert.	Maths	German	Maths	German
Variable type:	*(scale)*	*(scale)*	*(dummy)*	*(dummy)*	*(dummy)*	*(scale)*	*(scale)*	*(scale)*	*(scale)*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Comprehensive schooling	0.054	0.048	-0.048	0.120*	0.003	-0.165*	-0.127	0.140	0.178
	(p = 0.49)	(0.63)	(0.61)	(0.07)	(0.97)	(0.08)	(0.17)	(0.41)	(0.33)
	[-0.10, 0.35]	[-0.14, 0.34]	[-0.15, 0.10]	[-0.03, 0.30]	[-0.16, 0.13]	[-0.35, 0.02]	[-0.25, 0.06]	[-0.13, 0.31]	[-0.19, 0.39]
Grade 7	-0.019	-0.203	-0.081	-0.076	0.072				
	(0.91)	(0.31)	(0.34)	(0.60)	(0.44)				
	[-0.37, 0.47]	[-0.60, 0.19]	[-0.30, 0.09]	[-0.31, 0.17]	[-0.19, 0.26]				
Indicator Compr. states	-0.242**	-0.018	-0.023	-0.041	0.095*				
	(0.03)	(0.78)	(0.28)	(0.35)	(0.08)				
	[-0.45, -0.07]	[-0.24, 0.17]	[-0.09, 0.04]	[-0.17, 0.06]	[-0.01, 0.26]				
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.036	0.047	0.036	0.080	0.094	0.031	0.050	0.067	0.061
N state clusters	12	12	12	12	12	12	12	12	12
N Compr. state students	834	819	471	799	799	503	499	486	487
N Tracked state students	3940	3870	2407	3869	3869	2010	2025	1850	1832

Notes: Columns 1–5 present OLS regression results for the unsaturated DD model with student and school covariates applied to different dependent variables in the NEPS DD sample of non-academic-track students, each time retaining all observations with non-missing values for the respective variable. 'Time spent on homework' is the student-reported average time spent on homework per week, standardised to mean zero and standard deviation one for Tracked states students within each grade level. 'Help from parents' is the student-reported frequency of help received with homework from their parents, standardised as before. 'Private tutoring' is a indicator variable equal to one if parents report that their child receives private tutoring. Educational aspirations are measured at two margins: in column 4 (column 5) the dependent variable is an indicator equal to one when students report aspiring higher than the low-track (intermediate-track) certificate. Columns 6–9 are based on the grade 7 cross-section only, as these outcomes were not measured in fifth grade. The dependent variables are regressed on the full set of controls and an indicator for the Comprehensive states. Helplessness is an index, standardised as before, with higher values indicating a higher degree of feeling helpless in the respective school subject. The variable averages 5 survey items, each measured on a 4-point Likert scale. Motivation is an index, standardised as before, with higher values indicating a higher intrinsic motivation for the respective school subject. The variable averages 4 survey items, each measured on a 4-point Likert scale. p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

design. I construct two indicator variables indicating that students aspire higher than the low- or intermediate-track degree, respectively. The results in columns 4 and 5 show that comprehensive schooling reduces the share of students with low educational aspirations, while there is no effect at the higher margin. Increased aspirations at the bottom seem to mirror the large benefits for low-achieving students found above. The remaining variables were only measured in grade 7, so that I am forced to revert to OLS regressions with large control sets to approximate the effect of comprehensive schooling. With this caveat in mind and despite their limited significance, the results in columns 6–9 suggest that students taught comprehensively are less helpless and more motivated. Appendix Table B8 shows very similar patterns in the IQB data; in particular, I find strong evidence for positive effects of comprehensive schooling on motivational outcomes. Although far from conclusive, these results suggest that socio-emotional effects of tracking are relevant.

The fifth and most palpable mechanism for the effects of any tracking policy is certainly peer effects – mind you that the stated goal of tracking is to homogenise classroom peers in terms of ability. According to tracking proponents, this should benefit all students by allowing for more tailored teaching (Duflo *et al.*, 2011). Tracking opponents argue that, instead of homogeneity, peers' ability level is what really matters: more able peers generate direct knowledge spill-overs, increase the quality of classroom interactions (including with teachers) and serve as role models. Numerous papers show positive effects of mean peer achievement on student achievement (e.g. Sacerdote, 2001; Whitmore, 2005; Ammermüller and Pischke, 2009; Carrell *et al.*, 2009; Lavy *et al.*, 2012; Burke and Sass, 2013; Garlick, 2018). A growing literature shows non-linearities in the effects of peers – in particular, very low-achieving peers seem to generate negative spill-overs (e.g. Figlio, 2007; Carrell *et al.*, 2018; Lavy *et al.*, 2012; Bietenbeck, 2020). In a similar vein, Bursztyn and Jensen (2015) present experimental evidence for the presence of peer pressures penalising effort in low-track (non-honours) classes that are absent in high-track (honours) classes. The costs of exposure to low-achieving peer environments might thus well be larger than the benefits of exposure to high-achieving peer environments.

Figure 4 showed that peer group composition differs significantly between the two tracking regimes: low-achievers attend lower achieving classrooms on average and, consequently, have a higher probability of being exposed to the lowest-achieving individuals in the Tracked states.

Moreover, anecdotes about low-track schools with negative peer dynamics that discourage learning are common. Peer effects are thus a likely candidate to explain the large gains from comprehensive schooling for lower achieving students. To test their role directly I perform a simple mediation analysis. Column 7 of Table 4 presents results from the DD model in the IQB data controlling for the mean of classroom peers' test scores.[49] In both reading and listening the effect of comprehensive schooling disappears when controlling for mean peer achievement, indicating the importance of peer effects. The found pattern of effect heterogeneity by previous achievement implies that peer effects are heterogeneous: low-achievers seem to be more sensitive to peer group composition than high-achievers – a result also found by Garlick (2018). Altogether, my results confirm the importance of peer effects in explaining the effects of tracking and, inversely, suggest that the homogeneity of classrooms at such an early age might be less important than commonly assumed.

5 Discussion and Conclusions

This paper set out to estimate the effect of early between-school tracking in secondary school on student achievement – an issue that, despite its enduring prevalence in educational policy debates, is still not fully understood. Theoretically, the question of tracked *versus* comprehensive schooling seems to involve a trade-off between countervailing forces. On the one hand, homogeneous learning environments are likely to facilitate skill and knowledge acquisition as content and teaching style can be more closely tailored to median classroom ability. On the other hand, the concentration of high ability students in certain schools might impair competence development of students in lower tracks through negative motivational consequences and peer effects. Identifying these effects is notoriously difficult due to the severity of the selection problems involved.

My identification strategy exploits differences in tracking between German federal states: in all states, about 40% of students transition to the academic track after comprehensive primary school. Depending on the state, the remaining student body is either divided between low- and intermediate-track schools or taught comprehensively for another two years. I estimate the effects of

[49] I am prevented from repeating the mediation analysis in the NEPS data because I do not observe students primary school (pre-tracking) classrooms. Note that the IQB data is better suited for the analysis of peer effects anyway, first, because of the large number of observed classes and, second, because participation in the IQB tests is mandatory such that *whole classes* are observed. Accordingly, mean peer achievement is measured much more accurately in the IQB data than in the NEPS.

these two years of comprehensive instead of tracked schooling on achievement in a triple-differences framework. The estimator compares achievement growth of comprehensively taught non-academic-track students with that of tracked ones, while controlling for tracking-regime-specific trends using unaffected students in the academic track.

I find that student achievement increases when non-academic-track students are not ability tracked between schools but taught comprehensively for another two years: the 95% confidence set for the effect on seventh-grade test scores is estimated to be [0.09, 0.28] SD in maths and [0.04, 0.48] SD in reading. These somewhat imprecisely estimated level differences are composed of large positive effects for low-achievers and null effects for high-achievers. Consequently, comprehensive schooling has an equalising effect on the distribution of test scores without trading off efficiency against equity. There is some fade-out in the effects but comprehensively taught students are still better off towards the end of lower secondary schooling. Auxiliary analyses suggest that students' school-related motivation and educational aspirations are higher in the comprehensive system and that peer effects play an important role in explaining the effects.

With respect to Germany, my results confirm the reform efforts of several West German states to abolish low-track schools and replace their three- with two-tiered school systems. In line with policy-makers intentions, this appears to generate better and more equitable outcomes. Beyond the German context, the effects in this paper are immediately relevant for other countries with multi-track between-school tracking systems, like e.g. Czech Republic, Netherlands and Slovakia.

With respect to countries with two-tiered tracking systems, caution must be exercised when extrapolating from my results to the effects of turning those into fully comprehensive school systems. This is because the variation in tracking practices I exploit concerns only the (negatively) selected group of non-academic-track students. Accordingly, my results might not translate to students in the academic track. However, note that the central dimension of effect heterogeneity is previous achievement, which overlaps considerably between tracks. Even for the top quartile of non-academic-track students, who would be medium-high achievers also in the academic track, I find no evidence for negative effects from comprehensive schooling (with *positive* point estimates). This suggests that, if there are negative effects at the young ages considered here, these are confined to the very top of students.

