DIW BERLIN

1775

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2018

Better Together? Heterogeneous Effects of Tracking on Student Achievement

Sönke Hendrik Matthewes

Opinions expressed in this paper are those of the author(s) and do not necessarily reflect views of the institute.

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Better Together? Heterogeneous Effects of Tracking on Student Achievement

Sönke Hendrik Matthewes[†]
WZB Berlin, DIW Berlin & Freie Universität Berlin

December 5, 2018

Abstract

This study estimates mean and distributional effects of early between-school ability tracking on student achievement. For identification, I exploit heterogeneity in tracking regimes between German federal states. After comprehensive primary school, about 40% of students are selected for the academic track and taught in separate schools in all states. The remaining students, however, are either taught comprehensively or further tracked into two different school forms depending on the state. I estimate the effects of this tracking on students' mathematics and reading test scores with a difference-in-difference-in-differences estimator to eliminate unobserved heterogeneity in achievement levels *and* trends between states. I find substantial achievement gains from comprehensive versus tracked schooling at ages 10–12. These average effects are almost entirely driven by low-achievers. I do not find evidence for negative effects of comprehensive schooling on the achievement of higher performing students. My results show that decreasing the degree of tracking in early secondary school can reduce inequality while *increasing* the efficiency of educational production.

JEL Codes: I24, I28, J24.
Keywords: Tracking, student achievement, inequality, triple differences

[†] WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin, Germany. Email: `soenke.matthewes@wzb.eu`. Tel: +49-(0)30/25491-348. I thank Jan Marcus, Heike Solga, Katharina Spieß, Benjamin Elbers, my colleagues at the 'Skill formation and labour markets' WZB research unit, as well as participants of the EALE conference 2018, COMPIE conference 2018, DFG SPP 174 workshop 2018 and the BeNA labour workshop 2017 for helpful comments on presentations and earlier drafts. I gratefully 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:SC3:7.0.1, and SC Kindergarten, doi:10.5157/NEPS:SC2:7.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the BMBF. As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

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 increases inequality by construction as it deprives lower track students of more able peers (Sacerdote, 2011).[2] 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 implementation is likely to be error-prone as ability is not directly observable (Brunello et al., 2007). In particular, there might be systematic bias against particular socio-economic groups, due to biased teacher expectations or selective parental involvement in the placement of their children (van Ewijk, 2010; Jackson, 2013). In addition to being unfair, both error and bias in track placement would dissipate the claimed benefits of tracking related to homogeneity of classrooms as school tracks would not actually reflect

[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).

[2] Conversely, a reference group of lower ability, implying a higher ordinal rank, might boost students' academic self-concept and in turn achievement (e.g. Marsh & Parker, 1984; Murphy & Weinhardt, 2014). This so-called 'big-fish-little-pond' effect might partially counteract the positive peer effects referred to here.

different ability levels.

Given these opposing mechanisms, the 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 benefit students of all ability levels. If the hypothesised negative effects along the equity dimension are at work, then tracked school systems should depress student achievement at the bottom. In that case, tracking might 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 effects of early between-school tracking has emerged in the empirical literature.[3] As both theoretically and empirically it is clear that high-track students have nothing to fear from tracking – if anything, they gain –, the fundamental question that remains is whether students assigned to lower tracks in fact lose out from being separated from their higher achieving peers; and, if so, how large these losses are.

This paper exploits unique within-country between-state variation in tracking practices in Germany that has emerged as a result of federalism to isolate the effect of early between-school tracking on student achievement in lower tracks. While primary school is comprehensive everywhere, the grouping of students in secondary school differs between states. In some states Germany's traditional three-tiered between-school tracking system is still in place and the secondary systems consists of low-, middle- and academic-track schools. Others have transformed their secondary school system into a two-tiered one by conflating low- and middle-track schools into one school form. The academic-track school form, called *Gymnasium*, has been left unaltered in all states. This means that while in all states about 40% of students transition to the academic track after comprehensive primary school, depending on the state, the rest of the student body is either further divided between low- and middle-track schools or taught comprehensively for another two years.[4]

My research design first exploits this variation in tracking in a difference-in-differences

[3] I provide a detailed review of the literature on the effects of tracking on student outcomes at the end of this section.

[4] This refers to all 12 federal states under investigation (out of 16 in total).

(DD) framework: I estimate how achievement of students in the non-academic tracks develops differently in the first two years of secondary school depending on whether they are taught comprehensively or separated. This strategy controls for time-constant heterogeneity between states and general achievement trends between grades. Because the DD estimate might still be confounded by state-specific achievement trends (i.e. state-grade-specific shocks to achievement) the differences in achievement growth for non-academic-track students are additionally compared 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. For the estimation, I use individual-level panel data on mathematical and reading competence of the German National Educational Panel Study (NEPS), which allow me to track students' progress from before tracking starts until four years after. Hence, I can also evaluate the persistence of effects through the end of secondary school for low-track students. After having thus established the mean effect, the second part of the analysis explores 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 ability, as measured by students' position in the pre-tracking achievement distribution.

I find substantial achievement gains from comprehensive versus tracked schooling in early secondary school. The effect of comprehensive schooling in grades 5 and 6 on grade 7 achievement is estimated to be 0.17 standard deviations (SD) in mathematics and 0.21 SD in reading. These results are invariant to the inclusion of academic-track students as an additional control group in the DDD model. A battery of robustness checks, such as comparing achievement trends in primary school, controlling for potential school-level confounders and excluding outlier states, further corroborates the validity of my findings. While the average effect of comprehensive schooling is clearly positive, it is somewhat imprecisely estimated. The heterogeneity analysis reveals that it is the lower tail of the initial achievement distribution that drives the average results. For low-achievers, the effects are precisely estimated, large and persistent. For high-achievers, in contrast, I find zero effects. Therefore, comprehensive early secondary schooling has an equalising effect on the distribution of test scores. Importantly, in my setting it does not trade off efficiency against equity, but enhances both. Auxiliary analyses suggest that motivational and socio-emotional mechanisms might underlie these effects.

Note that the treatment effect identified in this paper applies 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 relation between classroom homogeneity and student learning. My results highlight large costs associated with early between-school tracking for 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 regarding the early tracking of students into different schools seem warranted.

This paper contributes to the literature on the impact of between-school tracking, which, in the face of the severity of the endogeneity issues involved, could only produce tentative evidence so far. The most credible results stem from two strands of the literature. The first exploits temporal *within-country* variation in tracking practices induced by de-tracking reforms. An array of papers analysing reforms in the Nordic countries find achievement gains for students from lower socio-economic backgrounds and no evidence that de-tracking lowers overall achievement (see Meghir & 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 estimated effects cannot be unequivocally attributed to changes in tracking, however. Analyses of Britain's de-tracking reform, which all use the fact that implementation was staggered across regions, have generated 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 and positive, whereas losses for the stronger students are small or absent (Guyon et al., 2012).[6]

Given that de-tracking reforms are rare, a second strand of the literature uses the considerable variation in tracking practices *between countries*. Naturally, this comes at the cost of potentially severe unobserved heterogeneity, which researchers try to circumvent in different ways. One 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 & 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 while secondary school is either comprehensive or tracked. These studies use DD to estimate how test scores change differently from elementary to secondary school in countries with tracked school systems versus those with comprehensive schooling. Most results indicate that tracking increases inequality in student achievement,[7] though Waldinger (2007)

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

[6] Piopiunik (2013) analyses another more narrow tracking reform implemented in the German state of Bavaria. The Bavarian pre-post differences closely resemble the (contemporaneous) differences in tracking analysed in this paper. Reassuringly, his findings based on a single state's reform (the reform-induced increase in tracking caused achievement losses in the lower tracks and widened the overall achievement distribution) are confirmed in this study for the whole of Germany.

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

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 assigning countries to 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 & 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 heterogeneity in tracking regimes 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 whom there is no difference in tracking between states. Third, I draw on individual-level panel data with test scores that are vertically scaled across waves.[8] Hence, in contrast to previous studies that pooled various achievement tests with different metrics, my effect sizes are meaningful and can be interpreted cardinally. Fourth, I am evaluating the impact of tracking along both the efficiency and the equity dimension by analysing effect heterogeneity and the distributional consequences next to mean effects. This is key given that the debate on tracking revolves around a perceived efficiency-equity trade-off.

Finally, in seeming contrast to the results presented here, Dustmann et al. (2017) find no effect of higher track placement on attainment or earnings in the German context. Crucially, however, their individual-level instrumental variables strategy identifies a local average treatment effect (LATE) of track placement for the group of students *at the margin between two tracks*. This effect margin is relevant for understanding the long-term consequences of initial misallocation of hard-to-assign (i.e. marginal) students to tracks, given an early between-school tracking system like that of Germany. Yet, if effects of tracking are heterogeneous, this LATE tells us little about whether tracking is desirable in the first place as it does not apply to the much larger group of non-marginal students. My results suggest that the zero effect found by Dustmann et al. (2017) might not be representative of the whole distribution of effects. Using state-level variation in tracking that shifts more than just marginal students, I find that the separation of students into different schools at an age as early as 10 depresses achievement for a sizeable group of low-achievers, thus putting them at a double disadvantage.[9]

[8] This means that the test scores are designed to be comparable over time. Achievement at different ages is measured on the same scale allowing the researcher to compare individual students' progress over time.

[9] As Dustmann et al. (2017) study cohorts that started secondary school between 1971 and 1986, another reason

The paper is structured as follows: Section 2 explains the institutional background. Section 3 sets out the empirical design. Section 4 describes my data source. Section 5 present the results and section 6 the robustness checks. Finally, Section 7 discusses implications and concludes.

2 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 Standing Conference of the Ministers of Education and Cultural Affairs of the Federal States (*Kultusministerkonferenz*; KMK) harmonises education policies between states considerably (KMK, 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. 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 generally lasts 4 years and is taught comprehensively with no ability grouping of students within or between schools.[10] Differences between states emerge thereafter (see Figure 1 for a schematic overview). In West Germany, after fourth grade, i.e. at the age of around 10, students have traditionally been sorted into one of three vertically ordered school types – *Hauptschule*, *Realschule* and *Gymnasium*, representing low, middle and high (academic) track – based on their performance in primary school.[11] 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 curriculum, lasts 8 or 9 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 sector (also in reputation). The middle track (i.e. *Realschule*) provides general knowledge, lasts 6 years and is supposed to prepare students for advanced vocational and professional education. If they complete the middle 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 applied general education, lasts 5 or 6 years and prepares students for technical vocational education. Also here, after completion, upgrading to middle-track schools is possible under specific conditions.

for different findings could be that the effect of tracking changed over time.

[10] In the two states of *Berlin* and *Brandenburg*, primary school lasts 6 years. For this reason, they are not part of the analysis.

[11] In all federal states students receive a track recommendation by their teacher based on their performance in primary school. Whether it is binding depends on the federal 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).

Coming from a comprehensive schooling tradition, after reunification the East German states did not adopt the three-tiered school system one-to-one.[12] While taking over the three-tiered differentiation in school-leaving certificates, they opted for a *two-tiered* school structure (Edelstein & Nikolai, 2013). The only difference between the two systems is that there are no separate low- and middle-track schools, but rather one school form, labelled School with Multiple Tracks (*Schule mit mehreren Bildungsgängen*; SMT). Here, all students not attending an academic-track *Gymnasium* school are taught together. If a student leaves an SMT school after 5 years (without failing the year, of course) she receives the low degree. If she has the required grades, she may stay on for another year to earn the intermediate degree. Hence, the difference between the two systems is one of tracking only.

