The Distributional Effects of Early School Stratification – Non-Parametric Evidence from Germany

CRED Research Paper No. 19

Marcus Roller
University of Bern,
CRED

Daniel Steinberg
University of Tübingen

December 2017

Abstract

Whether early school stratification is conducive or detrimental to scholastic performance has been subject to controversial debates in educational policy and science across many countries. We exploit a unique exogenous variation in Lower Saxony, Germany, where performance based tracking was preponed from grade 7 to grade 5 in 2004, i.e. with the completion of primary school. In particular, we measure the long-run effects of this reform on PISA achievement test scores based on a difference-in-differences setup. In order to disentangle average from distributional achievement effects, we complementarily rely on a changes-in-changes framework. Our results indicate that preponed school tracking increased test scores at the upper tail of the skill distribution and lowered test scores at the lower tail of the skill distribution, compensating each other on average.*

Key words: financial development, credit constraints, international trade, inequality.

JEL classification: F160, F650

*This study does not represent a comprehensive evaluation of the exploited reform in Lower Saxony in 2004.
1 Introduction

While most countries teach students comprehensively in primary school, some countries track students into different tiers in secondary school. Other countries rely on comprehensive school systems in secondary school as well or even run both systems in parallel. The effect of school stratification on scholastic achievement is controversially discussed in policy and research across many countries. On the one hand, proponents of integrated learning emphasise positive spillover and peer-group effects in favour of low achievers, and therefore a decline in intergenerational path dependencies regarding educational attainment (see Dustmann, 2004). On the other hand, opponents of integrated learning emphasise negative spillover and peer-group effects at the cost of high achievers and an equalization of school achievements more generally (see Ariga et al., 2005; Brunello et al., 2004).

In order to examine whether integrated learning impinges on educational attainment, several studies have relied on international cross-country contexts. For instance, Hanushek and Woessmann (2006), Entorf and Lauk (2008), OECD (2004) as well as Schütz et al. (2008) show that preponed school tracking results in lower student achievement on average while the variance of school achievements increases. Apart from cross-sectional studies, further lines of research have been based on longitudinal data within countries. Scandinavia in particular has served as a distinct laboratory to investigate the link between school tracking and various outcome variables. In Norway as well as Sweden several educational reform packages were imposed in the mid-20th century. At the heart of these school reforms was the expansion of compulsory education and the equalization of curricula on a national level, as well as a partial integration of the two-tiered school system entailing an academic pillar and a non-academic pillar. According to Meghir and Palme (2005) and Aakvik (2003), these reforms led to longer educational participation and higher incomes in Sweden along with lower intergenerational educational correlations within families in Norway. This implies that being born in affluent households is less relevant for educational attainment in the aftermath of the reform. However, as part of the reform package, several reforms were put in place simultaneously, contaminating the identification of the specific causal channel. Hall (2012) examines a further educational reform in Sweden in the 1990s. The main ingredient of the reform was an extended portion of academic curricula in the apprenticeship track in secondary school, which also qualified pupils to studies. According to Hall (2012), further academic curricula propelled educational participation in secondary school. A similar reform package was put in place in Finland, postponing the stratification of pupils until they turned 15. According to Kerr et al. (2013), the reform elicited an increase in the achievements of pupils on average, especially of those with adverse socio-economic backgrounds. Turning to the US, Zimmerman (2003) sheds light on peer rather than tracking effects on educational attainment. Starting with the assumption that first year students are matched randomly

---

2Integrated learning is used in the sense of performance based school tracking rather than integrating pupils with certain disabilities.
with their roommates, he compares prior SAT scores with subsequent academic achievements for different matches. The author shows that negative peer effects worsen the roommates’ academic achievement, though this relationship is apparent exclusively for verbal SAT scores. Complementarily, Hoxby (2000) emphasises the relevance of reference group effects within class rooms; namely, “a credibly exogenous change of 1 point in peers’ reading scores raises a student’s own score between 0.15 and 0.4 points, depending on the specification” (Hoxby, 2000, p.1). Regarding Germany, Lehmann and Nikolova (2005a,b) compare the achievements in mathematics and reading comprehension in primary school and Gymnasium, the highest track in Germany’s 3-tier school system, in class 5 and 6. Thereby, the authors exploit a peculiarity of the educational system in Berlin, according to which high achieving students can leave primary school after the 4th grade, switching over to the Gymnasium. As pupils who switch to the Gymnasium are positively selected, it does not come as a surprise that school switchers perform better on average. In order to account for self-selection effects, Baumert et al. (2009b) rely on a propensity score matching strategy. For pupils with the same initial achievements, the authors can not disclose any disparities in achievement between the two groups.

Complementarily, Mühlenweg (2008) looked at the state of Hessen. In Hessen some schools offer a special track, the so called “Förderstufen”, which students pass into directly after primary school, bringing together students with different achievements ex ante. In parallel, pupils are generally stratified on three tracks according to their previous achievements in primary school. Mühlenweg (2008) does not detect any major disparities in the PISA-test achievements of participants in the “Förderstufen” compared to stratified students, though students with inferior socio-economic backgrounds seem to benefit from the postponed tracking. However, the results in Hessen might be biased due to self-selection effects. As the students can choose the school on their own, they might sort themselves into schools based on specific characteristics.