Overall, my results provide a cautionary tale about early and rigid forms of vertical differentiation in schools applicable to all between-school tracking settings. They show that there are limits to efficiency gains from classroom homogeneity as other mechanisms, such as peer effects, motivation and aspirations, start to depress achievement at the bottom once a selective system becomes too differentiated. Accordingly, policy-makers need to carefully balance these forces when determining the degree of vertical differentiation in their school systems and the age at which it starts.

Finally, note that many papers on *within-school* streaming report positive effects for students selected for high-ability classrooms without negative effects for those in regular classrooms (e.g. Card and Giuliano, 2016; Duflo *et al.*, 2011; Figlio and Page, 2002). Rather than contradicting my and previous findings on between-school tracking, this suggests that costs of ability grouping increase with the degree of vertical differentiation between tracks. It makes intuitive sense that mechanisms relating to peer effects, motivational factors and educational aspirations are more pronounced when students are separated between schools. Consequently, forming (subject-specific) *classrooms* based on ability from a certain age onwards, but eschewing vertical differentiation between *schools* to avoid creating detrimental learning environments for low-track students, might allow reaping efficiency gains from homogeneity without incurring large costs in terms of equity.

References

Aakvik, A., Salvanes, K.G. and Vaage, K. (2010). 'Measuring heterogeneity in the returns to education using an education reform', *European Economic Review*, vol. 54(4), pp. 483–500.

Abadie, A., Athey, S., Imbens, G.W. and Wooldridge, J. (2017). 'When should you adjust standard errors for clustering?', NBER Working Paper No. 24003.

Abdulkadiroğlu, A., Angrist, J. and Pathak, P. (2014). 'The elite illusion: Achievement effects at Boston and New York exam schools', *Econometrica*, vol. 82(1), pp. 137–196.

Ammermüller, A. (2013). 'Institutional features of schooling systems and educational inequality: Cross-country evidence from PIRLS and PISA', *German Economic Review*, vol. 14(2), pp. 190–213.

Ammermüller, A. and Pischke, J.S. (2009). 'Peer effects in european primary schools: Evidence from the Progress in International Reading Literacy Study', *Journal of Labor Economics*, vol. 27(3), pp. 315–348.

Angrist, J.D. and Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.

Bald, H. (2011). 'Realschule–Erweiterte Realschule–Mittelschule usw.–eine Problemanzeige', *Theo-Web. Zeitschrift für Religionspädagogik*, vol. 10, pp. 80–102.

Bellenberg, G. (2005). 'Wege durch die Schule - Zum Zusammenhang zwischen institutionalisierten Bildungswegen und individuellen Bildungsverläufen im deutschen Schulsystem', *Bildungsforschung*, vol. 2(2).

Bellenberg, G. (2012). *Schulformwechsel in Deutschland. Durchlässigkeit und Selektion in den 16 Schulsystemen der Bundesländer innerhalb der Sekundarstufe I*, Gütersloh: Bertelsmann-Stiftung.

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

Betts, J.R. (2011). 'The economics of tracking in education', pp. 341–381, vol. 3 of *Handbook of the Economics of Education*, Amsterdam: Elsevier.

Bietenbeck, J. (2020). 'The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR', *Journal of the European Economic Association*, vol. 18(1), pp. 392–426.

Blossfeld, H.P., Rossbach, H.G. and von Maurice, J. (2011). 'Education as a lifelong process: The German National Educational Panel Study (NEPS)', *Zeitschrift für Erziehungswissenschaft*, vol. 14 (special issue).

Brunello, G. and Checchi, D. (2007). 'Does school tracking affect equality of opportunity? New international evidence', *Economic Policy*, vol. 22(52), pp. 782–861.

Brunello, G., Giannini, M. and Ariga, K. (2007). 'The optimal timing of school tracking: A general model with calibration for Germany', in (L. Wössmann and P. Peterson, eds.), *Schools and the Equal Opportunity Problem*, pp. 129–156, Cambridge: MIT Press.

Burke, M.A. and Sass, T.R. (2013). 'Classroom peer effects and student achievement', *Journal of Labor Economics*, vol. 31(1), pp. 51–82.

Bursztyn, L. and Jensen, R. (2015). 'How does peer pressure affect educational investments?', *The Quarterly Journal of Economics*, vol. 130(3), pp. 1329–1367.

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

Card, D. and Giuliano, L. (2016). 'Can tracking raise the test scores of high-ability minority students?', *American Economic Review*, vol. 106(10), pp. 2783–2816.

Carrell, S.E., Fullerton, R.L. and West, J.E. (2009). 'Does your cohort matter? Measuring peer effects in college achievement', *Journal of Labor Economics*, vol. 27(3), pp. 439–464.

Carrell, S.E., Hoekstra, M. and Kuka, E. (2018). 'The long-run effects of disruptive peers', *American Economic Review*, vol. 108(11), pp. 3377–3415.

Contini, D. and Cugnata, F. (2016). 'Learning inequalities between primary and secondary school. Difference-in-difference with international assessments', University of Turin 'Cognetti de Martiis' Working Paper No. 07/16.

Duflo, E., Dupas, P. and Kremer, M. (2011). 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya', *American Economic Review*, vol. 101(5), pp. 1739–74.

Dumont, H., Protsch, P., Jansen, M. and Becker, M. (2017). 'Fish swimming into the ocean: How tracking relates to students' self-beliefs and school disengagement at the end of schooling', *Journal of Educational Psychology*, vol. 109(6), pp. 855–870.

Dustmann, C. (2004). 'Parental background, secondary school track choice, and wages', *Oxford Economic Papers*, vol. 56(2), pp. 209–230.

Dustmann, C., Puhani, P. and Schoenberg, U. (2017). 'The long-term effects of early track choice', *The Economic Journal*, vol. 127(603), pp. 1348–1380.

Edelstein, B. and Nikolai, R. (2013). 'Strukturwandel im Sekundarbereich. Determinanten schulpolitischer Reformprozesse in Sachsen und Hamburg', *Zeitschrift fuer Pädagogik*, vol. 59(4), pp. 482–494.

Eisenkopf, G. (2007). 'Tracking and incentives', Thurgau Institute of Economics Research Paper No. 22.

European Commission (2014). 'European Vacancy and Recruitment Report', *Directorate-General for Employment, Social Affairs and Inclusion*.

Figlio, D.N. (2007). 'Boys named Sue: Disruptive children and their peers', *Education Finance and Policy*, vol. 2(4), pp. 376–394.

Figlio, D.N. and Page, M.E. (2002). 'School choice and the distributional effects of ability tracking: Does separation increase inequality?', *Journal of Urban Economics*, vol. 51(3), pp. 497–514.

Fischer, L., Rohm, T., Gnambs, R. and Carstensen, C. (2016). 'Linking the data of the competence tests', NEPS Survey Paper No. 1.

Galindo-Rueda, F. and Vignoles, A.F. (2004). 'The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability', IZA Discussion Paper No. 1245.

Gamoran, A. and Mare, R.D. (1989). 'Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality?', *American Journal of Sociology*, vol. 94(5), pp. 1146–1183.

Garlick, R. (2018). 'Academic peer effects with different group assignment policies: Residential tracking versus random assignment', *American Economic Journal: Applied Economics*, vol. 10(3), pp. 345–369.

Guyon, N., Maurin, E. and McNally, S. (2012). 'The effect of tracking students by ability into different schools a natural experiment', *Journal of Human Resources*, vol. 47(3), pp. 684–721.

Hall, P. and Kang, K.H. (2001). 'Bootstrapping nonparametric density estimators with empirically chosen bandwidths', *Annals of Statistics*, pp. 1443–1468.

Hanushek, E.A. and Wössmann, L. (2006). 'Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries', *The Economic Journal*, vol. 116(510), pp. C63–C76.

Helbig, M. and Nikolai, R. (2015). *Die Unvergleichbaren: Der Wandel Der Schulsysteme in Den Deutschen Bundesländern Seit 1949*, Bad Heilbrunn: Julius Klinkhardt.

A. C. Kerckhoff, K. Fogelman, D. Crook and D. Reeder, eds. (1996). *Going Comprehensive in England and Wales: A Study of Uneven Change*, London: Routledge.

Kerr, S.P., Pekkarinen, T. and Uusitalo, R. (2013). 'School tracking and development of cognitive skills', *Journal of Labor Economics*, vol. 31(3), pp. 577–602.

Kultusministerkonferenz (2014). *The Education System in the Federal Republic of Germany 2012/2013*, Bonn: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Laender in the Federal Republic of Germany.

Lavy, V., Silva, O. and Weinhardt, F. (2012). 'The good, the bad, and the average: Evidence on ability peer effects in schools', *Journal of Labor Economics*, vol. 30(2), pp. 367–414.

Leschinsky, A. (2008). 'Die Realschule - Ein zweischneidiger Erfolg', in (K. S. Cortina, J. Baumert, A. Leschinsky, K. U. Mayer and L. Trommer, eds.), *Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick*, Reinbek bei Hamburg: Rowohlt.

Mackinnon, J.G. and Webb, M.D. (2017). 'Wild bootstrap inference for wildly different cluster sizes', *Journal of Applied Econometrics*, vol. 32(2), pp. 233–254.

MacKinnon, J.G. and Webb, M.D. (2018). 'The wild bootstrap for few (treated) clusters', *The Econometrics Journal*, vol. 21(2), pp. 114–135.