Due to a number of reforms, there is no longer a geographical divide between a two-tiered East and a three-tiered West, however. Over time, low-track schools have become increasingly stigmatised due to falling student numbers and the lack of prospects for its graduates (Helbig & Nikolai, 2015). Consequently, public pressure grew to reform the traditional three-tiered school structure in West Germany. This led to several Western states reforming their school system along the lines of the two-tiered East German system. Just like in the East, the three different school-leaving certificates, as well as a distinct academic track consisting of *Gymnasium* schools, were retained, but separate low- and middle-track schools were abolished and replaced with so-called 'comprehensive schools' (*Gesamtschule*). Thus, just like SMT schools in the East, these schools comprise all non-academic-track students.[13]

Importantly, neither East German SMT schools nor West German comprehensive schools specifically assign students to the low or middle track during the first two years of secondary school (grades 5 and 6) (Leschinsky, 2008). Instead, in these two years, classrooms continue to be formed disregarding ability or previous performance. It is only from grade 7 onwards that these schools may group students by ability by forming track-specific classes.[14] Note that the first two years in non-academic-track secondary schools are strongly harmonised: official information of the KMK (2014) shows that curriculum and learning goals for the non-academic tracks in grades 5 and 6 focus on the acquisition of a standard set of basic general knowledge that is virtually indistinguishable between states. Even between low- and middle-track schools there is no difference in the topics covered, although treatment might be at a slightly higher level in middle-track schoools.[15]

[12] Except for *Mecklenburg-Vorpommern*, which did initially adopt the three-tiered system.

[13] The difference between the East German SMT schools 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 & Nikolai, 2015). In practice, this difference is only relevant in much later grades than those studied here.

[14] Schools that use within-school streaming from grade 7 onwards are labelled 'cooperative' while those that continue to teach comprehensively are called 'integrative'. In most states, it is up to the individual school which model to implement. To my knowledge there is no information on which model is more prevalent.

[15] The first two years of secondary school have a special status in the German education system. Labelled 'orientation stage' (*Orientierungsstufe*), grades 5 and 6 formally allow for the possibility to switch between tracks. In practice, this happens quite rarely (about 5% of students switch according to Bellenberg (2012)). To ensure

Altogether, these developments mean that in 2010 there were 5 federal states where, after a four-year comprehensive primary school period, students still entered the traditional three-tiered secondary school system and 7 federal states where, after a four-year comprehensive primary school period, students entered a two-tiered secondary school system (see Figure 2 for a map).[16] As should have become clear, this dichotomy is relevant only for non-academic-track students: students sorted into the academic track are separated from their lower-achieving peers after fourth grade to attend *Gymnasium* in both regimes. The remaining non-academic-track students, in contrast, are either further *tracked* into two different school forms (in states with the three-tiered regime; henceforth, called the 'Tracked' states) or taught *comprehensively* for at least another two years (in states with the two-tiered regime; henceforth, called the 'Comprehensive' states).[17]

I conceptualise the two tracking regimes as two treatment conditions. Receiving the treatment of comprehensive instead of tracked schooling may affect achievement through three main channels. The first one is peer effects, as schools (and classes) will be more heterogeneous in terms of ability in the Comprehensive than the Tracked states. Tracking proponents claim that heterogeneity has a negative effect on achievement as teaching is less tailored. Tracking opponents, in contrast, claim that what really matters is the average ability of one's peers: high-ability peers are beneficial and low-ability peers harmful to learning. Second, for students who would be assigned the low track in the three-tiered system, the treatment increases academic standards, whereas for students who would be assigned the middle-track standards decrease. Note that, in practice, the former is the first-order effect because of a strong emphasis on not putting future middle-track students at a disadvantage by lowering their performance standards in early secondary school. Low-achievers might either lose out from being held to excessive academic standards or benefit if they grow with these standards. Third, by attending either a low- or a middle-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 might affect their academic self-concept, motivation and, subsequently, achievement. All of this holds true for 2 years. Thereafter, within-school tracking commences.

compatibility, the content taught is focused on basic general knowledge and quite harmonised across states and tracks (e.g. any specialisations only occur after grade 7).

[16] The analysis excludes 4 of 16 states: *Berlin*, *Brandenburg* and *Mecklenburg-Vorpommern* because the tracking decision is made after grade 6 instead of after grade 4; *Rheinland-Pfalz* because the state was transitioning from a three-tiered to a two-tiered system during the period under investigation and, hence, its treatment status is ambiguous. While *de jure* all separate low- and middle-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 statistics show that the cohort's share of students in a low- or middle-track school declined from 8% in 2010 to 6% in 2011 to 3% in 2012. A robustness check where *Rheinland-Pfalz* is assigned the Tracked states (as initially there was some tracking) leaves all results unchanged (results available on request).

[17] For completeness, it should be mentioned that in some of the Tracked states municipalities are allowed to offer comprehensive schools (IGS), where all three degrees can be eaerned, 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 4).

3 Empirical Strategy

The goal of this paper is to identify the effect of between-school tracking on student achievement. Given the described differences in tracking practices between states in Germany, an intuitive empirical strategy is to compare achievement outcomes (of non-academic-track students) between the Comprehensive and Tracked states. Clearly, however, whether states adopted the two-tiered instead of the traditional three-tiered tracking regime is not random. Accordingly, a state's tracking regime likely correlates with other factors determining student achievement, such as early childhood education policies or student body composition. As not all of these factors are known or can be observed by the researcher, simple cross-sectional comparisons between states, even if they condition on large control sets, are likely to give biased estimates of the causal effect of tracking on achievement.

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 (grade 7). In primary school, before the grade 5 test, all students were taught comprehensively. Hence, the grade 5 scores can be conceptualised as a 'pre-treatment' measure of achievement, which can be used to control for time-constant achievement differences between states unrelated to tracking.[18] In the two school years between the achievement tests, those students not attending the academic track are either split between low- and middle-track schools or taught comprehensively depending on their state of residence. Thus, to isolate the effect of comprehensive versus tracked schooling on achievement, I compare Comprehensive and Tracked states merely in terms of the progress (non-academic-track) students make between grades 5 and 7. Causal interpretation then hinges on the assumption that in the absence of differences in tracking achievement would have developed in parallel between the two state groups.

The corresponding difference-in-differences (DD) can be estimated by the following regression model for the test score Y_{isg} of non-academic-track student i in state s and grade g:

$$Y_{isg} = \theta_s + \lambda Grade7_g + \beta_{DD} Compr_{sg} + u_{isg}, \tag{1}$$

where θ_s is a state fixed effect that captures level achievement differences at the end of primary school between states. $Grade7_g$ is a dummy variable for grade 7 scores, such that λ captures general achievement growth between grade 5 and grade 7. The treatment indicator $Compr_{sg}$

[18] This is unless there are anticipation effects of tracking practices in secondary school that affect students' competence development in primary school (i.e. their grade 5 scores). Such effects might arise if children (or their parents), knowing that they will be placed in different tracks depending on their performance, already work harder pre-tracking in (more) tracked regimes. This mechanism could easily introduce non-negligible bias in the cross-country DD estimates (Eisenkopf, 2007). However, the fact that in both the Tracked and Comprehensive states the highest-performing children are assigned the academic track dramatically limits its importance in my setting: highly ambitious students/parents have strong incentives to perform in both systems.

takes value one for grade 7 observations from the Comprehensive states. Accordingly, the DD estimate β_{DD} captures how non-academic-track students develop differently when taught comprehensively compared to being tracked in the first two years of secondary school. Given the group-level treatment variable, all standard errors are clustered at the state level, allowing for correlation of errors within states across grades (Bertrand, et al., 2004).[19] Comparing the estimate for β_{DD} with that for λ allows us to benchmark the effect of comprehensive schooling.

The DD model improves on a simple cross-sectional comparison in that it removes differences in achievement *levels* between states that would have been observed even in the absence of differences in tracking practices. Yet, unobserved heterogeneity that would have caused differential competence *growth* between the Comprehensive and Tracked states, even in the absence of differences in tracking, might remain. For example, schooling inputs in grades 5 and 6 could differ systematically between the two state groups. Econometrically, this would mean that the error term in equation (1) contains state-grade-specific shocks that correlate with a state's tracking regime: $u_{isg} = v_{sg} + e_{isg}$ and $\mathbb{E}[Compr_{sg}\, v_{sg}] \neq 0$, such that the OLS estimate of β_{DD} would be biased.

Importantly, in the current setting we can directly examine whether such correlated shocks might play a role. As explained in the previous section, the distinction between the Comprehensive and Tracked States is only meaningful for students of the non-academic tracks, who in grades 5 and 6 are either taught comprehensively or further tracked. For academic-track students, in contrast, there is no difference between the two regimes, as they enter *Gymnasium* schools after grade 4 no matter the state. Under the assumption that the selection into the academic track does not differ between Tracked and Comprehensive states, they can, therefore, be used as a control group to correct for potential regime-specific trends in achievement that the DD model does not pick up.[20]

The 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 and hence adds the subscript *t* for track (*academic* versus *non-academic*):

$$Y_{istg} = v_{sg} + \phi_{tg} + \psi_{st} + \beta_{DDD} Compr_{stg} + e_{istg} \tag{2}$$

v_{sg}, ϕ_{tg} and ψ_{st} are state-grade, track-grade and state-track fixed effects, respectively. The treatment indicator $Compr_{stg}$ takes value one for grade 7 observations of non-academic-track students in the Comprehensive states. The DDD estimate β_{DDD} measures how non-academic-track students progress differently in the first two years of secondary school depending on the tracking regime net of state-specific achievement trends as approximated by academic-track

[19] A robustness check will recalculate standard errors using a wild cluster bootstrap to ensure that inference is robust to the few clusters problem inherent to my setting.

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

students. If our estimates for β_{DDD} and β_{DD} are roughly identical this indicates that achievement growth in the academic track is roughly the same between Tracked and Comprehensive states. This should increase our confidence that there are no state-specific trends confounding the DD estimate from above and that the assumptions for DD to be interpreted causally hold. If the two estimates differ, then there appear to be diverging achievement trends between the two state groups. In that case, the causal interpretation of the DDD estimate would hinge on the assumption that achievement trends of academic-track students provide an good approximation of non-academic-track students' counterfactual achievement trends.

The DD and DDD models estimate the mean effect of comprehensive versus tracked schooling. Yet, at the heart of the debate about tracking lie concerns about its distributional consequences and the question regarding winners and losers. Mean effects could easily mask very heterogeneous effects in the population. In order to answer the question as to how tracking changes the achievement distribution, we can extend the logic of the DD estimator and, instead of limiting attention to the mean, inspect how the whole achievement distribution changes differently between Comprehensive and Tracked states from grades 5 to 7.[21] Let $f_g^C(Y)$ be the density of non-academic-track students' grade g test scores (standardised to mean zero and variance one within grade level g) for the Comprehensive states. The difference $f_7^C(Y) - f_5^C(Y)$ measures the change in the density between grades 5 and 7 at each point Y for this group. $f_7^T(Y) - f_5^T(Y)$ is the equivalent difference for the Tracked states. Comparing these two quantities gives a DD estimator of the effect of comprehensive versus tracked schooling on the density at each (standardised) test score level Y:

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

The four densities in this expression are estimated non-parametrically at 100 equally spaced points for Y using a standard Kernel estimator to give a complete picture of how the achievement distribution changes due to tracking.

Finally, I explore the question of who wins and who loses from tracking by estimating treatment effect heterogeneities by previous achievement. For this, I exploit the panel nature of my data, which allows me to rewrite the DD model of equation (1) into 'gain-score' or first-differenced (FD) form:[22]

$$\Delta Y_{is} = Y_{is7} - Y_{is5} = \lambda + \beta_{DD} Compr_{s7} + \Delta u_{is} \tag{4}$$

[21] Neumark et al. (2005) proposed this method to estimate the effect of minimum wages on the distribution of family income.

[22] Note that repeated cross-sections suffice to estimate the DD model. The panel-based FD estimator and the repeated cross-sections-based DD estimator are numerically identical in a balanced panel but deviate when there is attrition (Lechner et al., 2016). In particular, Lechner et al. (2016) show that (under the assumption of constant treatment effects) DD might be inconsistent while FD remains consistent when panel non-response is non-random. As there is attrition in my data, comparing the FD and DD estimates serves as a robustness check for potential bias in the DD estimates stemming from selective panel attrition.

The FD formulation makes it easy to check if and how the effect of comprehensive schooling depends on students' position in the pre-tracking achievement distribution. Specifically, we can allow β_{DD} to vary by quartile Q_q ($q = 1, \dots, 4$) of the grade 5 test score distribution:[23]

$$\Delta Y_{is} = (\lambda + \kappa_1) + \sum_{q=1}^{4} \mathbb{1}[i \in Q_q]\, \beta_{DD}^{q} Compr_{s7} + \sum_{q=2}^{4} \mathbb{1}[i \in Q_q] \kappa_q + \Delta u_{is} \tag{5}$$

I include a separate intercept κ_q for each quartile in order to separate treatment effect heterogeneity from general heterogeneity in test score growth between different performance groups.[24] Clearly, the choice of four groups is arbitrary. Hence, in robustness checks I split the pre-tracking achievement distribution into different numbers of groups.