In order to address these problems, randomised experiments appear to be more suitable in precluding self-selection effects. Piopiunik (2014) relied on such a quasi-experimental research design in Bavaria. As of 2000, students were stratified into three rather than two tracks based on their previous performance in primary school. In coherence with the theoretical predictions, Piopiunik (2014) concludes that further tracking gave rise to lower state PISA-test scores based on a state-level difference-in-difference research design. However, in his framework, there might be self-selection in or out of the untreated third track which would bias the results. Duflo et al. (2011) address the self-selection issue by exploiting experimental data of first year primary school students in Kenya. After the end of one year of tracking, the authors find that high achievers are particularly better off in the course of tracking, though changes in teachers’ incentives might make low achievers better off as well. Although their identification strategy is strong, it is questionable how their findings translate to developed countries. First, their tracking experiment took place in the first grade. None of the OECD countries tracks its students in primary school. Second, the Kenyan school drop out rates are severe while they do not play a major role in developed countries. Third, the authors measure the immediate effects
of school tracking on test scores. The high drop-out rates and the abolition of tracking in most schools after grade 1 make their results for one year after the program unlikely to be valid for developing countries.

In order to examine the distributional effects of school stratification in the long run, we combine a theoretical model with an empirical investigation. Theoretically, we set out a simple model of human capital development involving positive spillover effects from high-skilled to low-skilled students and a penalty that punishes teaching targeted at a distant student in the skill distribution (congruency effects). Hence, the net effect of preponed school tracking is ambiguous in the lower tail and unambiguously positive in the upper tail of students’ achievement. Students in the lower track experience a loss of positive spillover effects in the course of preponed school tracking but positive congruency effects. Students in the upper track, however, exclusively experience positive congruency effects in the course of a shift from a comprehensive school to a stratified school system. Augmenting the framework with negative spillover effects as suggested by Lavy et al. (2012) even strengthens these theoretical predictions. Finally, the model suggests that the trade-off we find exists for any pair-wise comparison of tracking systems, since all tracking systems in the model are shown to be Pareto-efficient.

Empirically, we rely on a quasi-randomised school experiment in Lower Saxony, where performance based streaming was preponed from grade 7 to grade 5 in 2004. The empirical study provides both a singular treatment in the respective time period and low probabilities of self-selection in and out of the reform group. Until 2004, all students in Lower Saxony were taught in a comprehensive school, the so-called “Orientierungsstufe” serving as an intermediary between primary and secondary school in grades 5 and 6. As of 2004, all students were streamed into three different tracks depending on their previous achievement directly after primary school. In order to test whether the theoretical predictions are persistent, we compare the individual PISA reading achievement test scores of student cohorts who were totally exposed to the policy intervention with the achievement of a control group made up of students in other German states who were not exposed to the intervention. In order to disentangle average and quantile achievement effects, we complement our difference-in-differences identification strategy with a changes-in-changes setup originally proposed by Athey and Imbens (2006) and extended by Melly and Santangelo (2015). The latter is inevitable as we expect heterogenous effects across the skill distribution, translating in heterogenous effects along the performance distribution in light of the theoretical predictions. In line with the theoretical predictions, we find negative performance effects in the lower tail of the performance distribution and positive effects in the upper tail, and hence insignificant effects at the mean. Furthermore, our findings suggest that there is no superior tracking system because in either case one tail of the distribution is worse off. Policy makers are therefore faced with a trade-off between optimizing the performance of low-skilled and high-skilled students.

3The PISA test takes place at the end of the ninth grade.
This study contributes to the empirical literature on early school tracking by providing causal estimates of the effect of early school tracking on the achievements of students. Furthermore, it is up to our knowledge the first study explicitly estimating the heterogenous effects along the entire distribution of students’ achievements. It also contributes to the theoretical literature on tracking by providing a model that is flexible about the skill level of the student at whom the teaching is targeted.

The paper is organised as follows. Section 2 derives a theoretical framework for the effect of school tracking on educational achievement. Section 3 provides a brief description of the institutional background with respect to educational federalism in Germany, the specific educational reform analysed and, more broadly, the PISA test environment. Section 4 confronts theory with data based on a difference-in-difference along with a changes-in-changes setup. Section 5 concludes.

2 Theory

In order to derive the link between school tracking and educational attainment, we have to contrast peer-group spillover effects and congruency effects. Regarding the former, individual performances are affected by the achievements of top classmates, i.e. low achievers experience positive spillover effects if they are grouped with high achievers. Regarding the latter, if teaching is targeted at the median student in class both high and low skilled students suffer, i.e. for the former the level of teaching is too low and for the latter it is too high (e.g. Brunello et al. (2004)).