Matthewes, S.H. (2018). 'Better together? Heterogeneous effects of tracking on student achievement', DIW Berlin Discussion Paper No. 1775.

Meghir, C. and Palme, M. (2005). 'Educational reform, ability, and family background', *American Economic Review*, vol. 95(1), pp. 414–424.

Murphy, R. and Weinhardt, F. (2020). 'Top of the Class: The Importance of Ordinal Rank', *The Review of Economic Studies, forthcoming*.

Neumark, D., Schweitzer, M. and Wascher, W. (2004). 'Minimum wage effects throughout the wage distribution', *Journal of Human Resources*, vol. 39(2), pp. 425–450.

Oakes, J. (1985). *Keeping Track: How Schools Structure Inequality*, Yale University Press.

Pant, H.A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. and Pöhlmann, C. (2013). *IQB-Ländervergleich 2012 - Zusatzmaterialien*, Münster: Waxmann.

Paulus, W. and Blossfeld, H.P. (2007). 'Schichtspezifische Präferenzen oder sozioökonomisches Entscheidungskalkül? Zur Rolle elterlicher Bildungsaspirationen im Entscheidungsprozess beim übergang von der Grundschule in die Sekundarstufe', *Zeitschrift für Pädagogik*, vol. 53(4), pp. 491–508.

Piopiunik, M. (2014). 'The effects of early tracking on student performance: Evidence from a school reform in Bavaria', *Economics of Education Review*, vol. 42, pp. 12–33.

Pischke, J.S. and Manning, A. (2006). 'Comprehensive versus selective schooling in England and Wales: What do we know?', NBER Working Paper No. 12176.

Roodman, D., Nielsen, M.Ø., MacKinnon, J.G. and Webb, M.D. (2019). 'Fast and wild: Bootstrap inference in Stata using boottest', *The Stata Journal*, vol. 19(1), pp. 4–60.

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

Sacerdote, B. (2001). 'Peer effects with random assignment: Results for Dartmouth roommates', *The Quarterly Journal of Economics*, vol. 116(2), pp. 681–704.

Sacerdote, B. (2011). 'Peer effects in education: How might they work, how big are they and how much do we know thus far?', pp. 249–277, vol. 3 of *Handbook of the Economics of Education*, Amsterdam: Elsevier.

Schütz, G., Ursprung, H.W. and Wössmann, L. (2008). 'Education policy and equality of opportunity', *Kyklos*, vol. 61(2), pp. 279–308.

Schwerdt, G. and Ruhose, J. (2016). 'Does early educational tracking increase migrant-native achievement gaps? Differences-in-differences evidence across countries', *Economics of Education Review*, vol. 52, pp. 134–154.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Slavin, R.E. (1990). 'Achievement effects of ability grouping in secondary schools: A best-evidence synthesis', *Review of Educational Research*, vol. 60(3), pp. 471–499.

P. Stanat, K. Böhme, S. Schipolowski and N. Haag, eds. (2016). *IQB-Bildungstrend 2015: Sprachliche Kompetenzen Am Ende Der 9. Jahrgangsstufe Im Zweiten Ländervergleich*, Münster: Waxmann.

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

Statistisches Bundesamt (2011). *Allgemeinbildende Schulen, Fachserie 11 Reihe 1, Schuljahr 2010/11*, Wiesbaden.

Statistisches Bundesamt (2012). *Allgemeinbildende Schulen, Fachserie 11 Reihe 1, Schuljahr 2011/12*, Wiesbaden.

Statistisches Bundesamt (2013). *Allgemeinbildende Schulen, Fachserie 11 Reihe 1, Schuljahr 2012/13*, Wiesbaden.

van Ewijk, R. (2011). 'Same work, lower grade? Student ethnicity and teachers' subjective assessments', *Economics of Education Review*, vol. 30(5), pp. 1045–1058.

Waldinger, F. (2007). 'Does tracking affect the importance of family background on students' test scores?', London School of Economics Working Paper.

Webb, M.D. (2014). 'Reworking wild bootstrap based inference for clustered errors', Queen's Economics Department Working Paper No. 1315.

Whitmore, D. (2005). 'Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment', *American Economic Review*, vol. 95(2), pp. 199–203.

Wössmann, L. (2016). 'The importance of school systems: Evidence from international differences in student achievement', *Journal of Economic Perspectives*, vol. 30(3), pp. 3–32.

A Data Appendix

The purpose of this appendix is to give a detailed description of the data sets and samples used throughout the paper. This discussion will be guided by Appendix Figure A1, which gives a schematic overview of how the different samples correspond to student cohorts and grade levels. The horizontal axis represents school years, divided into first and second term to show whether students were surveyed at the beginning or end of a school year. The vertical axis represents grade levels: the first four grades correspond to primary school, after which students transition to secondary school. Secondary school finishes after grade 9 in the low track, grade 10 in the intermediate track and grade 12 or 13, depending on the state, in the academic track. Cells containing survey names indicate the timing of testing/surveying. The shading shows the progression of sampled cohorts through grade levels within the German school system.

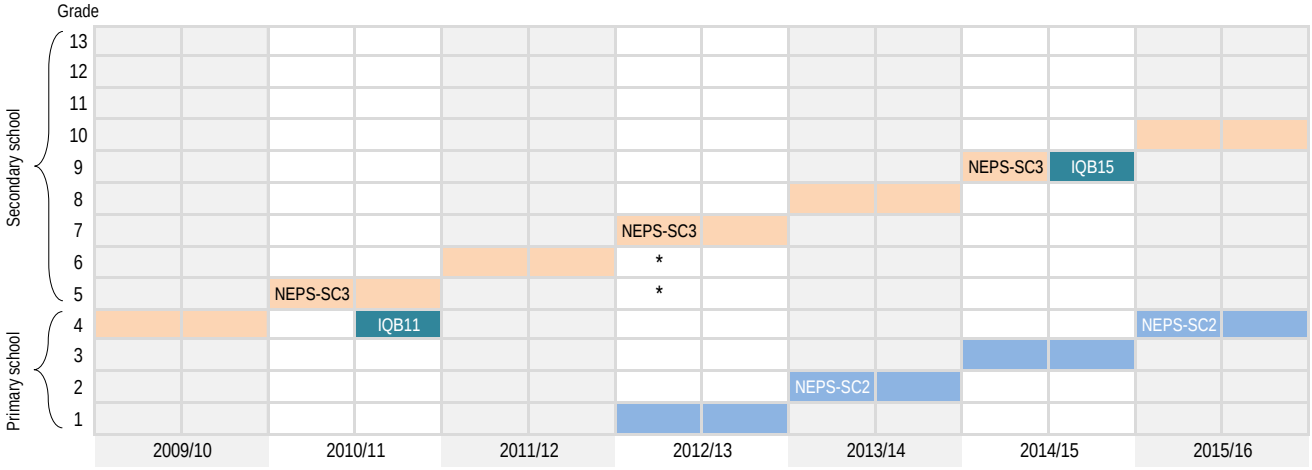

Figure A1. *Overview of data sets*

A.1 National Educational Panel Study

The German National Educational Panel Study (NEPS) is a study carried out by the Leibniz Institute for Educational Trajectories at the University of Bamberg. The NEPS collects longitudinal data on the development of competencies, educational processes, educational decisions, and returns to education for six different 'starting cohorts' (SC): newborns (SC1), kindergarten/primary school students (SC2), lower secondary school students (SC3), upper secondary school students (SC4), university students (SC5) and adults (SC6). For details on the project see Blossfeld *et al.* (2011).

A.1.1 NEPS-SC3

The main dataset of this paper is the lower secondary NEPS Starting Cohort 3 (NEPS-SC3), a random sample of newly minted fifth-graders in the school year 2010/11. Students were sampled according to a multi-stage process: (1) Random sampling (with probabilities proportional to scale) from the population of all German schools at lower secondary level. (2) Random selection of two grade 5 classes within the selected schools. (3) All students of the selected classes were invited to participate in the study. Participating students were surveyed and tested for the first time in the autumn of 2010, at the beginning of their first year in secondary school. Counting only students in regular schools, attending fifth grade for the first time in 2010/11, with non-missing test scores in mathematics and reading, the total size of the 'NEPS grade 5 cross-section' is 4,448, of which 2,303 are non-academic-track students, of whom 330 are from the Comprehensive states.

I use information from student and parent questionnaires to construct student-level control variables: age, a binary indicator for sex, a binary indicator for migration background, a binary indicator for single parent household, a binary indicator for foreign language spoken at home, an index for home possessions (from the student questionnaire), highest level of parental education measured in four categories, monthly household income, a binary indicator for receipt of unemployment benefits (from the parent questionnaire). While the student questionnaire variables are observed for almost everyone (< 1% missing values), parents' answers are missing for about 30% of the sample.

I use information from the school principal questionnaire to construct the following proxies for school quality/schooling inputs: average teacher age, student-teacher ratio, average class size, private school and four standardised indices for a school's facilities[50], extracurricular programmes[51], educational support offers[52] and quality control measures[53]. One additional school-level covariate

[50]The index sums the following binary items about the school's facilities: presence of gym; swimming pool; language laboratory; auditorium; common rooms; individual work stations; student library; teacher library.