4 Data Description

4.1 Analysis Samples

The empirical analysis is based on data from Starting Cohort 3 of the German National Educational Panel Study (NEPS) (Blossfeld et al., 2011). This survey is a random sample of fifth graders in Germany in fall 2010, who are followed throughout their school career through annually conducted waves of surveys and tests. Teacher, principle and parent questionnaires provide additional background information. Sampling followed a two-stage process, with schools as primary sampling units.

The analysis is restricted to students with non-missing test scores in regular schools in one of the 12 states under investigation. While the NEPS is, in principle, a panel data set, repeated cross-sectional data suffice for estimating the main DD and DDD models defined in equations (1) and (2). Hence, in order to maximise sample sizes, I do not restrict the estimation sample to students who are observed repeatedly across grades, but also include students who are observed only once. This includes students who dropped out of the survey between waves[25] and students who were drawn as part of a large refreshment sample in grade 7 to compensate for attrition[26]. The main estimation sample for the DD model, which includes only non-academic-track students, comprises 5,019 (student × grade) observations (split roughly equally between grades 5 and 7). This sample is also used for estimating the densities in (3). The main estimation sample for the DDD model, which adds academic-track students, comprises 9,660 (student × grade)

[23] Note that this is more flexible than simply interacting the grade 5 score with the comprehensive schooling indicator. The latter would impose a monotonous and linear relationship between the effect of comprehensive schooling and ability.

[24] The first quartile Q_1 serves as the baseline category and is absorbed in the overall intercept $(\lambda + \kappa_1)$.

[25] Panel attrition between grades 5 and 7 is 16% or 694 students.

[26] The sampling design of the refreshment sample is roughly identical to that of the main sample. For details see Steinhauer and Zinn (2016). Inclusion of the refreshment sample adds about 1,600 observations to the grade 7 sample.

observations. A potential caveat of using the DD estimator with repeated cross-sections instead of a balanced panel is that the common trend assumption might fail if panel non-response correlates with the treatment (Lechner et al., 2015). Therefore, I corroborate the DD results by estimating the FD model in (4) with the smaller balanced panel sample (73% of non-academic-track students tested in grade 5 are tested again in grade 7, leading to a panel of 1,670 students). The panel sample is also used to estimate effect heterogeneities by previous achievement as defined in (5).

Most of the analyses limit attention to information on students' track, state of residence and their mathematics and reading test scores. These two outcomes are selected for two reasons. First, these are the standard measures in the literature to quantify the effectiveness of school-based education. Second, achievement tests in these two domains were administered in the first wave of the study in fall 2010 (i.e. at the beginning of the school year 2010/11), just as the sampled students entered secondary school. Hence, they can be conceptualised as a pre-tracking measure of achievement, as required by my identification strategy.[27] Thereafter, students were tested every two years in both domains, such that we have test scores for grades 5, 7 and 9. Grade 7 scores are the main outcome of interest as they measure achievement right after the two years during which students are exposed to either treatment condition. Grade 9 scores are used to assess persistence in the impact of tracking. Importantly, these test scores are vertically equated, meaning that achievement is measured on one scale across grade levels.[28] Although often overlooked in the economics literature, this is necessary to meaningfully compare achievement growth over time between students (Contini & Cugnata, 2016).[29] Scores are standardised to have zero mean and variance of one across all grades.[30] The descriptive statistics in Table 1 show that on average students progress about half a test score standard deviation per two years of schooling in both maths and reading.

Further, Table 1 shows that about 80% of the sample comes from the Tracked states. This simply reflects the fact that these states are larger and more populous (see also the map in Figure 2). Other than that, the samples seem reasonably balanced in terms of composition and initial achievement. The average age at the time of the first test is 10.8 years in both state groups and about 48% of both samples are female. However, as expected, the Comprehensive states – composed mostly of the poorer East German states and city states – score slightly worse on socio-economic variables like household income, parental education and unemployment. Moreover, the share of students with a migration background is much lower in these states (17% vs. 27%), mainly reflecting the different migration histories of West and East Germany.

[27] In fact, students have been in secondary school, and hence the tracking system, for about 3 months at the time of the test. Note that this biases the estimates of the impact of comprehensive schooling slightly toward zero.

[28] The psychometric linking of test score scales across grades is achieved through the recurrence of certain anchor items in every wave of the competence test. For details on the design and linking of NEPS competence tests see Fischer et al. (2016).

[29] See Lang (2010) for an elaborate discussion on test score scaling and common pitfalls.

[30] Merely for the density DD estimator defined in (3), test scores are standardised within grade levels.

These differences highlight that simple cross-sectional comparisons between the two tracking regimes are probably not very revealing.

4.2 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 the selection into the academic track does not differ between the two tracking regimes. Otherwise, neither the academic-track nor the non-academic-track student bodies would be comparable, invalidating both the DD and the DDD estimator. Administrative records show that 40% and 43% of the cohort in question attend the academic track in the Tracked and Comprehensive states, respectively (Statistisches Bundesamt, 2016). Table 1 reveals that academic-track students are slightly oversampled in the NEPS. Reassuringly, however, both in the population and in my sample, these shares are very close between Comprehensive and Tracked states.

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 form necessarily comprises all low-achievers. This might amplify average ability differences between academic and non-academic tracks in two-tiered versus three-tiered systems. In order to test for the presence of such differences in selection, Figure 3 plots the pre-tracking (grade 5) test score distributions both overall and for academic and non-academic-track students separately by tracking regime. In both maths and reading, the differences between academic- and non-academic-track students are virtually identical in Tracked and Comprehensive states. This is confirmed by Kolmogorov–Smirnov tests.[31] Therefore, I conclude that there are no differences in the selection into the academic track between the two state groups and confidently compare their non-academic-track students' progress while using academic-track students to control for state-specific achievement trends.[32]

4.3 Distribution over Non-Academic Tracks

Figure 4 displays the distribution of non-academic-track students over different school forms by tracking regime. In the Tracked states one third of students attend a low-track school (HS) and about half attend middle-track schools (RS). In the counterfactual scenario of a two-tiered

[31] I implement 6 Kolmogorov–Smirnov tests for differences in the test score distributions of non-academic, academic and all students between Tracked and Comprehensive states for both maths and reading scores. All but one (academic-track reading; $p = 0.07$) fail to be rejected at the 10% level.

[32] The fact that the alternative choice options appear largely irrelevant for the selection into the academic track might seem puzzling at first. However, it can be explained by the special status that the academic-track *Gymnasium* holds 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 & Blossfeld, 2007). In my sample, 78% of students with college-educated parents attend the academic track.

tracking regime, 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 (14%) attends comprehensive schools where all three degrees can be obtained and there might or might not be within-school streaming. In the language of the treatment effects literature, these students can be thought of as 'always-takers' (there are no 'never-takers' or 'defiers' by construction). This introduces a slight downward bias in the estimates of the effect of comprehensive versus tracked schooling on achievement. In the Comprehensive states, there are no low- and middle-track schools. The vast majority of non-academic-track students in these states (74%) attends a SMT school. The remaining 26% attend a comprehensive school (IGS).[33] As previously mentioned, there is no within-school streaming in grades 5 and 6 in any of these schools.[34] It is the effect of this comprehensive schooling for non-academic track students in the two-tiered regime, as compared to the further tracking of this group in the three tiered-regime, that this paper aims to estimate.

5 Results

5.1 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. Before turning to the estimation results of the DD and DDD models derived above, for illustrative purposes, I begin by comparing students' progress between Tracked and Comprehensive states graphically. Figure 5 presents kernel density estimates of the distributions of grade 5 to 7 gain-scores, $\Delta_7 Y_{is} = Y_{is7} - Y_{is5}$, for academic- and non-academic-track students in both tracking regimes. The picture is quite striking: whereas in both maths and reading academic-track students' progress is very similar between Tracked and Comprehensive states – if anything, those in the Tracked states learn slightly more – for non-academic-track students the distribution of gains in the Comprehensive states appears almost monotonically shifted to the right as compared to that in 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 DD model from equation (1) for maths (Panel A) and reading (Panel B). This double-differences estimator compares achieve-

[33] These sample shares are fairly close to the true population shares, which are as follows: 32% HS, 51% RS and 17% IGS in Tracked states; 63% SMT and 37% IGS in Comprehensive states (Statistisches Bundesamt, 2016).

[34] As also previously mentioned, from grade 7 onwards some of these schools form track-specific classes. Unfortunately, the NEPS data does not provide information on which model of ability grouping schools follow from grade 7 onwards.

ment changes between grades 5 and 7 of non-academic-track students in the Comprehensive states with the corresponding changes of students in the Tracked states. Through the inclusion of state fixed effects, which absorb pre-tracking differences in achievement levels between states, the DD model controls for time-constant heterogeneity between states. The coefficient estimates for the grade 7 dummy imply that during the first two years of secondary school in the Tracked states students improve by about half a standard deviation (SD) in both subjects. In the Comprehensive states, students progress about 0.17 SD more in maths and 0.23 SD more in reading. Hence, these results confirm the finding from the graphical comparison above, suggesting positive effects on achievement from teaching students comprehensively in early secondary school.

Yet, as explained before, we cannot be sure that controlling for level differences in pre-tracking achievement between tracking regimes removes all confounding factors that might bias the comprehensive schooling estimate. Perhaps different student body compositions would have caused differential achievement growth in Tracked and Comprehensive states also in the absence of differences in tracking practices. To control for potential state-grade-specific achievement shocks that would violate the assumption of parallel counterfactual achievement trends, the DDD model from equation (2) adds academic-track students as an additional control group. Reassuringly, the triple-differences results, presented in Column (2), are very similar to those from before: for maths the effect of comprehensive schooling increases marginally to 20%, while for reading it remains virtually constant. The similarity of the DD and DDD estimates indicates that there are no divergent achievement trends between the Comprehensive and Tracked states, thus corroborating the causal interpretation of the estimates. Consequently, most analyses in the remainder of the paper restrict attention to the sample of non-academic-track students.

In summary, the results from Table 2 suggest substantial positive effects of comprehensive schooling on student achievement. In fact, the point estimates translate to 30% higher achievement growth between grades 5 and 7 in maths and close to 50% in reading. Note, however, that there is considerable uncertainty around these estimates (see standard errors in Column 1) due to the relatively small sample sizes in the Comprehensive states. Accordingly, one should exercise some caution in interpreting the point estimates at face value. Nevertheless, the fact that the coefficients for both subjects are significant and positive strongly rejects tracking proponents' claim that comprehensive schooling impedes achievement.[35] As, by definition, 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, which would have predicted a negative effect. For this pre-selected group of non-academic-track students, comprising low achievers and, as is visible from the pre-tracking achievement distributions displayed in Figure 3, also substantial shares of medium to high achievers, studying together for an additional two years generates achievement gains.

[35] The significance of the results also holds when using wild cluster bootstrap instead of cluster-robust standard errors to calculate significance levels. See section 6.

5.2 Distributional Consequences

The debate on tracking revolves around questions of equity. Hence, it is crucial to go beyond mean effects and explore the distributional consequences of tracked versus comprehensive schooling. To this end, this section describes how the shape of achievement distribution changed differently between grades 5 and 7 in the Comprehensive versus the Tracked states. For this exercise, test scores are standardised per grade level to ensure that the distribution of grade 5 scores is comparable to that of grade 7.

Panel A of Figure 6 presents non-parametric density estimates of non-academic-track students' grade 5 and grade 7 maths scores for the Comprehensive states. Panel B does the same for the Tracked states. In the Tracked states the distribution appears to have stayed relatively constant between grade levels, whereas in the Comprehensive states it appears to have tightened slightly. As the differences between the densities are small relative to the scale, Panel C summarises the information from the upper two panels into one graph by plotting the vertical distance between the grade 5 and 7 densities for the two state groups.[36] Thus, these lines describe how the shape of their maths score distributions changed between grades. The vertical distance between these two lines, displayed in Panel D, is the DD estimate of the effect of comprehensive versus tracked schooling at each point of the standardised maths score distribution. A clear picture emerges: comprehensive schooling seems to shift probability mass from the bottom end of the distribution (approximately from the range [-3, -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 A.1). Therefore, next to the positive average effect established before, comprehensive schooling appears to have an equalising effect on the achievement distribution.