Formally, suppose there is a continuum of students and their initial ability $\theta$ follows a uniform distribution, $\theta \sim U(\bar{\theta}, \bar{\theta})$. Students are tracked into $1, \ldots, J$ tiers according to their initial ability. Hence, all students with $\theta \in (\bar{\theta}_{j-1}, \bar{\theta}_j]$ are taught in tier $j$. Unlike in Duflo et al. (2011), human capital, i.e. their potential to score in tests, for a student with initial ability $\theta$ taught in tier $j$ is deterministic and given by:

$$h(\theta, \bar{\theta}_j, \tilde{\theta}_j) = f(\theta) + e(\theta, \tilde{\theta}_j) = f(\theta) + s(\theta, \bar{\theta}_j) + c(\theta, \tilde{\theta}_j) = f(\theta) + s(\theta, \bar{\theta}_j) + c = z - k(\theta, \tilde{\theta}_j)$$

with $f(\theta)$ strictly increasing in initial ability, $\partial f(\theta)/\partial \theta > 0$. $s$ is the spillover effect which depends positively on the distance to the student with the highest ability in class $j$: $\partial s(\theta, \bar{\theta}_j)/\partial (\bar{\theta}_j - \theta) > 0$. Moreover, we postulate that the second derivative is positive, $\partial^2 s(\theta, \bar{\theta}_j)/\partial (\bar{\theta}_j - \theta)^2 \geq 0$. Namely, the marginal gains in spillover effects are non-decreasing in the distance to the highest skill in class. $c$ captures the congruency effect which is maximised for the targeted students.

4 We abstract from possible positive spillovers of social skills through the entire paper, since PISA tests exclusively account for problem solving skills.
at \( z \). \( k(\theta, \tilde{\theta}_j) \) depends positively on the absolute distance to the median skill \( \tilde{\theta}_j \) in class \( j \): 
\[
\frac{\partial k(\theta, \tilde{\theta}_j)}{\partial \left( |\tilde{\theta}_j - \theta| \right)} > 0.
\]
Without loss of generality, we also posit that \( s(a, b) = 0 \) and \( k(a, b) = 0 \) if \( a = b \). The assumption that teaching in each class is targeted at the median skill student in the respective class is easy to legitimise as long as teachers do not have a different incentive to target their teaching towards a specific group of students within the class unlike in Duflo et al. (2011). Note that teaching levels do not serve as a direct argument of the production function because curricula were not affected in the course of the reform. Therefore, better teaching in our case is meant in the sense of more congruent teaching, providing a better fit to the initial abilities in class. Moreover, we postulate that congruency effects are more than compensated by spillover effects for increasing negative distances to the targeted student:
\[
\frac{\partial s(\theta, \tilde{\theta}_j)}{\partial (\tilde{\theta}_j - \theta)} > \frac{1}{2} \frac{\partial k(\theta, \tilde{\theta}_j)}{\partial (|\tilde{\theta}_j - \theta|)}
\]
(2)

We highlight the role of this assumption in the discussion of the results below. Finally, we assume that the tracking system does not affect the observability of inherent skills, \( \theta \), which is a reasonable assumption in light of our empirical setting (see table 9 in the appendix which is discussed in the empirical section).

Based on the setup, we proceed with an analysis of achievement effects in the course of a transition from a comprehensive school (N) to a stratified school system (I) based on three tracks according to individual skill levels \( \theta \). All students with ability \( \theta \in [\bar{\theta}_1, \tilde{\theta}_1] \) are taught in class 1, where \( \theta_1 \) is defined such that one third of the students have lower ability, \( U(\theta_1) = 1/3 \). Students with ability \( \theta \in (\bar{\theta}_1, \tilde{\theta}_2] \) are taught in class 2, where \( \theta_2 \) is defined such that two thirds of the students have lower ability, \( U(\theta_1) = 2/3 \). The remaining students with \( \theta \in (\tilde{\theta}_2, \tilde{\theta}_3] \) are taught in class 3. Table 1 summarises the results stated in proposition 1-3. Figure 1 presents an illustrative example according to which both spillover and the congruency effects are linear in distances.

Table 1: Theoretical Effects of School Tracking

<table>
<thead>
<tr>
<th>Effect</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spillovers</td>
<td>( s(\theta, \tilde{\theta}_1) &lt; s(\theta, \tilde{\theta}) )</td>
<td>( s(\theta, \tilde{\theta}_2) &lt; s(\theta, \tilde{\theta}) )</td>
<td>( s(\theta, \tilde{\theta}) = s(\theta, \tilde{\theta}) )</td>
</tr>
<tr>
<td>Congruency</td>
<td>( c(\theta, \tilde{\theta}_1) &gt; c(\theta, \tilde{\theta}) )</td>
<td>( c(\theta, \tilde{\theta}) = c(\theta, \tilde{\theta}) )</td>
<td>( c(\theta, \tilde{\theta}_3) &gt; c(\theta, \tilde{\theta}) )</td>
</tr>
<tr>
<td>Total</td>
<td>( e(\theta, \tilde{\theta}_1, \tilde{\theta}_1) &lt; e(\theta, \tilde{\theta}, \tilde{\theta}) )</td>
<td>( e(\theta, \tilde{\theta}_2, \tilde{\theta}_2) &lt; e(\theta, \tilde{\theta}, \tilde{\theta}) )</td>
<td>( e(\theta, \tilde{\theta}, \tilde{\theta}_3) &gt; e(\theta, \tilde{\theta}, \tilde{\theta}) )</td>
</tr>
</tbody>
</table>