[51]The index sums the following binary items about the schools afternoon programme: extracurricular homework supervision; remedial teaching for students with non-German background; instruction for students with non-German background; courses in maths; science; German or literature; foreign languages; sports; music or arts; politics or philosophy; handicrafts; offers in technology or media; community activities; social learning; inter-cultural learning; required free-time activities; voluntary free-time activities; project days; project weeks; hot lunches; long-term projects.

[52]The index sums the following binary items about the individual educational support offered by the school: courses in learning techniques; participation in projects or competitions; homework coaching; tutoring; other forms of coaching.

[53]The index sums the following binary items about quality control: complete school mission statement; written school profile; written specification of quality indicators; written specification of performance standards; standardised performance testing; systematic appraisal of data; school brochure; harmonised exams across classrooms.

is retrieved from the teacher questionnaire (but averaged by school to reduce the number of missing values): days of further training received over the past year. School covariates are missing for about 10% of the sample.

The NEPS being a panel, the same students were tested again two years later, in the autumn of 2012, when they had just started seventh grade according to schedule. Students that repeated a grade but remained in the same school are included in the testing, which in Figure A1 is indicated by two asterisks at grade levels 5 and 6 (note that to still be in fifth grade students would have to had repeated twice, a case that is not actually observed in the data). Students that switched school are not part of the grade 7 sample, as testing was tied to students remaining in their initial school context. All analyses that use grade-5-to-7 gain-scores as outcomes are based on the panel sample of students who have non-missing test scores in these first two waves of the NEPS-SC3 survey. This sample is referred to as the 'NEPS 5-to-7 panel sample' in the text.

There is substantial attrition in the NEPS-SC3 panel: of the 4,448 students tested in fifth grade 3,521 are tested again in 2012, of whom 1,646 are non-academic-track students, of whom 269 are from the Comprehensive states. This amounts to an overall panel attrition rate of 21%. In the non-academic tracks the attrition rate is 29%, compared to only 13% in the academic track, indicating that panel drop-out is negatively associated with achievement. Indeed, limiting attention to the non-academic-track sample, drop-outs have 0.06 SD lower maths scores ($p < 0.01$) and 0.04 SD lower reading scores ($p < 0.01$) than their peers (they are also 4 percentage points more likely to be migrants ($p = 0.06$) and 8 percentage points more likely to have low SES ($p < 0.01$)). Further, 49% of panel drop-outs in the Tracked states are from low-track schools, whereas only 35% of all non-academic-track students in the Tracked states belong to the low track.

The reasons for drop-out in the non-academic-track sample are schools withdrawing their participation in the NEPS study (36%), schools or classes being closed (8%) and students switching school (35%). The remaining 21% drop out for an unknown reason, i.e. either because of absence on the day of testing or because students or their parents withdrew their participation in the survey. Overall, attrition is higher in the Tracked states (30% compared to 18% in the Comprehensive states). However, excluding panel drop-out due to schools withdrawing their participation in the survey and schools closing (as these are due to administrative reasons at the school level, unlikely

to be related to schooling policy at the state level and clearly not driven by self-selection at the student level), there are no significant differences in any of these shares between Comprehensive and Tracked states.

In addition to students part of the panel sample, the 2012/13 'NEPS grade 7 cross-section' is augmented with a large random refreshment sample of seventh-graders. The refreshment sample was drawn to counteract selective attrition: of the 1,795 additional students, a large majority of 1,125 are from non-academic-track schools (of whom 283 are from the Comprehensive states). Accordingly, the refreshment sample balances the higher rate of attrition in the non-academic tracks and ensures that the NEPS sample remains representative of the student population in both segments of the school system. Together with the 3,521 students from the panel sample, the NEPS grade 7 cross-section has, in total, 5,316 observations, of which 2,771 are in the non-academic tracks, of which 552 are from the Comprehensive states.

The main difference-in-differences (DD) and triple-differences (DDD) regressions presented in Table 2 pool the NEPS-SC3 grade 5 and grade 7 cross-sections. The DD model uses non-academic-track students only and, thus, relies on 2,303 fifth-grade and 2,771 seventh-grade student observations (for 5074 student×grade observations in total). This is referred to as the (grade 7) 'NEPS DD sample' in the text. The DDD model adds academic-track students students as an additional control group for an additional 2,145 fifth-grade and 2,545 seventh-grade observations (for 9,764 student×grade observations in total). This is referred to as the (grade 7) 'NEPS DDD sample' in the text. These sample sizes are summarized in Appendix Table A1.

Table A1. *Sample sizes of NEPS cross-sections*

	Non-academic tracks		Academic track	
	Grade 5	Grade 7	Grade 5	Grade 7
Tracked states	1,973	2,219	1,797	2,064
Compr. states	330	552	348	481
	DD sample			
	DDD sample			

After the second wave, students were tested again two years later, in the school year 2014/15, when the cohort attended ninth grade according to schedule. Limiting attention to the non-academic tracks, of the initial panel sample there are 1,286 observations left in mathematics, of which 186 are from the Comprehensive states, and 1,255 in reading, of which 191 are from the Comprehensive

states (testing in reading happened later in the year, explaining the difference in the number of observations). All analyses that use grade-5-to-9 gain-scores as outcomes are based on these 'NEPS 5-to-9 panel samples'.

The 'NEPS grade 9 cross-section' of non-academic-track students, which on top of the grade-5-to-9 panel sample includes students from the grade 7 refreshment sample still participating in the survey, comprises 2,149 student observations, of which 433 are from the Comprehensive states. Analogously to the seventh-grade DD regressions, for the ninth-grade DD model of Table 3 the NEPS grade 5 and grade 9 cross-sections are pooled.

A.1.2 NEPS-SC2

The NEPS Starting Cohort 2 (NEPS-SC2), a random sample of German primary school students, is used as an additional data source for two reasons: (i) to provide information on primary school inputs and (ii) to investigate achievement trends before tracking starts. The sampling design is very similar to that of NEPS-SC3, with schools as primary sampling units.[54]

A concern for the validity of the DD estimates is that school inputs might have changed differently between primary and secondary school between Tracked and Comprehensive states. Accordingly, it is important to probe their robustness to the inclusion of school input controls. However, only the secondary school environment is observed in the NEPS-SC3, which logically can only affect the post-treatment (grade 7) scores. For the pre-treatment (grade 5) scores, the relevant schooling inputs are those from primary school, which are missing because because I do not observe the primary school students came from. To impute the missing primary school/pre-treatment schooling inputs in the NEPS-SC3 DD sample, I use the SC2 principal questionnaires. In particular, I calculate state-level averages for all the above-mentioned school-level controls, using the earliest available principle questionnaire for each school to minimise the distance between my main cohorts primary school time and the time the primary school information is recorded (in vast majority of cases this means 2012), and assign each grade 5 observation in the NEPS DD sample its state-level average.[55]

[54]Note that the NEPS-SC2 panel commenced two years prior to primary school, when children were still in kindergarten. Still, primary schools served as primary sampling units. As the earlier waves of the panel are irrelevant for the purpose of this paper, they are ignored in this description.

[55]Fortunately, primary school (SC2) principals were asked the same questions as those in secondary school (SC3).

For the analyses of primary school achievement trends, I use the NEPS-SC2's student-level achievement data. As shown in Appendix Figure A1, the first measurement point used here is the beginning of the school year 2013/14, when the surveyed cohort has just entered second grade. Counting only students in regular schools with with non-missing test scores in mathematics, the total size of the 'NEPS grade 2 cross-section' is 5,384, of which 979 are from the Comprehensive states.

Students were tested again two years later, in the autumn of 2015, when they had just started fourth grade according to schedule. Those that repeated a grade but remained in the same school were included in the testing. Students that switched school are not part of the grade 4 sample. Analyses that use grade-2-to-4 gain-scores as outcomes are based on the panel sample of students who have non-missing test scores in these first two waves of the NEPS-SC2. This sample is referred to as the 'NEPS 2-to-4 panel sample' in the text. It comprises 4,676 observations in total, of which 849 are from the Comprehensive states. Hence, the attrition rate between second and fourth grade is 13% overall and in both state groups. Again, panel drop-out is negatively related to performance: drop-outs have 0.18 SD lower maths scores ($p < 0.01$).

In addition to the panel sample the 'NEPS grade 4 cross-section' includes 1,141 newly sampled students[56] for a total sample size of 5,817, of which 1,059 are from the Comprehensive states. The DD model presented in Appendix Table B3 pools the NEPS-SC2 grade 2 and grade 4 cross-sections for a total sample size of 11,201 student×grade observations.

A.2 IQB National Assessment Studies

For auxiliary analysis the paper draws on two large German cross-sectional educational assessment studies carried out by the Institute for Educational Quality Improvement (IQB) at the behest of the Standing Conference of the Ministers of Education and Cultural Affairs of the States (KMK): the IQB National Assessment Study 2011 (*IQB Ländervergleich in der Primarstufe 2011*; IQB11) and the IQB National Assessment Study 2015 (*IQB-Bildungstrend 2015 in der Sekundarstufe I*; IQB15). The purpose of the IQB studies is to monitor in how far students meet nationally defined

[56]Unlike in the NEPS-SC3, the newly sampled students are not part of a refreshment sample. These are students that were part of the SC2 kindergarten sample but then attended a primary school that did not participate in the NEPS. For financial reasons, after kindergarten these students were only tested again in fourth grade.

educational standards for the primary and lower secondary level. Participation in the IQB tests is mandatory for all sampled students.[57] In contrast to the NEPS' sampling design, the IQB studies do not randomly sample from the population of all German students in a particular grade level but, instead, draw separate random samples within each state. Accordingly, smaller states are heavily overrepresented in the IQB data and the use of student sampling weights is necessary to obtain estimates representative of Germany.