5.3 Effect Heterogeneities

The fact that I find a positive average effect and an equalising effect on the overall achievement distribution does not tell us who is driving these effects. It might very well be that certain groups of students lose out from not being tracked, but that these are losses are compensated by the gains of other groups. Therefore, this section explores treatment effect heterogeneities by previous achievement and socio-economic status (SES).

The analysis in this section is based on the FD or gain-score model defined in equation (4) and, hence, on the smaller panel sample (of non-academic-track students). In a balanced panel, FD is equivalent to DD, but they differ when there is panel attrition as DD also uses students who are observed only once. Hence, comparing the FD and DD estimates allows us to gauge the severity of bias in the DD estimates due to non-random panel non-response (Lechner et al., 2015). Note that this issue should be partially alleviated by the fact that the repeated cross-

[36] More precisely, the figure plots the difference of the non-parametric maths score density estimates at 100 equally spaced points between the minimum and maximum value of the standardised scores.

sectional DD sample includes students from the randomly drawn (and, hence, representative) grade 7 refreshment sample. However, under treatment effect heterogeneity there is a second order effect. Theory and previous empirical evidence suggest the hypothesis that treatment effects of comprehensive schooling are larger for low-ability students. If those students are more likely to drop out between waves, panel-based FD might underestimate the true average treatment effect. Indeed, attrition is much higher for students in the bottom quartile of the grade 5 achievement distribution (44% compared to 27% for all non-academic-track students) and for low-track students (in the Tracked states, 42% of low-track students, compared to 29% for all non-academic-track students, drop out between waves).

Columns (1) and (4) of Table 3 present the FD estimates for maths and reading, respectively. Reassuringly, the effect estimates of 0.14 SD (maths) and 0.21 SD (reading) are very close to the results from above. However, as to be expected under the hypothesised form of treatment effect heterogeneity together with selective attrition, they are somewhat smaller than the DD estimates. The following regressions investigate effect heterogeneity by previous achievement explicitly. In particular, I allow the effect of comprehensive schooling to vary by quartile of the pre-tracking achievement distribution by interacting the comprehensive schooling indicator with respective dummies as demonstrated in equation (5). To the extent that grade 5 achievement captures ability, the achievement quartiles can be thought of as representing ability groups.

The results in columns (2) and (5) reveal a steep gradient in the effect of comprehensive schooling with respect to previous achievement. This is most pronounced for maths where the effect is strictly decreasing with previous achievement. The estimate of 0.3 SD for the lowest group is more than double the size of the average effect. For the middle two groups the effects are positive but insignificant (the point estimates are 0.12 and 0.1 SD, respectively), whereas for the the highest group we get a fairly precise zero. Large and significant coefficients of 0.27 SD for the low and 0.19 SD for the mid-low groups indicate that also in reading low-achieving students benefit from comprehensive schooling. Here, the gradient is somewhat less pronounced, given that the point estimates also remain positive for the upper two groups. These are, however, imprecisely estimated and insignificantly different from zero.

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. Importantly, we do not find a negative effect for any of the achievement groups, meaning that higher achievers do not seem to lose out from learning together with their lower achieving peers.[37]

Lastly, given that much research on tracking focuses on socio-economic inequalities, columns (3) and (6) check for effect heterogeneity by SES. The dummy variable 'Low SES' indicates that

[37] These results are confirmed when investigating effect heterogeneity by *terciles* (instead of quartiles) of the grade 5 achievement distribution (see Appendix Table A.1).

a student reported having less than the median number of books at home (for the non-academic-track sample).[38] It is interacted with the comprehensive schooling indicator (and added to the regression) analogously to the previous achievement dummies before. The results reported in columns (3) and (6) show little interaction between the effect of tracking and students' socio-economic background. This is at odds with previous studies that find that especially low-SES students benefit from comprehensive schooling (e.g. Kerr et al., 2013). My result is likely explained by the fact that non-academic-track students are a negatively pre-selected group, where 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. The socio-economic dividing line runs more between the academic- and the non-academic tracks, rather than between different school forms of 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. In turn, it is unsurprising that SES-inequalities are not significantly affected by comprehensive schooling.

5.4 Persistence of Effects

This section presents the results for grade 9 outcomes – the grade level after which students can leave school with a low-track degree (conditional on obtaining the required grades). Remember that many SMT and comprehensive schools in the Comprehensive states start using within-school streaming starting from grade 7 onwards. Hence, differences in tracking between the two state groups diminish from that grade level onwards. The comparability of school systems between states decreases more generally with grade level as schooling policies are less harmonised for higher grades (KMK, 2014). Accordingly, the purpose of this section is mainly to obtain a rough idea of the persistence of effects, while acknowledging that effect estimates might be contaminated by other factors.

Table 4 presents the results for the FD model for grade 5 to 9 gain-scores, $\Delta_9 Y_{is} = Y_{is9} - Y_{is5}$, in analogous fashion to Table 3. Due to panel attrition, the sample sizes are substantially smaller than before, which reduces precision in the estimates. The average effects displayed in columns (1) and (4) continue to show an advantage for those students taught comprehensively in grades 5 and 6, but they lose their significance. Point estimates are about half the size than those for grade 7, indicating gradual fade-out of the large average short-run effects.[39] However, inspection of

[38] Many background variables from the parent questionnaire measure SES more directly than the student reported number of books at home. However, all of these variables have a high portions of missing values (about 25%). To avoid reducing the sample size any further, I refer to this variable from the student questionnaire to proxy for SES. The conclusions are the same when using other SES proxies in smaller samples. Results for these regressions are available on request.

[39] 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. Using the larger cross-sectional sample and estimating the grade 9 level effects by DD confirms the fade-out of mean differences. As results between DD and FD do not differ, the DD results are omitted for brevity.

columns (2) and (5) reveals that the small and insignificant average effects mask large and persistent effects of comprehensive schooling for the lower ability groups. Again, this is most clear for maths where the very large short-run effect for the bottom achievement group is now halved to 0.16 SD, but remains highly significant – both statistically and economically. The coefficient for the mid-low group is positive but insignificant and the coefficients for the upper half of the ability distribution are very imprecisely estimated but closer to zero. Similarly, for reading we find two sizeable positive coefficients (one of which is significant) in the bottom half and insignificant coefficients close to zero for the top half of the achievement distribution. The SES interactions in columns (3) and (6) remain insignificant.

5.5 Mechanisms

The effect of tracking, or conversely comprehensive schooling, on achievement might operate through various channels. As explained above, the three most important channels in this context are peer effects, academic standards and motivational consequences. I cannot discriminate between these with the data at hand and all three are likely to play a role. Nevertheless, regarding peer effects, the results indicate that classroom heterogeneity might be less important than commonly assumed; otherwise a positive effect of comprehensive schooling could not be explained. The fact that benefits for low-achievers are large while high-achievers are unaffected by comprehensive schooling suggests non-linear peer effects: contact with higher achieving peers appears more beneficial than contact with lower achieving ones harmful. Given that curricular differences between low- and middle-track schools are relatively small in the first two grades of secondary school, it is unlikely that academic standards are the primary driver behind the results. However, the negative image of low-track schools in Germany might both discourage students assigned to these schools and negatively influence teacher expectations about their potential. Such stigma, negatively selected peers and slightly less demanding standards might present a serious mix of obstacles to achievement that explains the large advantage of low-achievers in the Comprehensive states.

These dynamics should find expression in differences in socio-emotional outcomes between Tracked and Comprehensive states. To explore these issues Table 5 presents evidence on differences in students' educational aspirations, their school-related motivation and feelings of helplessness in school. Educational aspirations were measured in grades 5 and 7, allowing for implementation of the difference-in-differences design. Given the categorical nature of this variable, I construct a dummy that indicates that a student aspires higher than the low-track school-leaving certificate. Column (1) indicates that in the Tracked states the share of students aspiring higher decreases by about 2 percentage points between grades, whereas in the Comprehensive states it increases by roughly 6 percentage points. Indeed, comprehensive schooling reduces the share of students with low educational aspirations. The other variables are measured in grade 7 only. Consequently, for these outcomes I revert to OLS regressions and large control sets to approximate the effect of comprehensive schooling.[40] Accordingly, I cannot fully rule out that the estimates are confounded by differences between states unrelated to tracking. With these caveats in mind and despite their limited significance, the results in columns (2) to (5) seem to suggest that the positive effect of comprehensive schooling on cognitive outcomes goes hand in hand with improved socio-emotional outcomes: students taught comprehensively are less helpless (at least in maths) and more motivated (at least in German). Although far from conclusive, these results are consistent with motivational consequences of tracking being important.

6 Robustness Checks

This section probes the robustness of my results by performing an array of additional tests. In particular, I consider (i) potential threats to identification and (ii) potential problems for inference.

6.1 Threats to Identification

The difference-in-differences model controls for (unobserved) factors influencing student achievement that have become manifest before the start of secondary school. Thus, it eliminates all time-constant between-state achievement differences. Consequently, causal interpretation of the comprehensive schooling effect relies on the assumption that, in the absence of differences in tracking, grade 5 to 7 achievement growth of non-academic-track students would have been the same in Tracked and Comprehensive states. Evidence favouring of the plausibility of this assumption has been presented in form of the triple-differences model, which confirms that, in the academic track, achievement growth is indeed parallel. Though reassuring, we do not have certainty that these parallel trends carry over to the non-academic tracks.

A further natural test of this assumption is to compare ‘pre-treatment’ achievement trends between Comprehensive and Tracked states. These are the trends in primary school, before tracking has started. Unfortunately, the NEPS Starting Cohort 3 panel, used for the analysis so far, commenced in fifth grade, meaning that we have no information on these students’ achievement growth in primary school. Hence, I revert to data from NEPS Starting Cohort 2, which is a random sample of first graders in 2013 followed throughout primary school.[41] Given that there were no major changes to primary education in Germany in this time period, the achievement trends of this cohort should be roughly similar to those of the cohort used

[40] Control variables include: student age, sex, migration background, number of siblings, parental years of education, household income, unemployment and single parent household. In order to retain the full sample, I add missing dummies for each control variable and replace missing values with variable means.

[41] The cohort of the main sample attended first grade in the school year 2006/07.

in the main analysis. With the primary school sample, I estimate the DD model defined in equation (1) using grade 2 and grade 4 maths scores (instead of grade 5 and 7 scores) to test if divergent achievement trends can already be detected in the two years prior to tracking.[42] Table 5 presents the estimation results. Reassuringly, the coefficient on the interaction between the Comprehensive state dummy and the grade 4 dummy is very close to zero, indicating parallel pre-treatment trends between Tracked and Comprehensive states.

Parallel trends in primary school and the academic track do not rule out the possibility of co-treatments. Perhaps home and school inputs received by non-academic-track students in the first two years of secondary school differ systematically between the two regimes. If so, these could be the real reason behind divergent achievement growth. We know that there are socio-economic differences between states and perhaps there are SES differences in parental investment into their children in grades 5 and 6. Further, even though curricula in lower secondary school are harmonised between states, it is conceivable that SMT and comprehensive schools differ systematically from low- and middle-track schools in terms of the school inputs they provide.

In order to test for such potential confounders, I draw on ample background data provided in the NEPS. In particular, I cumulatively add (i) basic individual controls from the student questionnaire (age, sex, migration background, number of siblings), (ii) SES controls from the parent questionnaire (parental years of education, household income, unemployment, single parent household) and (iii) class-level controls from the teacher and principal questionnaires (teacher experience, teacher further education, class size) to the FD model defined in equation (4).[43] If the results are really being driven by systematically different home or schooling inputs between Tracked and Comprehensive states then the estimated effect of comprehensive schooling should drop upon the inclusion of these controls. Tables 6 shows the results for maths (Panel A) and reading (Panel B). The row with values for the R^2 indicates that these variables add explanatory power to the model. Nevertheless, in both subjects the effect estimate remains almost completely stable throughout all four specifications. As any other home or schooling inputs that are not included here will most certainly correlate with those I do observe, the fact that there is no substantial drop in the effect estimates is strong evidence for the robustness of my results.