\(^5\)We postulate three classes together containing the entire distribution of students, which ensures that the expenditures of both tracking systems coincide.
**Proposition 1.** Students in the upper track, class 3, are strictly better off in the stratified school system compared to a comprehensive school: \( h(\theta, \tilde{\theta}, \bar{\theta}_3) > h(\theta, \tilde{\theta}, \bar{\theta}) \ \forall \theta \in (\bar{\theta}_2, \bar{\theta}) \)

**Proof:** Students in the upper track do not experience changes in spillover effects in the course of the transition from a comprehensive school to a stratified school system, \( s(\theta, \tilde{\theta}) = s(\theta, \tilde{\theta}) \). However, the distance to the class median is lower than the distance to the overall median for all ability levels in the upper class, \( |\tilde{\theta}_3 - \theta| < |\tilde{\theta} - \theta| \ \forall \theta \in (\bar{\theta}_2, \bar{\theta}) \) which implies that skills are more congruent for all of them, \( c(\theta, \tilde{\theta}_3) > c(\theta, \tilde{\theta}) \). In total, all students are better off under the tracking system: \( e(\theta, \tilde{\theta}, \bar{\theta}_3) > e(\theta, \tilde{\theta}, \bar{\theta}) \) and \( h(\theta, \tilde{\theta}, \bar{\theta}_3) > h(\theta, \tilde{\theta}, \bar{\theta}) \). ■

**Figure 1:** Theoretical effects of school tracking

![Graph](image)

**Notes:** Differences in spillover effects, congruency effects, and overall effects between a no tracking system with a three tier tracking system. Example with linear spillover and congruency effects.
**Proposition 2.** Students in the middle track, class 2, are strictly worse off in the stratified school system compared to a comprehensive school: \( h(\hat{\theta},\tilde{\theta}_2,\hat{\theta}_2) < h(\hat{\theta},\tilde{\theta},\hat{\theta}) \quad \forall \theta \in (\hat{\theta}_1,\hat{\theta}_2] \)

**Proof:** Students in the middle track do not experience changes in congruency effects in the course of the transition from a comprehensive school to a stratified school system, \( c(\theta,\hat{\theta}) = c(\theta,\tilde{\theta}) \). However, the distance to top student in class is lower than the distance to the overall top student for all ability levels in the middle class, \( \hat{\theta}_3 - \theta < \tilde{\theta} - \theta \quad \forall \theta \in (\hat{\theta}_1,\hat{\theta}_2] \), which implies that spillover effects are lower for them, \( s(\hat{\theta},\tilde{\theta}_2) < s(\hat{\theta},\tilde{\theta}) \). In total, students in the middle track are worse off under the tracking system: \( e(\theta,\tilde{\theta}_2,\hat{\theta}_2) < e(\theta,\tilde{\theta},\hat{\theta}) \) and \( h(\theta,\tilde{\theta}_2,\hat{\theta}_2) < h(\theta,\tilde{\theta},\hat{\theta}) \).

**Proposition 3.** Students in the lower track, class 1, are strictly worse off in the stratified school system compared to a comprehensive school: \( h(\theta,\tilde{\theta}_1,\hat{\theta}_1) < h(\theta,\tilde{\theta},\hat{\theta}) \quad \forall \theta \in [\tilde{\theta},\hat{\theta}_1] \)

**Proof:** The top student in class 1 is worse off because his distance to the class median and the overall median is the same, \( |\hat{\theta}_1 - \tilde{\theta}_1| = |\hat{\theta} - \tilde{\theta}| \), and hence the congruency is equal, \( c(\tilde{\theta}_1,\hat{\theta}_1) = c(\tilde{\theta}_1,\hat{\theta}) \). However, the distance to the top student in class is lower than the distance to the overall top student for all ability levels in the middle class, \( \tilde{\theta}_1 - \theta = 0 < \hat{\theta} - \tilde{\theta}_1 \) which implies that spillover effects are lower, \( s(\tilde{\theta}_1,\hat{\theta}_1) < s(\tilde{\theta}_1,\hat{\theta}) \). We can conclude that: \( c(\tilde{\theta}_1,\hat{\theta}_1,\hat{\theta}_1) < c(\tilde{\theta}_1,\hat{\theta},\hat{\theta}) \).

The loss of the median class 1 student can be expressed as follows:

\[
\Delta(\tilde{\theta}_1) = s(\tilde{\theta}_1,\hat{\theta}) - s(\tilde{\theta}_1,\hat{\theta}_1) - k(\tilde{\theta}_1,\hat{\theta}) + k(\tilde{\theta}_1,\hat{\theta}_1)
\]

\[
= s(\tilde{\theta}_1,\hat{\theta}) - k(\tilde{\theta}_1,\hat{\theta}) + \int_{\tilde{\theta}_1}^{\tilde{\theta}_1} \frac{\partial s(\theta,\tilde{\theta})}{\partial \theta} - \frac{\partial s(\theta,\hat{\theta}_1)}{\partial \theta} d\theta
\]

while \( s(\tilde{\theta}_1,\hat{\theta}) > k(\tilde{\theta}_1,\hat{\theta}) \) follows from the compensation assumption. The integral is also positive because we assumed non-decreasing returns to distance in spillover effects. Thus, we can conclude that: \( e(\tilde{\theta}_1,\hat{\theta}_1,\hat{\theta}_1) < e(\tilde{\theta}_1,\hat{\theta},\hat{\theta}) \).