A.2.1 IQB11

The IQB11 was the first primary-level National Assessment Study (see Stanat *et al.*, 2012, for details). It tested fourth-graders in mathematics, reading and listening at the end of the 2010/11 school year, when students were at the end of their primary school time. As can be seen in Appendix Figure A1, this is one cohort later than the NEPS-SC3 cohort. Within each state, sampling followed a multi-stage process: (1) Random sampling of primary schools. (2) Random selection of one fourth-grade class within selected schools. (3) All students in the selected class were obliged to participate.

Again retaining only students with non-missing test scores on regular schools, the total sample size of the 'IQB11 grade 4 cross-section' is 18,904, of which 11,187 are from the Comprehensive states. Note that the number of sampled schools in each state was chosen depending on earlier estimates of the variance in student performance to achieve similar level of precision in each state. Accordingly, the number of observations is different per state and not proportional to the actual share of a state's schools (or students) of the overall German population.

Most of the analysis restricts attention to non-academic-track students. Hence, classifying students as academic- or non-academic-track is crucial. This presents a challenge for using the IQB11 data as there is is no official assignment of students to tracks in primary school yet. Fortunately, however, the IQB11 survey was conducted at the very end of the school year: data collection ran from the end of May until mid July. Fourth-grade students receive their track recommendation with their mid-term reports in January and then start applying for secondary schools. The application period typically ends in March. Accordingly, at the time of the survey it

[57] However, participation in the accompanying student, parent and teacher questionnaires is not mandatory in some states, such that, control variables have more missing values.

had been decided which secondary school, and hence track, students would attend in the coming year already. The IQB11 survey asked parents directly which school their child will attend in the coming year. As, unfortunately, this variable has about 20% missing values due to parental non-response I also use information on students' track recommendation, which is reported by the school and has almost no missing values (1%), to classify students as academic or non-academic (see below). The resulting IQB11 grade 4 cross-section of non-academic-track students comprises 11,158 observations, of which 6,573 are from the Comprehensive states.

In order to classify students as accurately as possible based on the two above-mentioned variables, I choose state-specific assignment rules that maximise the fit between the state-specific academic-track shares estimated from my sample and the true shares, obtained from administrative records (Statistisches Bundesamt, 2012). In seven states where the track recommendation is non-binding,[58] I assign all students whose parents report that their child will attend an academic-track school in the coming year to the academic track. Among those students with a missing parent answer, I classify students with an academic track recommendation as academic. The rest is classified as non-academic. In the remaining five states where the track recommendation is binding, two (slightly) different assignment rules emerge as best predictors: In Bavaria and Baden-Württemberg, I only assign students to the academic track if they have both an academic-track recommendation and their parents report that they will attend an academic-track school in the coming year. All others are classified as non-academic. In Thuringia, Saxony and Saxony-Anhalt, I assign all students whose parents report that they will attend an academic-track school in the coming year to the academic track, unless they fail to have an academic-track recommendation (however, a missing value on this variable is fine). Of those with a missing parent answer, those with an academic-track recommendation are assigned to the academic track. The rest is classified as non-academic.

A.2.2 IQB15

The IQB15 tested ninth-graders in reading and listening at the end of the 2014/15 school year, towards the end of lower secondary schooling (see Stanat *et al.*, 2016, for details).[59] As can be seen

[58]These are Bremen, Hamburg, Hesse, Lower Saxony, North Rhine-Westphalia, Saarland and Schleswig-Holstein.
[59]Unfortunately, the IQB15 did not test students in maths.

in Appendix Figure A1, this is the same cohort as the NEPS-SC3 cohort used for the main analysis. Sampling and survey design is largely identical to the IQB11 survey. Within each state, sampling followed a multi-stage process: (1) Random sampling of secondary schools. (2) Random selection of one ninth-grade class within selected schools. (3) All students in the selected class were obliged to participate.

All analysis based on the IQB15 data restrict attention to students on regular *non-academic-track* schools with non-missing test scores. The total sample size of the non-academic-track 'IQB15 grade 9 cross-section' is 13,742 students, of whom 7,009 are from the Comprehensive states.

The DD regressions for reading and listening competencies presented in Table 4, as well as the DD regressions for non-cognitive outcomes presented in Table B8, pool the non-academic-track IQB11 grade 4 and IQB15 grade 9 cross-sections for the 'IQB DD sample' of 24,900 student×grade observations.

B Additional Tables and Figures

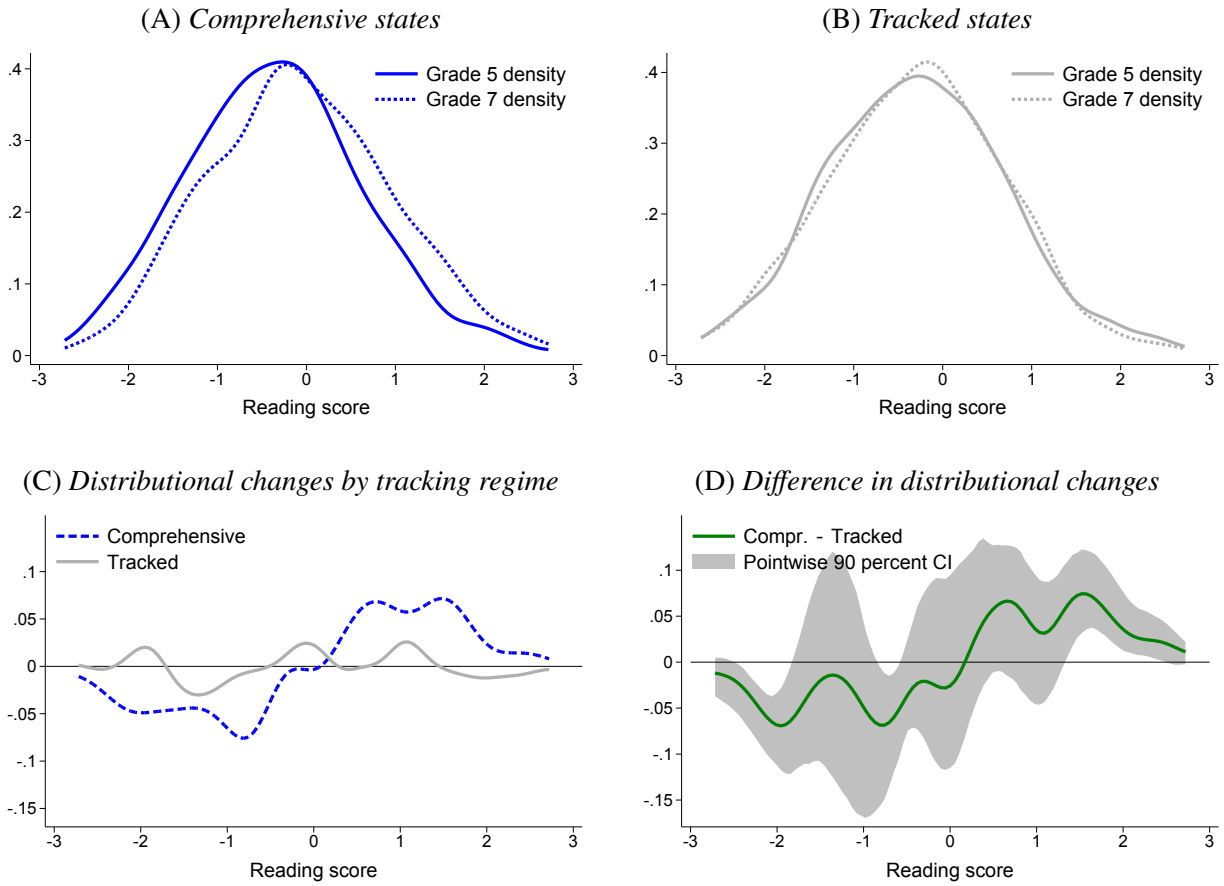

Figure B1. *Reading score distributions before and after treatment exposure.*

Notes: Figure B1 describes how the test score distribution in reading changes differently between grades 5 and 7 depending on the tracking regime. The same notes as in Figure 8 apply.