6.2 Threats to Inference

It is well known that clustered data can cause problems for inference whenever there is intra-cluster correlation of errors and/or independent variables (MacKinnon & Webb, 2017). My data has a clear group-level structure: students are nested in classrooms, which are nested in schools, which are nested in states. The primary sampling unit of the NEPS is the school. Hence, for correct inference the correlation between errors of students of the same school is the

[42] The DD model cannot be estimated for reading scores as these are only available for grade 4.

[43] To not reduce the sample size I add missing dummies for each control variable and replace missing values with variable means.

minimum that needs to be taken into account in the calculation of standard errors. As treatment assignment depends on the state (and its interaction with track), I followed standard practice for conservative inference and clustered standard errors at the state level in all analyses, allowing for arbitrary within-state across-grade dependence of errors (Bertrand et al., 2004; Cameron & Miller, 2015).

However, the asymptotics of the standard cluster robust variance estimator (CRVE) are based on the number of clusters going to infinity while my analyses only include 12 states. Finite sample distributions might only poorly approximate asymptotic ones if the number of clusters is small. On top of this, states differ widely in size meaning that, potentially problematically, clusters are also quite heterogeneous in my case (see e.g. Carter, Schnepel & Steigerwald, 2017). Hence, in order to ensure that the confidence of my results has not been overstated despite clustering at the state level, Table 6 compares CRVE- with wild cluster bootstrap-based p-values for the DD and DDD models for maths and reading (Cameron et al., 2008).[44] Both methods lead to very similar results, indicating that inference is robust.

Lastly, I address the concern that my results are driven by particular outlier states, whose performance diverged extremely from the others, instead of reflecting a causal effect of comprehensive schooling. To this end, I perform a simple leave-one-out analysis, dropping one state at a time and re-estimating the DD and DDD models. Figure 7 plots the accordant distribution of the point estimates for the effect of comprehensive schooling with confidence intervals. The significance of the results is slightly affected when some larger states are dropped but, in general, the main effects are robust to the exclusion of any particular state.[45]

7 Discussion & Conclusion

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

[44] The wild cluster bootstrap proposed by Cameron et al. (2008) 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 & Webb, 2017) or a schools-as-'sub-clusters'-of-states (MacKinnon & Webb, 2016) version of the bootstrap.

[45] Due to data confidentiality the names of the federal states cannot be revealed.

peer effects. Moreover, potential error and bias in track placement could attenuate theoretical positive effects of tracking in practice. Identifying these effects is notoriously difficult due to the severity of the selection problems involved.

I circumvent such issues of endogeneity by exploiting 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, however, the remaining student body is either divided between low- and middle-track schools or taught comprehensively for another two years. I estimate the effects of these two years of comprehensive instead of tracked schooling on achievement in a difference-in-difference-in-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.

The effect of comprehensive versus tracked schooling in grades 5 and 6 is estimated to be 0.17 SD in mathematics and 0.23 SD in reading. While there is some non-negligible uncertainty around these average effects they are clearly non-negative. The heterogeneity analysis reveals that it is the lower tail of initial achievement distribution that drives these average effects. For them the effects are precisely estimated and large. Large effects for low-achievers and zero effects for high-achievers suggest a pronounced ability gradient in the effect of comprehensive schooling. It has an equalising effect on the distribution of test scores without trading off efficiency against equity. The effects for the bottom-end are persistent: even though many schools use within-school streaming from grade 7 onwards, in grade 9 – towards the end of secondary school for those that leave school with the lowest degree – these students are still considerably better off. Auxiliary analyses suggest that students' school-related motivation and educational aspirations are higher in the comprehensive system.

The fact that for low-achievers the disadvantage from early tracking is large and persistent is likely due to the fact that this group concentrates in low-track schools, which, following the negative selection, become environments of low standards, expectations, motivation and performance. Unsurprisingly, in Germany, these schools have become increasingly stigmatised. If, instead, these students are integrated into schools with higher achieving peers and exposed to higher scholastic demands, their achievement improves substantially without impairing that of their peers. Accordingly, the reforms of several West German states to abolish low-track schools and establish two-tiered school systems should be applauded from an efficiency and an equity standpoint.

The German school system being exemplary for the early between-school tracking systems that are found across Europe (e.g. Austria, Belgium, Hungary, Luxembourg, Netherlands) (Bol & van de Werfhorst, 2013), these results carry direct relevance beyond the German context. Globally, many governments seek to reduce the share of low achieving school-leavers that their school systems produce. My results highlight that vertical differentiation, especially at such early ages, impedes on this goal as it depresses performance at the bottom.

Nevertheless, one should be careful with extrapolating from these results to the effects of fully comprehensive school systems as the variation in tracking practices I exploit only concerns a negatively preselected group of students. We can only speculate about the effects of comprehensive schooling would be for the best students on academic-track schools. The fact that I find a negative ability gradient suggests that the effect could eventually turn negative – especially as students become older. There is natural appeal to the idea of creating homogeneous classrooms through some form of tracking at one point of secondary schooling. However, the fact that even for the best 25% of non-academic-track students – who would be medium-high achievers even in academic-track schools (see Figure 3) – I do not find negative effects suggests that the costs to classroom heterogeneity might be smaller than commonly assumed. The evidence presented in this paper, therefore, highlights that homogeneity in terms of ability is only part of the story. Peer effects, motivational consequences of tracking and curricular demands are also important determinants of student achievement; otherwise these findings could not be explained. 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 differentiation starts.

Finally, note that the literature on *within-school streaming* mostly finds positive effects for students selected for high-ability classrooms without negative effects for those in regular classrooms (e.g. Card & Giuliano, 2016; Duflo et al., 2011; Figlio & Page, 2002). Rather than contradicting my and previous findings on *between-school tracking*, this suggests that the costs to tracking increase convexly with the degree of vertical differentiation between tracks. It makes intuitive sense that mechanisms like peer effects, motivational factors and academic expectations are more pronounced when students are separated between schools. Combining these insights about the interplay between the grouping of students and educational production suggests that 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 policy-makers to reap efficiency gains from homogeneity without incurring large costs in terms of equity.

References

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

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

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management*, *15*(4), 623–645.

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

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

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

Betts, J. R. & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review*, *19*(1), 1–15.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (2011). *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Springer.

Bol, T. & van de Werfhorst, H. (2013). The measurement of tracking, vocational orientation, and standardization of educational systems: a comparative approach. *AIAS, GINI Discussion Paper*, *81*.

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

Brunello, G., Giannini, M., & Ariga, K. (2007). The optimal timing of school tracking: A general model with calibration for Germany. volume 1 of *Schools and the Equal Opportunity Problem* (pp. 129–156). MIT Press.

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

Cameron, A. C. & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

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

Carter, A. V., Schnepel, K. T., & Steigerwald, D. G. (2017). Asymptotic behavior of a t-test robust to cluster heterogeneity. *The Review of Economics and Statistics*, *99*(4), 698–709.

Contini, D. & Cugnata, F. (2016). Learning inequalities between primary and secondary school. Difference-in-difference with international assessments. *University of Turin 'Cognetti de Martiis' Working Paper*, *07/16*.

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

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

Dustmann, C., Puhani, P., & Schönberg, U. (2017). The long-term effects of early track choice. *The Economic Journal*, *127*(603), 1348–1380.

Edelstein, B. & Nikolai, R. (2013). Strukturwandel im Sekundarbereich. Determinanten schulpolitischer Reformprozesse in Sachsen und Hamburg. *Zeitschrift für Pädagogik*, *59*(4), 482–494.

European Commission (2014). European vacancy and recruitment report. *Directorate-General for Employment, Social Affairs and Inclusion*.

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

Fischer, L., Rohm, T., Gnambs, R., & Carstensen, C. (2016). Linking the data of the competence tests. *NEPS Survey Paper*, *1*.

Galindo-Rueda, F. & Vignoles, A. F. (2004). The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability. *IZA Discussion Paper*, *1245*.

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

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

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

Helbig, M. & Nikolai, R. (2015). *Die Unvergleichbaren: Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949*. Julius Klinkhardt.

Jackson, M. (2013). *Determined to Succeed? Performance versus Choice in Educational Attainment*. Stanford University Press.

Kerr, S. P., Pekkarinen, T., & Uusitalo, R. (2013). School tracking and development of cognitive skills. *Journal of Labor Economics*, *31*(3), 577–602.

Kultusministerkonferenz (KMK) (2014). The education system in the Federal Republic of Germany 2012/2013. In *Eurydice National Dossier*. Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany.

Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. *Journal of Economic Perspectives*, *24*(3), 167–182.

Lechner, M., Rodriguez-Planas, N., & Kranz, D. F. (2016). Difference-in-difference estimation by FE and OLS when there is panel non-response. *Journal of Applied Statistics*, *43*(11), 2044–2052.

Mackinnon, J. G. & Webb, M. D. (2016). The subcluster wild bootstrap for few (treated) clusters. *Queen's Economics Department Working Paper*, *1364*.

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

Marsh, H. & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*(1), 213–231.

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

Murphy, R. & Weinhardt, F. (2014). Top of the class: The importance of ordinal rank. *CESifo Working Paper*, *4815*.

Neumark, D., Schweitzer, M., & Wascher, W. (2005). The effects of minimum wages on the distribution of family incomes: A nonparametric analysis. *Journal of Human Resources*, *40*(4), 867–894.

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

Paulus, W. & 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*, *53*(4), 491–508.

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

Pischke, J.-S. & Manning, A. (2006). Comprehensive versus selective schooling in England and Wales: What do we know? *National Bureau of Economic Research (NBER) Working Paper*, *12176*.

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

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

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

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

Statistisches Bundesamt (2016). Fachserie 11.1 – Allgemeinbildende Schulen. Schuljahr 2015/16. Technical report, Statistisches Bundesamt, Wiesbaden.

Steinhauer, H. W. & Zinn, S. (2016). NEPS technical report for weighting: Weighting the sample of starting cohort 3 of the national educational panel study (waves 1 to 5). Technical report, Leibniz Institute for Educational Trajectories.

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

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*, *1315*.

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

Figures & Tables

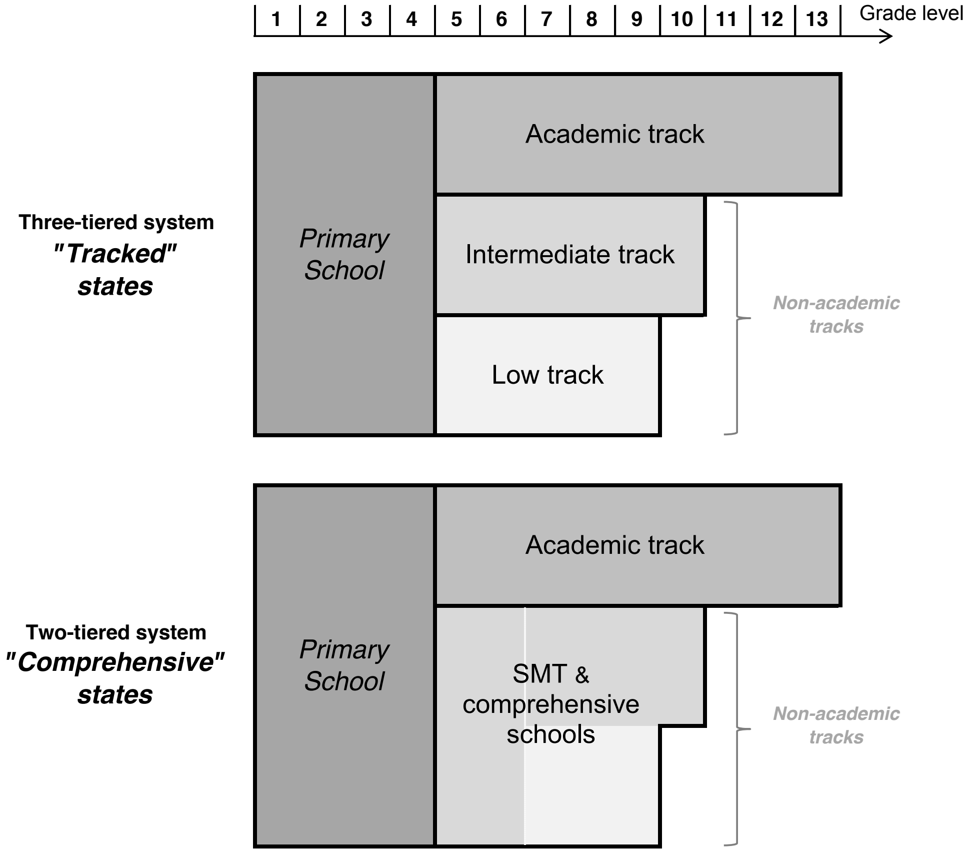

Figure 1: Schematic overview of the two tracking regimes in Germany.