For students for which \( \tilde{\theta}_1 < \theta < \hat{\theta}_1 \) we can formulate the loss in the course of a transition from a comprehensive school to a stratified school system as:

\[
\Delta(\theta) = s(\theta,\hat{\theta}) - s(\theta,\hat{\theta}_1) - k(\theta,\hat{\theta}) + k(\theta,\hat{\theta}_1)
\]

\[
= s(\theta,\hat{\theta}) - k(\theta,\hat{\theta}) + k(\theta,\hat{\theta}_1) + \int_{\theta}^{\theta_1} \frac{\partial s(\theta,\hat{\theta})}{\partial \theta} - \frac{\partial s(\theta,\hat{\theta}_1)}{\partial \theta} d\theta
\]

Since \( k(\theta,\hat{\theta}) < k(\tilde{\theta}_1,\hat{\theta}) \), and all other addends are positive, we can directly conclude that: \( e(\theta,\tilde{\theta}_1,\hat{\theta}_1) < e(\theta,\tilde{\theta},\hat{\theta}) \quad \forall \theta \in (\tilde{\theta}_1,\hat{\theta}_1) \).

For students characterised by \( \theta \leq \theta < \tilde{\theta}_1 \) it holds:

\[
(\hat{\theta} - \theta) - (\tilde{\theta}_1 - \theta) = \frac{2}{3}(\hat{\theta} - \theta)
\]

\[
(\hat{\theta} - \theta) - (\tilde{\theta}_1 - \theta) = \frac{1}{3}(\hat{\theta} - \theta)
\]
Because students with $\theta = \tilde{\theta}_1$ are strictly better off, the two equations above together with the compensating assumption and the non-decreasing returns to distance assumption for spillovers imply that: $e(\theta, \tilde{\theta}_1, \tilde{\theta}_1) < e(\theta, \tilde{\theta}, \tilde{\theta}) \quad \forall \theta \in [\hat{\theta}, \tilde{\theta}_1)$.

It follows that all students in class 3 are worse off because the total effect for all of them is lower in the stratified school system: $e(\theta, \bar{\theta}_1, \tilde{\theta}_1) < e(\theta, \bar{\theta}, \tilde{\theta}) \quad \forall \theta \in [\theta, \bar{\theta}_1]$. ■

Our model also implies that there is no unambiguously optimal degree of tracking if we compare the two systems since some students are better off while others are worse off. This result holds for any pairwise comparison of possible tracking systems along the skill distribution. Thus, each tracking system is Pareto efficient. In order to make an optimal decision, policy makers therefore have to contrast the costs and benefits of the respective students in the skill distribution. This is proven below.

**Proposition 4.** All kinds of tracking systems (including no tracking) are Pareto-efficient.

*Proof:* In order to prove proposition 4, we have to distinguish between two scenarios. In the first scenario, we investigate differing thresholds for a given number of tiers, while in the second scenario we ascertain the transition to a system with a different number of tiers. In order to prove Pareto-efficiency, we just have to show that in either case there is somebody strictly better (worse) off.

**Scenario 1:** Suppose there exists a system with $J$ tiers. Let $\bar{\theta}_j$ denote the threshold that separates tier $j$ from tier $j + 1$ which is the first tier from the right where the upper threshold remains unchanged. Now suppose that this threshold is shifted to the right such that $\bar{\theta}_j' > \bar{\theta}_j$. The students in tier $j + 1$ with $\theta > \bar{\theta}_j'$ are strictly better off because their spillovers do not change and they experience higher congruency effects. If the threshold $\bar{\theta}_j$ is shifted to the right and the overall number of tiers stays the same, there must exist a tier $j - i$ to the left where $\bar{\theta}_j' - i > \bar{\theta}_j - i$. The student with exactly $\bar{\theta}_j' - i$ may now experience higher congruency effects but looses all spillover effects. Since he cannot be compensated for the loss of spillover effects, he is clearly worse off. Consequently, adapting the threshold for a given number of tiers always makes at least one student worse and one student better off in either system.

**Scenario 2:** Suppose there is a change in the number of tiers. Without loss of generality, we posit that both systems have exactly the same number of tiers but the upper tier is divided into two tiers in system 2. All students to the right of the additional threshold are better off (proof: proposition 1), whilst those to the left are strictly worse off (proof: proposition 3). Finally, we can conclude that there is no pairwise comparison of tracking systems according to which in one system none of the students is worse off than in the other system. Therefore, every tracking system is Pareto-efficient. ■

The effects shown in the model are economically only relevant if they are persistent over time and are not equalised by the subsequent tracking that takes place in either case. Hence, our
empirical study focuses on the long-run effects measuring the reading achievements in the PISA test 3 years after the students went through either a tracked or comprehensive 5th and 6th grade. Unfortunately, there is no comparable test at the end of grade six that would even allow us to tackle the degree of persistence.