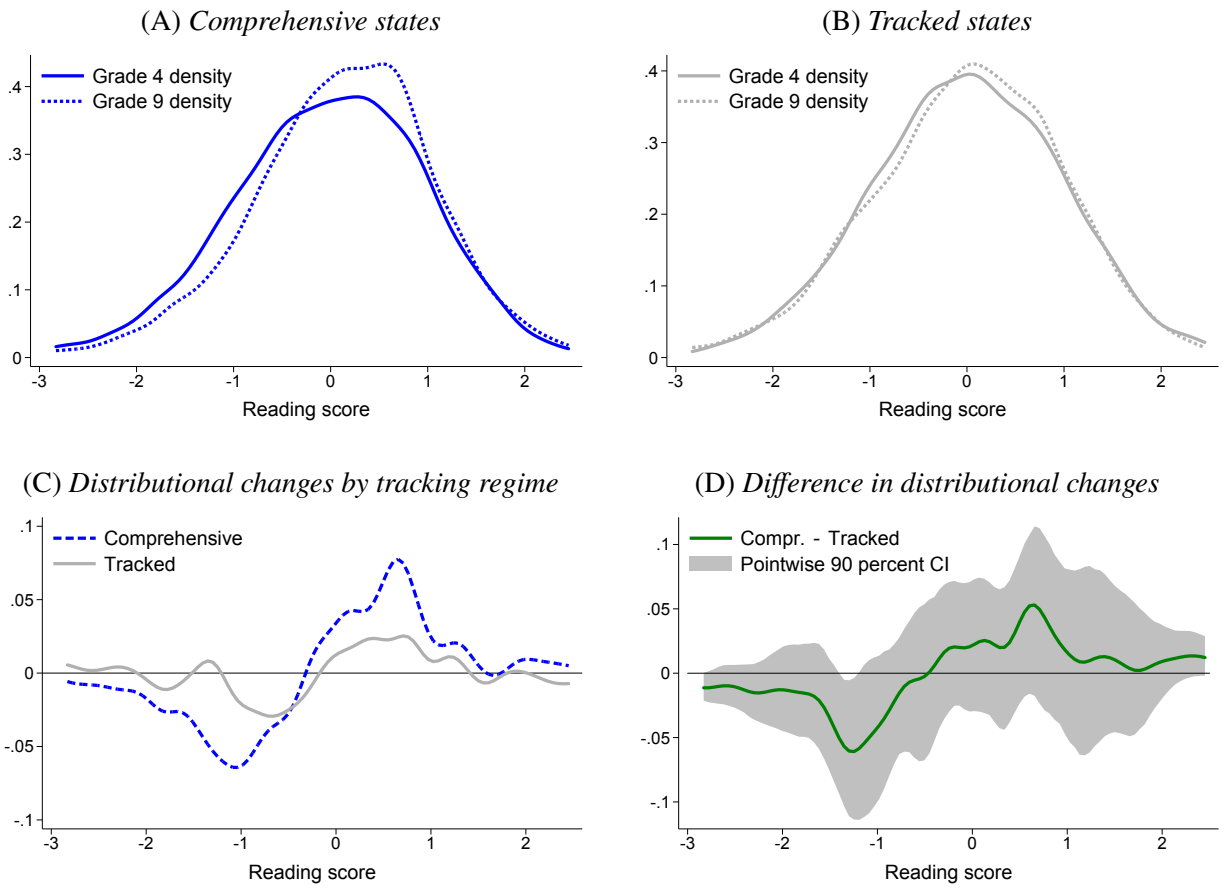

Figure B2. *Distributional analysis for reading scores in the IQB data.*

Notes: Figure B2 repeats the distributional analysis from Figure 8 using fourth and ninth grade reading scores in the IQB DD sample ($N = 20,139$). The density estimates are based on the first plausible value and apply student sampling weights. Otherwise the curves are constructed analogously to Figure 8.

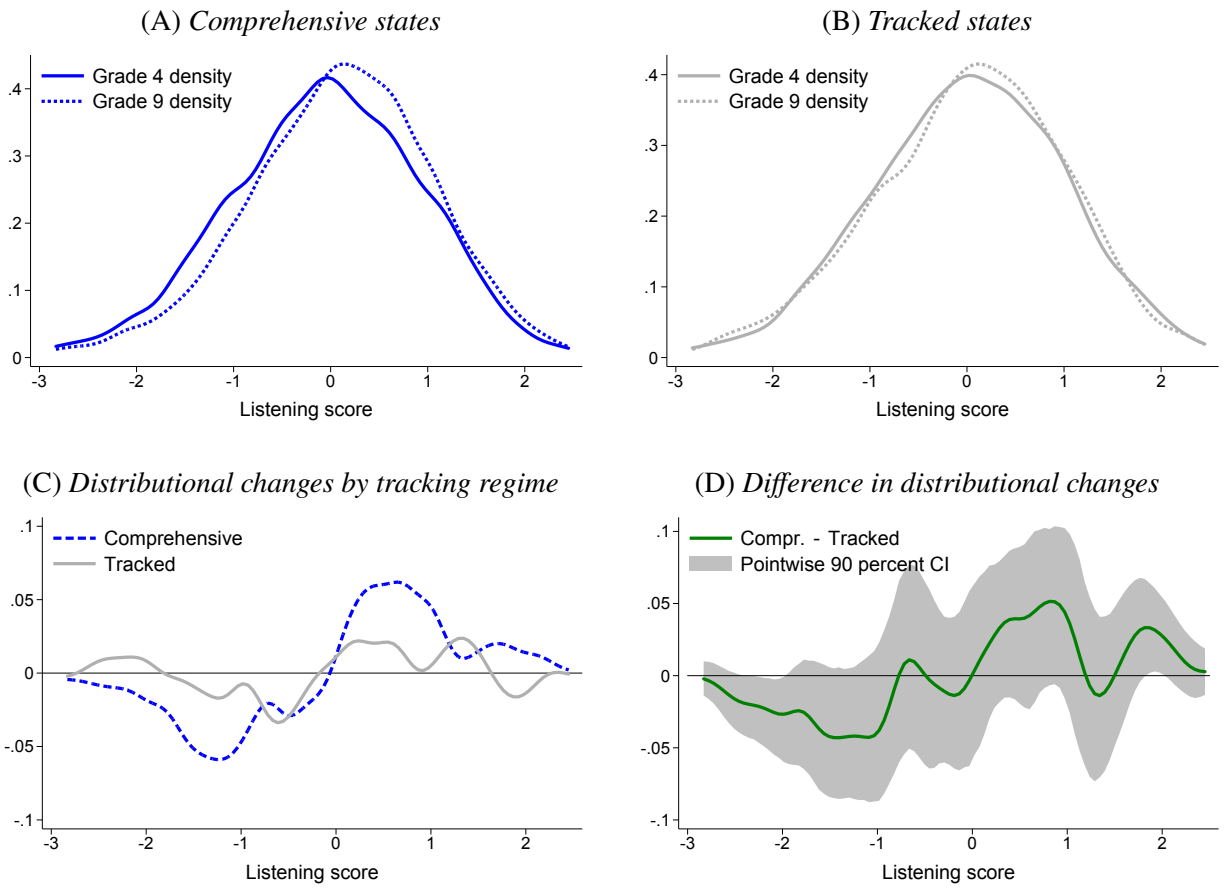

Figure B3. *Distributional analysis for listening scores in the IQB data.*

Notes: Figure B3 repeats the distributional analysis from Figure 8 using fourth and ninth grade listening scores in the IQB DD sample ($N = 20, 139$). The density estimates are based on the first plausible value and apply student sampling weights. Otherwise the curves are constructed analogously to Figure 8.

Table B1. *Summary statistics and balance test for school characteristics.*

	Primary school			Secondary school			Double difference	
	Compr. states	Tracked states	*p*-value (1)=(2)	Compr. states	Tracked states	*p*-value (4)=(5)	$\hat{\beta}^{std}_{DD}$	*p*-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: NEPS data								
Teacher age (years)	47.52	46.64	(0.38)	48.82	47.36	(0.33)	0.14	(0.61)
Further training past year (index)	-0.03	0.02	(0.87)	0.31	0.12	(0.57)	0.14	(0.39)
School size (students)	36.79	47.08	(0.20)	63.92	75.48	(0.52)	-0.04	(0.72)
Student-teacher ratio	14.02	14.66	(0.52)	11.03	12.56	(0.18)	-0.31	(0.46)
School equipment (index)	-0.14	0.07	(0.47)	0.25	0.12	(0.50)	0.31	(0.31)
Educational support (index)	0.14	-0.05	(0.39)	0.92	0.21	(0.02)	0.53	(0.11)
Extracurriculars (index)	0.25	-0.05	(0.29)	-0.10	-0.30	(0.66)	-0.10	(0.89)
Quality control (index)	0.02	0.01	(0.97)	0.05	0.10	(0.91)	-0.06	(0.45)
N schools	62	261		29	133		485	
Panel B: IQB data								
Teacher job experience (years)	23.56	18.27	(0.10)	20.81	13.92	(0.11)	0.14	(0.34)
Further training past two years (hours)	25.60	27.12	(0.58)	25.16	29.50	(0.42)	-0.08	(0.57)
Private school (binary)	0.06	0.02	(0.14)	0.07	0.08	(0.88)	-0.21	(0.20)
Class size	18.89	20.23	(0.06)	23.26	24.10	(0.52)	0.11	(0.66)
All-day schedule/week (days)	2.15	1.94	(0.85)	2.86	2.73	(0.80)	-0.04	(0.95)
Homework support (binary)	0.78	0.72	(0.43)	0.75	0.69	(0.41)	-0.00	(0.99)
Extracurricular learning (binary)	0.32	0.24	(0.13)	0.24	0.28	(0.49)	-0.26	(0.09)
N schools	578	377		340	298		1593	

Notes: Table B1 reports means of school covariates by state group for primary schools (columns 1–2) and non-academic-track secondary schools (columns 4–5). Panel A refers to the NEPS data, where the primary school information comes from the NEPS-SC2 data set and the secondary school information comes from the (main) NEPS-SC3 data set (see Appendix A for details). Panel B refers to the IQB data, where the primary school information comes from the fourth-grade IQB11 data set and the secondary school information from the ninth-grade IQB15 data set. All indices are normalised to mean zero and standard deviation one, separately by grade level. Columns 3 and 6 report *p*-values for tests for zero mean differences in each covariate between the Tracked and Comprehensive states' primary and secondary schools, respectively. Column 7 reports normalised double-differences for each variable, which equal the second difference between Comprehensive and Tracked states between primary and secondary school, divided by the variable's standard deviation. Column 8 reports *p*-values testing for a zero double-difference between state groups and grades. All tests are based school-level regressions and 999 wild cluster bootstraps iterations with Webb weights, clustering at the state level. All figures and tests in this table use school weights provided in the respective data sets.