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 Figure 4 for details). Academic track = *Gymnasium*, Intermediate track = *Realschule*, Low track = *Hauptschule*, School with multiple tracks (SMT) = *Schule mit mehreren Bildungsgängen*, Comprehensive schools = *Integrierte Gesamtschule*

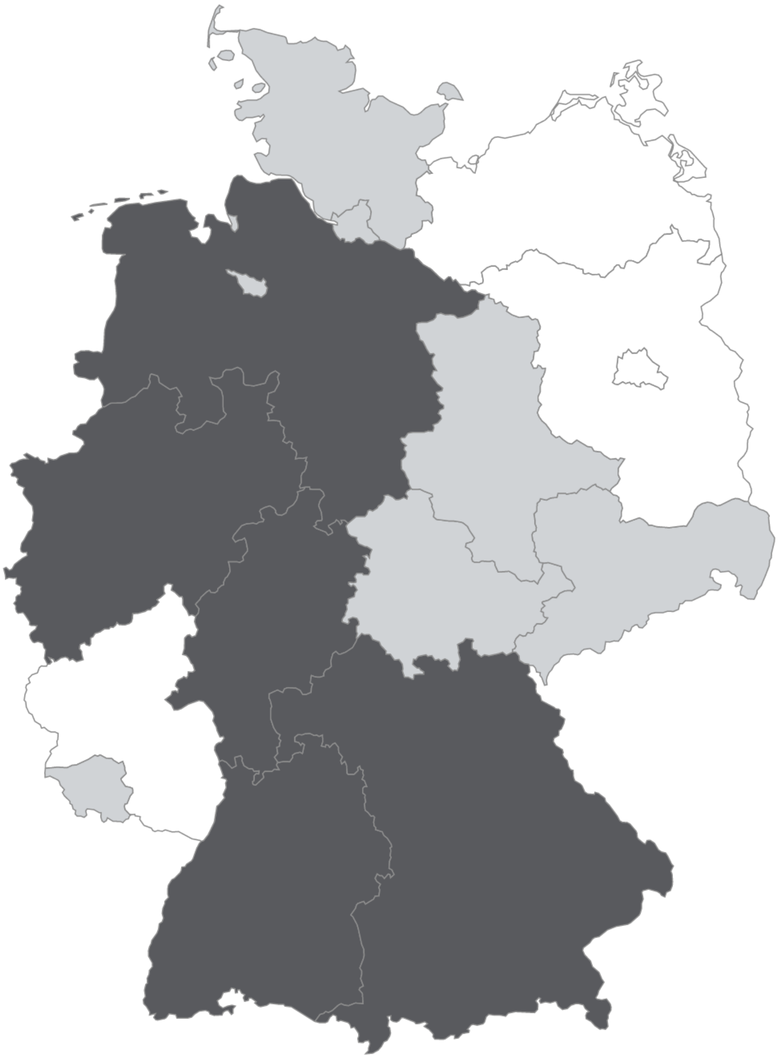

Figure 2: German federal states coloured by tracking regime.

Notes: Comprehensive states = light grey. Tracked states = black. States excluded from the analysis = white.

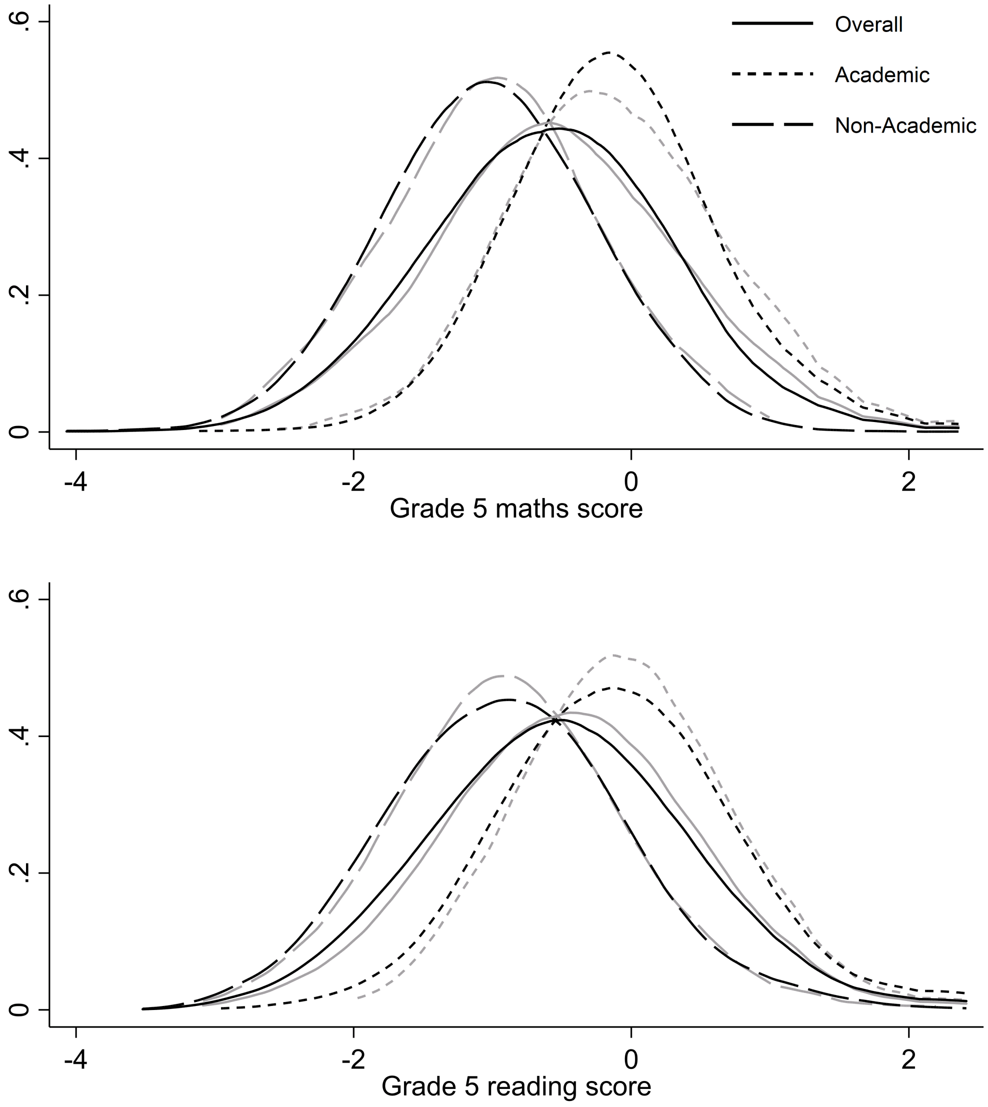

Figure 3: Pre-tracking (grade 5) distribution of test scores by track and tracking regime.

Notes: Comprehensive states = grey lines. Tracked states = black lines. Densities are estimated based on the panel sample using a Epanechnikov kernel estimator with optimal bandwidth (for normally distributed variables) (Silverman, 1986).

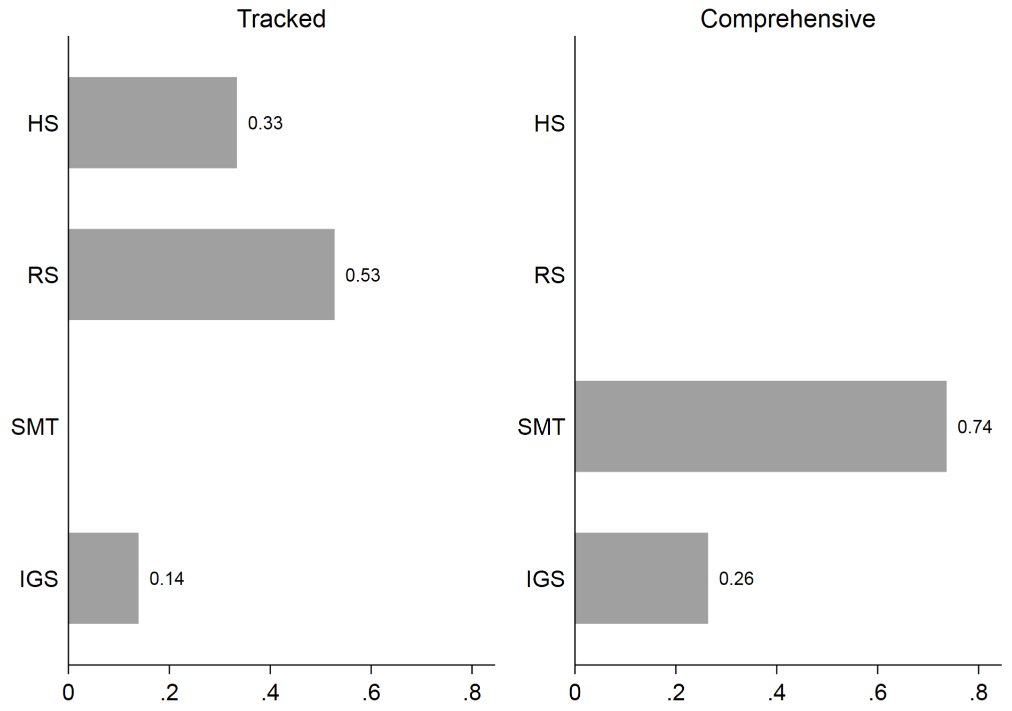

Figure 4: Grade 5 distribution of non-academic-track students over different school forms by tracking regime.

Notes: HS = low-track school (*Hauptschule*), RS = intermediate-track school (*Realschule*), SMT = school with multiple tracks (*Schule mit mehreren Bildungsgängen*), IGS = comprehensive school (*Integrierte Gesamtschule*)

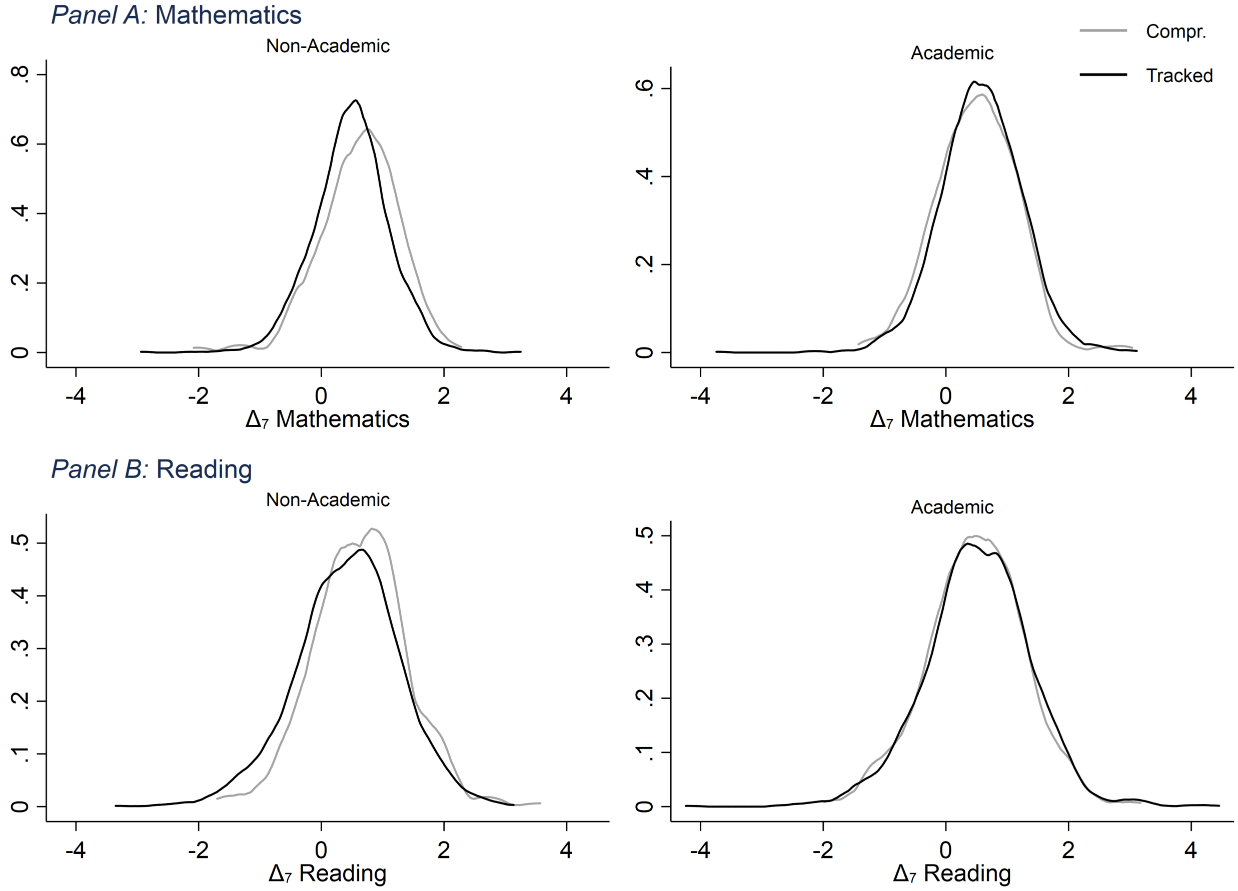

Figure 5: Distribution of gain-scores, $\Delta_7 Y_{is} = Y_{is7} - Y_{is5}$, by track and tracking regime.