In the following sections, we contrast theoretical predictions with data based on an exogenous policy change in Lower Saxony, Germany. Before setting out the econometric model, we provide a brief description of the educational system in Germany in general, and the educational reform in Lower Saxony in particular, along with a description of the Programme for International Student Assessment (PISA).

3 Institutional Background

3.1 Educational Federalism in Germany

In Germany, each state is independently in charge of educational policies. Therefore, the educational system in Germany is characterised by several federal elements basically leading to 16 different educational systems on a state level. However, the general system of schooling is similar across states: In essence, while nursery school for children between 2 and 6 years is optional, pupils thereafter attend a compulsory primary school. Based on their achievement in primary school and upon teachers’ recommendations, pupils decide upon the specific track in secondary school. Most of the former West-German states have three tracks of secondary schools, namely the “Hauptschule” (HS), serving students interested in apprenticeship programs and vocational training, the more demanding “Realschule” (RS) and the “Gymnasium” (GY), which qualifies them for academic studies. The former East-German states track their students into only two tiers. Complementarily, all states allow comprehensive schools combining the different tracks.

While there are differences in the tracking itself across states, there are also differences in the timing of the tracking. Most of the states track their students as of the 5th grade. In Berlin and Brandenburg however the primary school lasts 6 years. Lower Saxony and Bremen used to be special cases where students attended an intermediary school after a 4-year primary school lasting 2 years, which was called “Orientierungsstufe” (OS) before being divided into three tracks. Both, Lower Saxony and Bremen abolished this school type in 2004 and 2005, respectively, preponing the tracking to the beginning of grade 5.

---

7 In some states the recommendation is binding.
3.2 School Tracking in Lower Saxony

In 2003, the state government put in place a major educational reform package, the so-called “Gesetz zur Verbesserung der Schulqualität und zur Sicherung von Schulstandorten”. At the heart of the educational enactment was the complete replacement of the “Orientierungsstufe” with secondary schools directly after the completion of primary school. By 2004, the school system in Lower Saxony was characterised by some idiosyncrasies as pointed out above. Namely, between 1973 and 2004 in grade 5 and 6 all students were educated jointly in a comprehensive school, the so-called “Orientierungsstufe” (OS), while other states normally stratified students directly after primary school.\(^8\) Although separate schools in different buildings, the class composition of students remained the same in grade 5 and 6, which results in the system effectively having a 6 year primary school. Hence, the reform package standardised the year of tracking across German states. 2 illustrates the preponed tracking in Lower Saxony graphically.

Figure 2: Preponed school tracking in Lower Saxony

![Diagram showing preponed tracking](image)

Notes: Visualization of the preponed tracking. HS: Hauptschule (lower track); RS: Realschule (middle track); GY: Gymnasium (upper track).

Besides preponed tracking, the reform package entailed a shortening of the Gymnasium from 9 years to 8 years. We will discuss this issue in the results section.

3.3 Control Group

Our identification strategy rests on a comparison of the change in students’ achievements in Lower Saxony from 2006 to 2009 and the achievements of those in other German states who

\(^8\)This holds with the exception of Berlin and Brandenburg, which offer primary schools over 6 years.
were not exposed to any intervention. In principle, all other German states could serve as part of the control group, but we need to preclude some of them because they either underwent a relevant reform during the same time period or their school system exhibits major idiosyncrasies. First, we need to dispense with Bremen, which abolished the “Orientierungsstufe” in 2005, such that the test cohort of 2009 is partially treated. Furthermore, we drop all former East-German states along with Saarland because they do not stream their students into three but only two school types. Keeping them in the control group would require additional common trend assumptions which might be questionable. Thus, the control group consists of Baden-Württemberg, Bavaria, Hamburg, Hessen, North Rhine-Westphalia, Rhineland-Palatinate and Schleswig-Holstein. Figure 2 visualises Lower Saxony as a treatment group while all other former West-German states besides Bremen and Saarland serve as part of the control group.

Figure 3: Treatment and control group

Before proceeding with the econometric specification, we provide a brief description of the data underlying our analyses.
4 Evidence

4.1 Data

4.1.1 Data Sets

Our outcome variable of interest is reading achievement in the PISA test of ninth graders. The Programme for International Student Assessment (PISA) goes back to an initiative of the OECD which attempts to measure the achievement of 15-year-old students in mathematics, reading and science across OECD member countries. As of 2000, international tests have been conducted every 3 years with a certain focus. Unlike alternative international tests like Trends in International Mathematics and Science Study (TIMSS), which measure specific mathematical skills related to certain curricula, PISA tests are meant to capture general skills applied to real-world problems.

Although administered by the OECD, the formulation of test exercises and the analysis of the results is led by an international consortium headed by the Australian Council for Educational Research. In Germany, tests are carried out on behalf of the combined state boards of education, the so-called “Kultusministerkonferenz”. The scientific responsibility was borne by the Max Planck Institute for Educational Research in 2000, the Leibniz Institute for Education in Science in 2003 and 2006, the German Institute for Educational Research in Leipzig in 2009 and, finally, the Center for International Comparative Educational Research in 2012.