Table B2. *Seventh-grade DD robustness checks: alternative model specifications and additional control variables.*

	Ind. controls	School controls one-by-one								All school controls				Further school & state controls					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
Panel A: Mathematics																			
Comprehensive	0.165*** (0.01)	0.183*** (0.00)	0.171*** (0.00)	0.135 (0.14)	0.269*** (0.00)	0.173*** (0.00)	0.166*** (0.00)	0.165** (0.03)	0.164*** (0.00)	0.259** (0.05)	0.319* (0.07)	0.246** (0.01)	0.111 (0.24)	0.259** (0.04)	0.307** (0.02)	0.246** (0.03)	0.313*** (0.01)	0.259** (0.04)	0.230** (0.03)
Panel B: Reading																			
Comprehensive	0.243** (0.02)	0.254** (0.04)	0.244** (0.02)	0.240** (0.03)	0.344*** (0.00)	0.251** (0.02)	0.243** (0.01)	0.242** (0.02)	0.238** (0.02)	0.363*** (0.00)	0.375** (0.01)	0.351** (0.01)	0.552* (0.09)	0.362*** (0.00)	0.390*** (0.00)	0.354*** (0.00)	0.365** (0.01)	0.363*** (0.00)	0.333*** (0.00)
Teacher age		✓								✓	✓	✓		✓	✓	✓	✓	✓	✓
Teacher further training			✓							✓	✓	✓		✓	✓	✓	✓	✓	✓
School size				✓						✓	✓	✓		✓	✓	✓	✓	✓	✓
Student-teacher ratio					✓					✓	✓	✓		✓	✓	✓	✓	✓	✓
Facilities						✓				✓	✓	✓		✓	✓	✓	✓	✓	✓
Educational support							✓			✓	✓	✓		✓	✓	✓	✓	✓	✓
Extracurriculars								✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
Quality control									✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Grade × School controls													✓						
Grade × Ind. controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Private school														✓					
Class size															✓				
Timing summer break																✓			
School expenditure																	✓		
Binding teacher rec.																		✓	
Rheinland-Palatinate																			✓
Student sampling weights											✓								
Saturated (w/ state FE)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Unsaturated (w/o state FE)												✓							
N	5074	5074	5074	5074	5074	5074	5074	5074	5074	5074	4890	5074	5074	5074	5074	5073	5074	5074	5360

Notes: Table B2 reports OLS regression results for variations of the DD model for grade 5 and 7 maths and reading test scores in the NEPS DD sample. Column 1 reposts the saturated DD model with grade-level-interacted student controls from column 4 of Table 2 for reference. Columns 2–9 separately add the main eight school controls to this model. Column 10 presents the full model where all school controls are added jointly, corresponding to column 5 in Table 2. Column 11 presents results from estimating the full model of column 10 with student sampling weights, provided by the NEPS. Column 12 presents the fully-controlled, *unsaturated* DD model, which instead of state fixed effects includes an indicator for the Comprehensive states. Column 13 returns to the saturated model of column 10 and adds interactions between grade level and all school controls. Columns 14–18 separately add the following variables to the full model of column 10: a school-level indicator for private schools (note: due to the imputation this is a state-level average for grade 5 observations), a school-level measure of average class size (note: state-level average for grade 5 observations), a state-level measure of the time (in months) since the end of the summer break at the day of testing, a state-level measure of per pupil public expenditure for schools, a state-level indicator for binding teacher recommendations. Column 19 presents the full model from column 10, adding observations from the otherwise excluded state Rheinland-Palatinate (RP) to the sample (RP is added to the Tracked states). p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3. *DD regression for primary school maths score.*

Dependent variable:	Mathematics (1)
Compr. state × Grade 4	-0.015
	(0.77)
	[-0.118, 0.125]
Compr. state	0.043
	(0.77)
	[-0.321, 0.345]
Grade 4	-0.000
	(1.00)
	[-0.161, 0.093]
Adjusted R^2	0.000
N state clusters	12
N Compr. state students	2038
N Tracked state students	9163

Notes: Appendix table B3 reports regression results for the unsaturated DD model applied to primary school maths achievement, i.e. from regressing grade 2 and grade 4 test scores on an intercept, an indicator for the Comprehensive states, an indicator for grade 4 observations and their interaction. This tests whether already in primary school (i.e. before students are exposed to different tracking regimes) mean achievement diverges between Comprehensive and Tracked states. *p*-values in parantheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4. *Heterogeneity analysis for seventh-grade achievement in DD and VA models.*

Dependent variable:	Level test scores				Grade-5-to-7 gain scores					
Model specification:	DD (1)	DD (2)	DD (3)	DD (4)	FD (5)	VA (6)	VA (7)	VA (8)	VA (9)	VA (10)
Panel A: Mathematics										
Comprehensive schooling	0.207***	0.226**	0.140	0.160**	0.156*	0.153**	0.115	0.144*	0.136**	0.039
	(0.01)	(0.03)	(0.12)	(0.01)	(0.06)	(0.04)	(0.22)	(0.05)	(0.03)	(0.57)
	[0.10, 0.50]	[0.01, 0.67]	[-0.06, 0.30]	[0.06, 0.29]	[-0.01, 0.34]	[0.00, 0.28]	[-0.07, 0.28]	[-0.00, 0.26]	[0.02, 0.29]	[-0.12, 0.23]
× Female		-0.020					0.078			
		(0.88)					(0.42)			
		[-0.28, 0.26]					[-0.16, 0.26]			
× Low SES			0.073					0.019		
			(0.47)					(0.85)		
			[-0.13, 0.49]					[-0.15, 0.35]		
× Migration background				0.026					-0.024	
				(0.91)					(0.87)	
				[-0.65, 0.46]					[-0.33, 0.51]	
× Below median gr. 5 score										0.223
										(0.15)
										[-0.13, 0.48]
Panel B: Reading										
Comprehensive schooling	0.277**	0.212	0.353**	0.209**	0.248**	0.228***	0.178	0.198	0.180*	0.179**
	(0.02)	(0.23)	(0.04)	(0.05)	(0.02)	(0.00)	(0.11)	(0.20)	(0.06)	(0.04)
	[0.05, 0.65]	[-0.10, 0.64]	[0.02, 0.62]	[0.00, 0.37]	[0.06, 0.47]	[0.09, 0.36]	[-0.07, 0.36]	[-0.18, 0.34]	[-0.01, 0.27]	[0.03, 0.38]
× Female		0.145					0.113			
		(0.32)					(0.38)			
		[-0.14, 0.43]					[-0.13, 0.41]			
× Low SES			-0.197					0.064		
			(0.20)					(0.67)		
			[-0.41, 0.10]					[-0.17, 0.58]		
× Migration background				0.178					0.166	
				(0.40)					(0.24)	
				[-0.42, 0.55]					[-0.12, 0.55]	
× Below median gr. 5 score										0.086
										(0.39)
										[-0.24, 0.30]
State FE	✓	✓	✓	✓						
Grade 5 score						✓	✓	✓	✓	✓
N Compr. state students	882	882	882	860	269	269	269	269	269	269
N Tracked state students	4192	4192	4192	4130	1377	1377	1377	1377	1371	1377

Notes: Columns 1–4 are based on the NEPS DD sample of non-academic-track students. Column 1 presents results from the saturated DD model without controls, corresponding to column 2 of Table 2. Columns 2–4 fully interact the DD model with indicators for female, below-median socio-economic status (SES) and migration background students, respectively. SES is operationalised as the first principal component of the following variables and their missing dummies: household income, highest parental years of education, home posessions index and the number of books at home. Columns 5–10 are based on the NEPS 5-to-7 panel sample of non-academic-track students. Column 5 presents results for the first-differenced (FD) model, i.e. from regressing grade-5-to-7 gain scores on an indicator variable for the Comprehensive states. The value-added (VA) model in column 6 adds grade 5 test scores as a regressor to the FD model. Columns 7–10 interact the Comprehensive states indicator in the VA model with indicators for female, low-SES, migration background and below-median grade 5 score students, respectively (while each time also adding the indicator as own regressor). p-values in parantheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B5. *Effect persistence until ninth-grade.*