Notes: Densities are estimated based on the panel sample using a Epanechnikov kernel estimator with optimal bandwidth (for normally distributed variables) (Silverman, 1986).

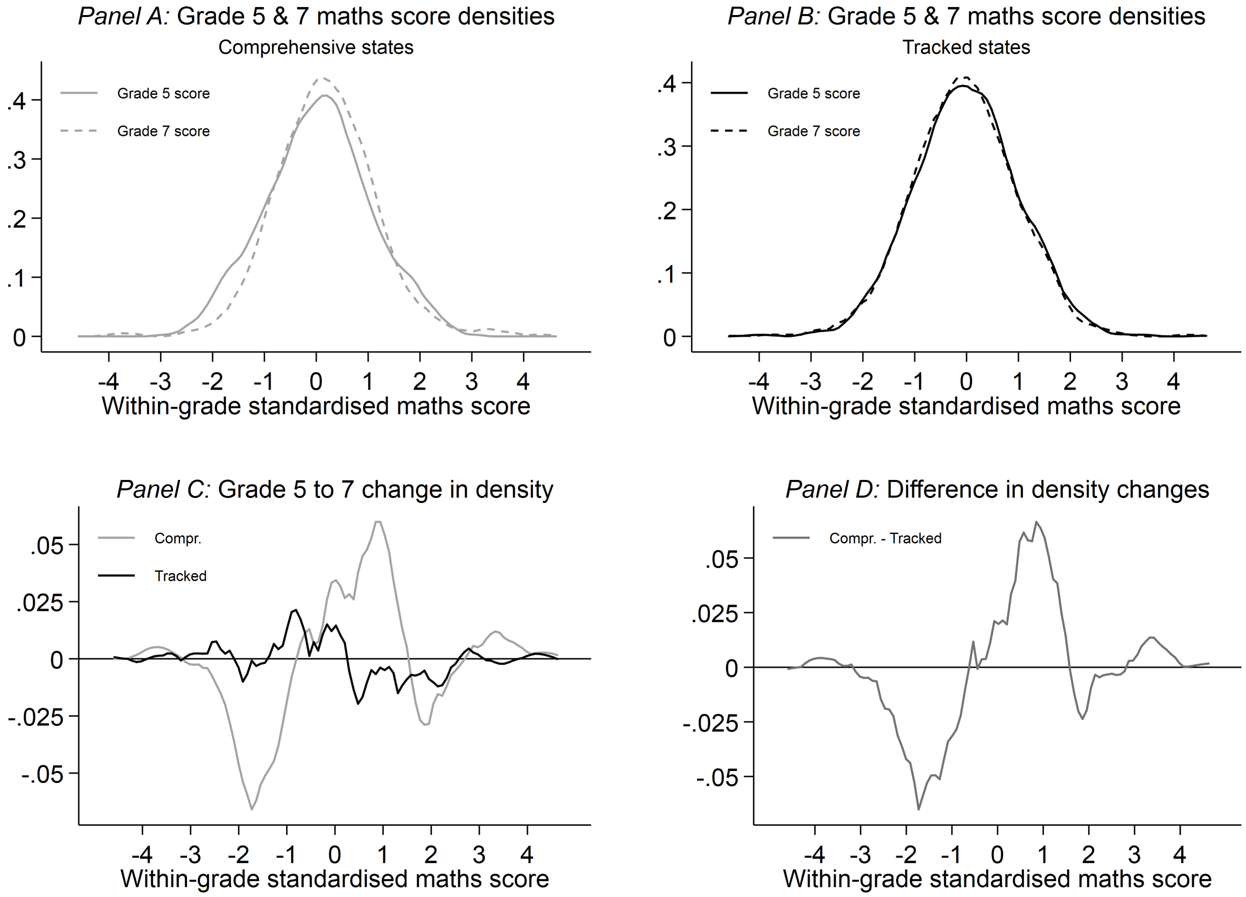

Figure 6: Changes in maths score distribution through comprehensive schooling vs. tracking.

Notes: The densities in Panel A and B are based on the cross-sectional grade 5 and grade 7 samples with within-grade standardised maths scores. Estimation uses a Epanechnikov kernel estimator with optimal bandwidth (for normally distributed variables) (Silverman, 1986). For Panel C, I calculate the difference between the non-parametric density estimates of the grade 7 and grade 5 scores at 100 equally spaced points between the minimum and maximum of the standardised maths score for Tracked and Comprehensive states. Panel D plots the difference in these differences.

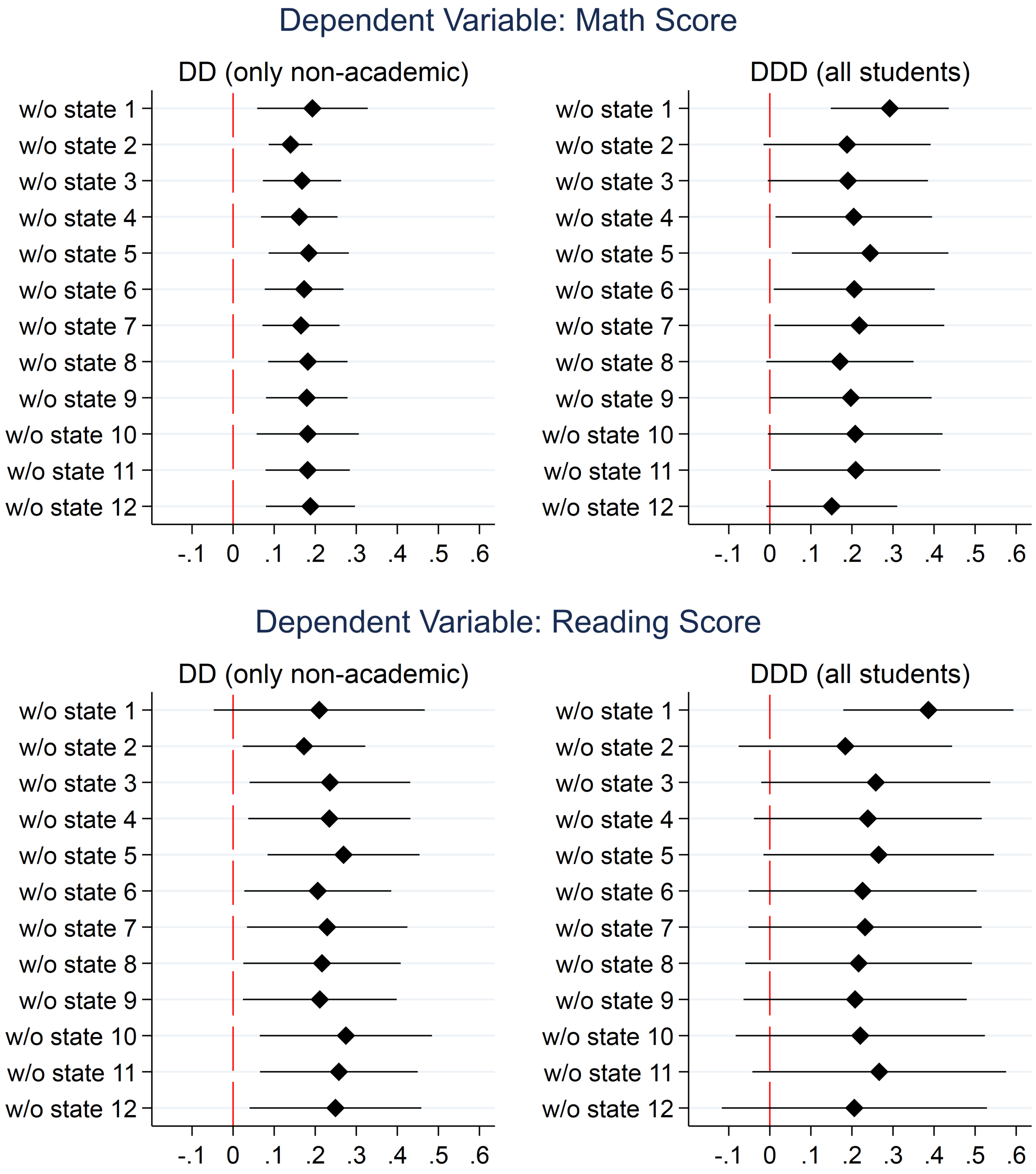

Figure 7: Robustness check IV: Leave-one-state-out DD and DDD estimates.

Notes: The figure presents point estimates including 95%-confidence intervals (based on clustered standard errors) for the effect of comprehensive schooling for maths (Panel A) and reading scores (Panel B) when dropping one state at a time.

Table 1: Descriptive Statistics by Tracking Regime

	Tracked States		Comprehensive States	
Variable	Mean	SD	Mean	SD
Grade 5 cross-section				
Mathematics score	-0.59	(0.85)	-0.53	(0.88)
Reading score	-0.50	(0.92)	-0.44	(0.88)
Academic track*	0.48	(0.50)	0.51	(0.50)
Female*	0.47	(0.50)	0.48	(0.50)
Age in fifth grade	10.79	(0.53)	10.80	(0.66)
Migration background*	0.27	(0.44)	0.17	(0.37)
Nr. of siblings	1.14	(0.96)	0.95	(0.86)
Single parent*	0.09	(0.29)	0.08	(0.28)
Unemployment*	0.10	(0.30)	0.14	(0.35)
Parental education (years)	14.53	(2.40)	14.49	(2.18)
Household income (1,000)	3.79	(2.32)	3.46	(1.54)
N	3704		679	
Grade 7 cross-section				
Mathematics score	0.00	(0.93)	0.06	(0.83)
Reading score	0.05	(1.01)	0.18	(0.91)
Academic track*	0.48	(0.50)	0.46	(0.50)
N	4255		1034	
Grade 9 cross-section				
Mathematics score	0.61	(0.90)	0.56	(0.83)
Reading score	0.48	(0.83)	0.47	(0.81)
Academic track*	0.48	(0.50)	0.46	(0.50)
N	3307		850	

Notes: Stars * indicate binary variables. Number of observations are based on the number of of non-missing test scores for each grade level.