The PISA test generally consists of two parts, a test session of 2 hours and a questionnaire session of 1 hour. During the test session, each student solves a maximum of 50 problem tasks entailing multiple choice and whole answer questions selected from a total pool of 165 tasks. Hence, not all students are confronted with the same questions. In essence, exercises are generally marked as right or wrong, though retrospectively certain exercises might be evaluated as exceptionally difficult, which attaches a certain weight to them. The national test results are forwarded to the Australian Council for Educational Research, which serves as a center for data analysis.

The number of countries participating in the test has increased dramatically. While in 2000 28 OECD and 4 additional countries participated in the test, in 2012 34 OECD and 31 additional countries took part. Meanwhile, the number of participating students rose from 265,000 in 2000 to 510,000 in 2012. The following figure displays average PISA test scores in three subjects spanning the period from 2000 to 2012.
Table 2: Average PISA scores in Germany

<table>
<thead>
<tr>
<th>Year</th>
<th>Focus</th>
<th>Mathematics</th>
<th>Science</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>Reading</td>
<td>490</td>
<td>487</td>
<td>484</td>
</tr>
<tr>
<td>2003</td>
<td>Mathematics</td>
<td>503</td>
<td>502</td>
<td>491</td>
</tr>
<tr>
<td>2006</td>
<td>Science</td>
<td>504</td>
<td>516</td>
<td>495</td>
</tr>
<tr>
<td>2009</td>
<td>Reading</td>
<td>513</td>
<td>520</td>
<td>497</td>
</tr>
<tr>
<td>2012</td>
<td>Mathematics</td>
<td>514</td>
<td>524</td>
<td>508</td>
</tr>
</tbody>
</table>

Source: Klieme et al. (2010).

Since international PISA test scores are comparable across countries but not necessarily across states within Germany, the German test committee complementarily relied on larger samples in order to generate representative samples even on a state level. These tests serve as a foundation for regional comparisons within Germany, commonly known as PISA-E-tests. On average, the sample sizes in PISA-E tests exceed the sample size of the international PISA test by the factor of 10. In Germany 46,000, 39,000 and 10,000 students took part in international tests in 2003, 2006 and 2009, respectively. While the international PISA-E test allows for international achievement comparisons, the PISA test allows for comparisons throughout Germany. However, the PISA-E-tests were phased out in 2006 and replaced by similar tests organised by the so-called “Institut für Qualitätssicherung im Bildungswesen (IQB)” founded by the “Kultusministerkonferenz”. The tests conducted by the IQB have a certain focus, alternating between reading, math and natural science every three years, though the results are highly correlated with PISA test scores. In our empirical strategy below we make use of both, PISA-E test scores prior to 2009 as well as IQB test scores post 2006. As these tests capture the same underlying skills and are highly correlated, they serve as a foundation for our changes-in-changes strategy. The changes-in-changes approach explicitly allows for transitions in test exercises as long as these tests measure the same sort of skills and the transition materialises in the treatment and control group (Melly and Santangelo, 2015).

Large representative samples for all states are only available for the reading test of students in grade 9. Hence, we will focus on the reading achievements of these students. However, we will make use of the 15-year-old students as well to provide more extensive pretreatment tests. Table 3 gives an overview of the data sets used along with the respective sample sizes.
Table 3: Overview data sets

<table>
<thead>
<tr>
<th>Year</th>
<th>Test</th>
<th>Sample</th>
<th>Observations</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>PISA-E</td>
<td>15-year-old</td>
<td>35,584</td>
<td>Baumert et al. (2009a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9th grade</td>
<td>34,754</td>
<td>Baumert et al. (2009a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9th grade</td>
<td>39,216</td>
<td>Prenzel et al. (2006)</td>
</tr>
<tr>
<td>2009</td>
<td>IQB-Test</td>
<td>9th grade</td>
<td>39,663</td>
<td>Köller et al. (2011)</td>
</tr>
</tbody>
</table>

Figure 4 shows how we can exploit these data sets for our identification. The first student cohort being affected by the intervention in Lower Saxony was tested in 2009. This cohort started school in 2000. The previous cohort was also partially treated because the OS was abolished immediately stopping after grade five. Earlier cohorts are unaffected. The last cohort before the treatment that was tested in the PISA test is the 1997 cohort which was tested in 2006. The 2009 test cohort and the 2006 test cohort therefore allow us to base or difference in difference strategy on these repeated cross-sectional observations.

Figure 4: Treated and untreated student cohorts

Notes: The figure shows the effect of the reform on the four student cohorts that started school from 1997 to 2000.
4.1.2 Test Scores