Dependent variable:	Level test scores		Grade-5-to-9 gain scores	
Model specification:	DD	DD	VA	VA
	(1)	(2)	(3)	(4)
Panel A: Mathematics				
Comprehensive schooling	0.054	0.084	0.088	-0.049
	(p = 0.53)	(0.39)	(0.26)	(0.81)
	[-0.10, 0.33]	[-0.27, 0.43]	[-0.08, 0.25]	[-0.39, 0.25]
× Below median gr. 5 score				0.257
				(0.38)
				[-0.22, 0.78]
N Compr. state students	753	753	186	186
N Tracked state students	3619	3619	1100	1100
Panel B: Reading				
Comprehensive schooling	0.198	0.290	0.163	0.133
	(0.20)	(0.30)	(0.35)	(0.37)
	[-0.23, 0.44]	[-0.30, 0.52]	[-0.14, 0.38]	[-0.29, 0.32]
× Below median gr. 5 score				0.061
				(0.69)
				[-0.25, 0.50]
N Compr. state students	754	754	191	191
N Tracked state students	3522	3522	1064	1064
Controls		✓		
State FE	✓	✓		
Grade 5 score			✓	✓

Notes: Columns 1–2 present OLS regression results for the saturated DD model for grade 5 and 9 test scores in maths (Panel A) and reading (Panel B), estimated on the grade 9 NEPS DD sample of non-academic-track students. Column 1 presents the model without controls and column 2 adds student covariates, interacted with grade level, and school covariates (see notes of Table 2). Column 3 presents results for the value-added (VA) model, i.e. from regressing grade-5-to-9 gain scores on an indicator for the Comprehensive states and the grade 5 score. The regressions use the panel sample of non-academic-track students for whom both grade 5 and grade 9 test scores are observed. Column 4 interacts the Comprehensive state indicator with an indicator for students with below-median grade 5 test scores (and adds this indicator as a separate regressor). p-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations, clustering at the state level and using Webb weights. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B6. *Summary statistics for the IQB DD sample.*

	Non-academic primary school students (IQB11)			Non-academic secondary school students (IQB15)		
	Compr. states	Tracked states	*p*-value (1)=(2)	Compr. states	Tracked states	*p*-value (4)=(5)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pre-treatment outcomes						
Grade 4 mathematics score	-0.09	-0.00	(0.62)			
Grade 4 reading score	-0.02	-0.00	(0.87)			
Grade 4 listening score	-0.08	0.00	(0.32)			
Self-concept German	0.00	-0.00	(0.94)			
Social integration	-0.13	-0.00	(0.11)			
Reading motivation	-0.09	-0.00	(0.05)			
Attitude towards reading	-0.02	0.00	(0.74)			
Private tutoring	0.23	0.24	(0.81)			
Panel B: Student characteristics						
Female (binary)	0.48	0.46	(0.10)	0.46	0.47	(0.82)
Age	10.63	10.52	(0.09)	15.62	15.66	(0.19)
Migration background (binary)	0.15	0.31	(0.06)	0.18	0.36	(0.06)
Foreign language at home (binary)	0.12	0.21	(0.07)	0.15	0.28	(0.10)
Parental education:						
Low	0.45	0.43	(0.25)	0.50	0.49	(0.60)
Mid	0.41	0.41	(0.88)	0.34	0.34	(0.93)
High	0.14	0.16	(0.13)	0.16	0.17	(0.47)
Parental HISEI score	43.28	44.29	(0.18)	44.82	44.68	(0.91)
N students	6573	4585		7009	6733	

Notes: Table B6 reports means of pre-treatment outcomes and student covariates by state group for primary school students classified as non-academic (columns 1–2) and non-academic-track secondary school students (columns 4–5). The former are based on the grade 4 IQB11 data and the latter on the IQB15 grade 9 data. All figures use student weights. Columns 3 and 6 report *p*-values for tests for zero mean differences between Tracked and Comprehensive states. Test are based on 999 wild cluster bootstraps iterations, clustering at the school level, using Webb weights.

Table B7. *IQB DD results by plausible value.*

Dependent variable:	Reading		Listening	
Plausible value	$\hat{\beta}_{DD}$	*p*-value	$\hat{\beta}_{DD}$	*p*-value
PV1	0.160	(0.12)	0.154	(0.04)
PV2	0.142	(0.21)	0.141	(0.06)
PV3	0.129	(0.25)	0.132	(0.08)
PV4	0.150	(0.14)	0.175	(0.02)
PV5	0.163	(0.18)	0.15	(0.05)
PV6	0.122	(0.24)	0.123	(0.10)
PV7	0.131	(0.18)	0.148	(0.08)
PV8	0.154	(0.12)	0.17	(0.02)
PV9	0.156	(0.14)	0.164	(0.02)
PV10	0.139	(0.13)	0.121	(0.08)
PV11	0.162	(0.10)	0.139	(0.04)
PV12	0.153	(0.11)	0.175	(0.02)
PV13	0.164	(0.11)	0.152	(0.06)
PV14	0.166	(0.10)	0.142	(0.04)
PV15	0.173	(0.07)	0.165	(0.03)
Average	0.151	(.15)	0.150	(.05)

Notes: Table B7 displays coefficient estimates with accordant wild cluster bootstrapped *p*-values for the saturated DD model without controls for reading and listening scores in the IQB sample separately by plausible value. For details about the model see Table 4.

Table B8. *DD regressions for behavioural and socio-emotional outcomes in the IQB data.*

Dependent variable:	Self-concept languages (1)	Reading motivation (2)	Attitude towards reading (3)	Social integration (4)	Private tutoring (5)
Comprehensive schooling	-0.012	0.133**	0.152**	0.068	-0.047
	(0.90)	(0.02)	(0.02)	(0.42)	(0.21)
	[-0.20, 0.18]	[0.04, 0.24]	[0.04, 0.26]	[-0.10, 0.24]	[-0.13, 0.02]
Controls	✓	✓	✓	✓	✓
State FE	✓	✓	✓	✓	✓
R^2	0.057	0.077	0.089	0.030	0.030
N state clusters	12	12	12	12	12
N Compr. state students	9465	6743	6907	9594	7872
N Tracked state students	7532	5752	5906	7671	6705

Notes: This table reports results for the fully controlled, saturated DD model applied to different non-cognitive outcomes in the IQB DD sample of non-academic-track students, each time retaining all observations with non-missing values for the respective dependent variable. 'Private tutoring' is an indicator variable equal to one if the student reports receiving private tutoring. All remaining variables are composite scores designed by the IQB, intended to measure the indicated psychological construct, each based on several survey items measured on 4-point Likert scales. I standardise all of them to mean zero and standard deviation one in the group of Tracked states' non-academic-track students, separately by grade. *p*-values in parentheses and 95%-confidence sets in brackets stem from 999 wild cluster bootstrap iterations using Webb weights, clustering at the state level. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1705	Giordano Mion Luca David Opromolla Gianmarco I.P. Ottaviano	Dream Jobs
1704	Gabriel M. Ahlfeldt Jason Barr	The Economics of Skyscrapers: A Synthesis
1703	Luke Milsom Vladimír Pažitka Isabelle Roland Dariusz Wójcik	Gravity in International Finance: Evidence from Fees on Equity Transactions
1702	Jonathan Colmer Dajun Lin Siying Liu Jay Shimshack	Why are Pollution Damages Lower in Developed Countries? Insights from High Income, High-Pollution Hong Kong
1701	Paul Cheshire Katerina Kaimakamis	Offices Scarce but Housing Scarcer: Estimating the Premium for London Office Conversions
1700	Francesco Campo Mariapia Mendola Andrea Morrison Gianmarco Ottaviano	Immigrant Inventors and Diversity in the Age of Mass Migration
1699	Mariana Spatareanu Vlad Manole Ali Kabiri Isabelle Roland	Bank Default Risk Propagation along Supply Chains: Evidence from the U.K.
1698	Georg Graetz Björn Öckert Oskar Nordström Skans	Family Background and the Responses to Higher SAT Scores
1697	Federico Cingano Fadi Hassan	International Financial Flows and Misallocation

1696	Piero Montebruno	Disrupted Schooling: Impacts on Achievement from the Chilean School Occupations
1695	Ester Faia Sébastien Laffitte Maximilian Mayer Gianmarco Ottaviano	Automation, Globalization and Vanishing Jobs: A Labor Market Sorting View
1694	Ulrich J. Eberle	Damned by Dams? Infrastructure and Conflict
1693	Abel Brodeur Andrew E. Clark Sarah Flèche Nattavudh Powdthavee	COVID-19, Lockdowns and Well-Being: Evidence from Google Trends
1692	Fabrice Defever José-Daniel Reyes Alejandro Riaño Gonzalo Varela	All These Worlds are Yours, Except India: The Effectiveness of Cash Subsidies to Export in Nepal
1691	Adam Altmejd Andrés Barrios-Fernández Marin Drlje Joshua Goodman Michael Hurwitz Dejan Kovac Christine Mulhern Christopher Neilson Jonathan Smith	O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries
1690	Michael Amior Alan Manning	Monopsony and the Wage Effects of Migration
1689	Frank Pisch	Managing Global Production: Theory and Evidence from Just-in-Time Supply Chains

The Centre for Economic Performance Publications Unit
Tel: +44 (0)20 7955 7673 Email info@cep.lse.ac.uk
Website: http://cep.lse.ac.uk Twitter: @CEP_LSE