Table 2: DD and DDD estimates of comprehensive schooling effect on grade 7 outcomes

Model specification:	Double Differences	Triple Differences
	(1)	(2)
Panel A:	Mathematics	
Comprehensive schooling	0.174^{***} (0.043)	0.204^{**} (0.086)
Grade 7	0.535^{***} (0.013)	
Panel B:	Reading	
Comprehensive schooling	0.229^{**} (0.085)	0.239^{*} (0.126)
Grade 7	0.470^{***} (0.037)	
State FE	✓	
State×grade FE		✓
State×track FE		✓
Track×grade FE		✓
Obs. (students×grade)	5014	9660

Notes: Column (1) presents OLS estimates for the DD model defined in equation (1) with either grade 5 and 7 mathematics scores as the dependent variable (Panel A) or grade 5 and 7 reading scores (Panel B). The sample includes all students on regular non-academic-track schools with non-missing test scores. Column (2) presents OLS estimates of the DDD model defined in equation (2). All academic-track students with non-missing test scores are added to the sample. Clustered standard errors (at the federal state level) are reported in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Effect heterogeneity by previous achievement and SES for grade 7 outcomes

Dependent variable:	Δ_7 Mathematics			Δ_7 Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
Comprehensive schooling	0.135^{**} (0.048)		0.079 (0.046)	0.210^{***} (0.060)		0.210^{**} (0.088)
× Low		0.304^{***} (0.088)			0.268^{***} (0.078)	
× Mid-lower		0.121 (0.086)			0.189^{***} (0.057)	
× Mid-upper		0.101 (0.096)			0.149 (0.115)	
× High		-0.021 (0.051)			0.197 (0.114)	
× Low SES			0.098 (0.069)			-0.000 (0.101)
Obs. (students)	1670	1670	1659	1670	1670	1659

Notes: All regressions are based on the panel sample of non-academic-track students who have non-missing grade 5 and grade 7 test scores. Columns (1) and (4) present OLS estimates for the FD model of equation (4) for mathematics and reading grade 5-to-7 gain-scores, respectively. Columns (2) and (5) present results from OLS estimation of the FD model of equation (5), which interacts the treatment variable with dummies for each quartile of the grade 5 test score distribution (and adds these dummies as separate regressors). The model underlying columns (3) and (6) interacts the treatment with a dummy variable for reporting less than the median number of books at home. Clustered standard errors (at the federal state level) are reported in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Persistence of effects until grade 9

Dependent variable:	Δ_9 Mathematics			Δ_9 Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
Comprehensive schooling	0.088 (0.066)		0.126* (0.067)	0.101 (0.082)		0.096 (0.084)
× Low		0.156*** (0.044)			0.134 (0.158)	
× Mid-low		0.193 (0.122)			0.183** (0.065)	
× Mid-high		-0.129 (0.168)			0.036 (0.110)	
× High		0.090 (0.104)			0.083 (0.128)	
× Low SES			-0.071 (0.078)			0.017 (0.102)
Obs. (students)	1282	1282	1281	1246	1246	1242

Notes: All regressions are based on the panel sample of non-academic-track students who have non-missing grade 5 and grade 9 test scores. Regressions for grade 9 gain-scores are analogous to those for grade 7 reported in Table 2. Clustered standard errors (at the federal state level) are reported in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Inspecting potential socio-emotional mechanisms

Model specification:	Double Diff.	Cross-sectional OLS			
Dependent variable:	Aspirations (*dummy*)	Helplessness (*scale*)		Motivation (*scale*)	
	> low certificate	Maths	German	Maths	German
	(1)	(2)	(3)	(4)	(5)
Comprehensive schooling	0.080^{*} (0.041)	-0.116^{**} (0.050)	-0.076 (0.049)	0.144 (0.092)	0.137^{*} (0.073)
Grade 7	-0.023^{**} (0.009)				
Constant		1.195^{***} (0.336)	1.274^{***} (0.288)	2.195^{***} (0.507)	1.956^{***} (0.163)
State FE	✓				
Basic & SES controls		✓	✓	✓	✓
Observations	(students×grade) 5021	(students) 2511	(students) 2523	(students) 2339	(students) 2322

Notes: Column (1) reports OLS estimates for the DD model of equation (1) with as dependent variable an indicator variable equal to unity when the student reports higher realistic educational aspirations than the low-track school-leaving certificate. The sample includes all grade 5 and grade 7 observations with non-missing values for the aspirations variable. The remaining columns are based on grade 7 cross-sectional samples, in each regression including all students with non-missing values for the dependent variable. These OLS regressions control for the following covariates: student age, sex, migration background, number of siblings, parental years of education, household income, unemployment and single parent household, with missing values in each of them replaced with variable means and the respective missing dummies added. Helplessness ranges from 1 to 4, with larger values indicating a higher degree of feeling of helplessness in the respective school subject. The variable averages 5 survey items, each measured on a 4-point Likert scale. Motivation ranges from 1 to 4, with larger 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. Clustered standard errors (at the federal state level) are reported in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Robustness check I: Common trends in primary school

Dependent variable:	Mathematics (1)
Compr. state × Grade 4	-0.013 (0.033)
Grade 4	1.361^{***} (0.029)
State FE	✓
Obs. (students×grade)	11190

Notes: In contrast to all other results, these results are based on the primary school sample of the NEPS (Starting Cohort 2). The table presents OLS estimates for the DD model of equation (1) for grade 2 and 4 mathematics scores. The sample includes all students on regular primary schools with non-missing test scores. Clustered standard errors (at the federal state level) are reported in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Robustness check II: Potential confounding through home and school inputs

Specification:	Baseline	+ basic controls	+ SES controls	+ class-level controls
	(1)	(2)	(3)	(4)
Panel A:	Dependent variable: Δ_7 Mathematics			
Comprehensive schooling	0.135^{**}	0.132^{**}	0.132^{**}	0.143^{**}
	(0.048)	(0.050)	(0.050)	(0.063)
R^2	0.006	0.008	0.014	0.020
Panel B:	Dependent variable: Δ_7 Reading			
Comprehensive schooling	0.210^{***}	0.222^{***}	0.219^{***}	0.215^{***}
	(0.060)	(0.051)	(0.049)	(0.039)
R^2	0.008	0.017	0.019	0.024
Obs. (students)	1670	1670	1670	1670

Notes: All regressions are based on the panel sample of non-academic-track students who have non-missing grade 5 and grade 7 test scores. Panel A presents the estimation results for grade 5-to-7 gain-scores in mathematics and Panel B those for reading. Columns (2) to (4) gradually add the following covariates to the baseline FD model of equation (4), whose results are presented in column (1): basic individual controls (age, sex, migration background, number of siblings), SES controls (parental years of education, household income, unemployment, single parent household) and class-level controls (teacher experience, teacher further education, class size). Missing dummies are added as before (see notes of Table 5). Clustered standard errors (at the federal state level) are reported in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Robustness Check III: Inference using wild cluster bootstrap

		p-value	
	Coefficient	Cluster-robust variance estimate	Wild cluster bootstrap
	(1)	(2)	(3)
Panel A:		Mathematics	
DD	0.174	[0.002]	[0.006]
DDD	0.205	[0.037]	[0.024]
Panel B:		Reading	
DD	0.230	[0.020]	[0.030]
DDD	0.239	[0.084]	[0.078]

Notes: Column (1) presents OLS estimates of the effect of comprehensive schooling for the DD model of equation (1) (using only the non-academic-track sample) and the DDD model of equation (2) (using all students). Column (2) displays *p*-values based on the conventional cluster-robust variance estimator (CRVE) and a $t(G-1)$ distribution (where G is the number of clusters). Column (3) displays *p*-values based on the (restricted) wild cluster bootstrap. Bootstraps estimates are based on 999 replication and weights from Webb's (2014) six-point distribution for the case of few clusters.

Appendix

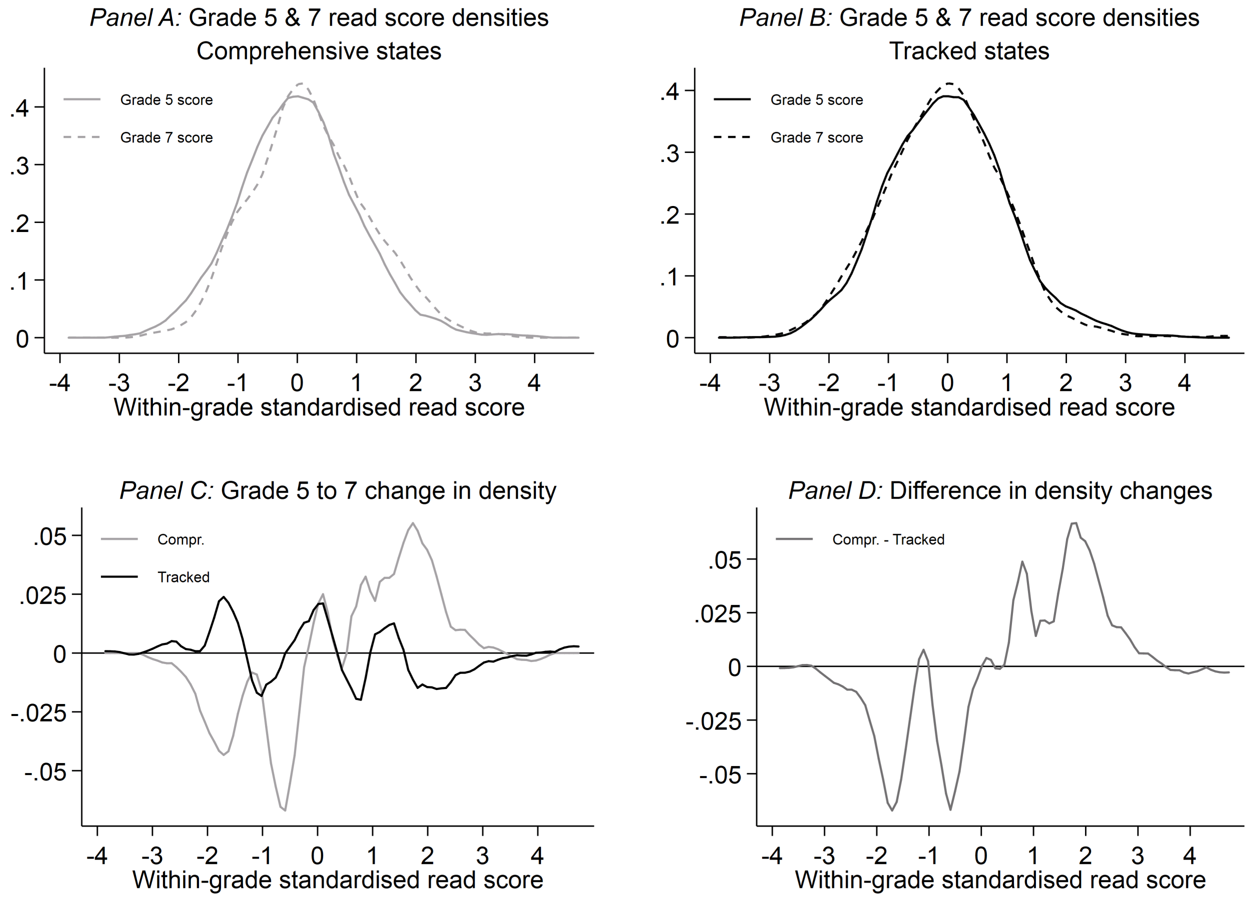

Figure A.1: Changes in reading score distribution through comprehensive schooling vs. tracking.

Notes: The densities in Panel A and B are based on the cross-sectional grade 5 and grade 7 samples with within-grade standardised reading scores. Estimation uses a Epanechnikov kernel estimator with optimal bandwidth (for normally distributed variables) (Silverman, 1986). For Panel C, I calculate the difference between the non-parametric density estimates of the grade 7 and grade 5 scores at 100 equally spaced points between the minimum and maximum of the standardised reading score for Tracked and Comprehensive states. Panel D plots the difference in these differences.

Table A.1: Effect heterogeneity by previous achievement tercile for grade 7 outcomes

Dependent variable:	Δ_7 Mathematics		Δ_7 Reading	
	(1)	(2)	(3)	(4)
Comprehensive schooling	0.135^{**} (0.048)		0.210^{***} (0.060)	
× Bottom		0.277^{***} (0.078)		0.248^{***} (0.070)
× Mid		0.090 (0.085)		0.119 (0.076)
× Top		0.030 (0.052)		0.227^{**} (0.076)
Obs. (students)	1670	1670	1670	1670

Notes: All regressions are based on the panel sample of non-academic-track students who have non-missing grade 5 and grade 7 test scores. Columns (1) and (3) present OLS estimates for the FD model of equation (4) for mathematics and reading grade 5-to-7 gain-scores, respectively. Columns (2) and (4) present results from OLS estimation of the interacted FD model like in equation (5) but with dummies for each *tercile*, instead of quartile, of the grade 5 test score distribution. Clustered standard errors (at the federal state level) are reported in parentheses. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.