In accordance with other large scale assessments, the goal of PISA and IQB-tests is not to provide a perfect measure of student skills on a micro level but rather a measure for the skill distribution of a whole population on a macro level. Hence, the plausible values provided as outcome variables allow us to investigate questions concerning the entire distribution without necessarily having precise estimates of individual skills (see von Davier et al., 2009). Usually, five plausible values are reported per student. Rather than taking the mean of these five values and performing descriptive analyses based on this mean, we are required to conduct the analyses with each item separately and take the average of the results. This procedure requires adjustments of standard errors according to the following formula (von Davier et al., 2009):

\[
SE(\hat{\mu}) = \left( 1 + \frac{1}{K} \left[ \frac{1}{K-1} \sum_{i=1}^{K} (\hat{\mu}_i - \hat{\mu})^2 \right] + \frac{1}{K} \sum_{i=1}^{K} \text{Var}(\hat{\mu}_i) \right)^{0.5}
\]

where \( K \) is the number of plausible values, \( \hat{\mu}_i \) the estimator for the mean using plausible value \( i \). \( \hat{\mu} \) is the average of all estimators: \( \hat{\mu} = \sum_{i=1}^{K} \mu_i \). This procedure is only applicable for descriptive analyses like means. Hence, we will rely on this procedure in order to check for common pretreatment trends and for the baseline difference-in-differences approach below. As soon as we are interested in multivariate analyses, each procedure has to be based on the first plausible value which requires a relatively large sample size compared to the univariate model (see Baumert et al., 2002). Since we base our estimates on PISA-E scores, we are confident that this requirement is met.

4.1.3 Sample Selection Issues

According to the quasi-randomised experiment, school tracking was preponed in Lower Saxony from grade 7 to grade 5 in 2004. As the PISA test is scheduled every three years for students in grade 9, the first student cohort to be affected by the intervention in Lower Saxony was tested in 2009. Ideally, we would like to compare long-run class compositions for Lower Saxony in both relevant years, 2006 and 2009. However, the composition of students generally differs between grade 1 and grade 9 because students have to repeat the respective class if they fall short of certain educational standards. While student cohorts preceding the test cohort and repeating a class before grade 7 become part of the test cohort and are at least partially exposed to the intervention, student cohorts repeating a class between grade 7 and 9 become part of the test cohort without being totally exposed to the intervention. However, since grade repetitions between grade 7 and 9 are unaffected by the reform, we do not encounter a self-selection problem. Yet, we must keep in mind that our estimates are biased towards zero because there are repeaters in the sample that are only partially affected by the reform (one repetition in grades 7 to 9) or not at all (two or more repetitions in grades 7 to 9). The former experienced
the same duration of stratified schooling as the test cohort, whilst the latter make up a very tiny share of the sample\(^9\), which should not drive or undermine our results. Complementarily, we have to illuminate changes in grade repetitions in grade 5 and 6 in the course of the reform as well. According to the panels on the left hand side in figure 5 and figure 6, both rates experienced a strong increase right after the reform in school year 2005/2006. At first glance, the rise in grade repetitions might lead to a sample selection of tested student cohorts. If grade repeaters are adversely selected, tested cohorts might be positively selected and we might overestimate the effect of preponed stratification on students’ performance. However, due to the preponed stratification, we have to contrast average repetition rates prior to the reform and cumulative repetition rates post of the reform. A decomposition of the repetition effects with respect to schools and years on the right hand side of figure 5 and 6 below puts sample selection issues into perspective.

**Figure 5:** Share of repetitions grade 5

![Graph showing the share of repetitions grade 5](image)

*Notes:* Shares of repeaters as percentages of total students attending grade 5.
Source: Federal Statistical Office.

\(^9\)In Lower Saxony there are only 4 students in our sample that repeated twice in grade 7 to 9.
Figure 6: Share of repetitions grade 6

Notes: Shares of repeaters as percentages of total students attending grade 6.
Source: Federal Statistical Office.

Apparently, prior to the reform, repeaters are made up of the lowest tail of achievement, while post reform even students in the middle or upper tail of overall achievement are forced to repeat. In other words, prior to the reform, student’s achievements are conditioned on the pooled average standard, while post reform, student achievements are conditioned on the corresponding standard in each school type. Hence, after the reform the lower tail of achievement in the Gymnasium is now forced to repeat, though they would not have been forced to repeat in a comprehensive school prior to the reform. The same holds for the lower tail of students in the Realschule. The lower tail of achievement in the Hauptschule, however, would have been forced to repeat a grade in a comprehensive school as well.

Since the rate already increases immediately for the sixth graders, we do not have a selection problem for those students because they already fall into the 2009 cohort with a higher rate. This is not the case for students that have to repeat grade 5. They fall out of the 2009 cohort at a higher rate but fall into it at the old low rate. The increase was almost entirely driven by students in the Realschule and there is no reason to assume that they bunch at a certain level after repeating. Neither is it entirely clear in which direction a possible bias goes. It might well be the case that the general selection is positive due to opposite tail effects. Even if this were the case, the bias is rather small because those students make up only about 1.5% of all students.

Consequently, although we generally discussed two potential sources of sample selection, we do not expect them to drive our results both because the number of affected students is extremely low and because the expected bias is rather towards zero. We control for age in the settings with covariates in order to account for both repetition issues.
4.1.4 Descriptive Statistics of Covariates

Table 4 reports descriptive statistics for all variables we make use of in our analyses except for the outcome variable, which is introduced in the course of pre-treatment tests below.\textsuperscript{10} With respect to covariates, we account for the age in months along with a dummy variable, which is 1 for male students and 0 otherwise, and a dummy variable that equals 1 if the student is enrolled in a certain track (Hauptschule, Realschule, Gymnasium) and 0 otherwise. We observe these variables for the entire sample. In the raw data females are slightly overrepresented in Lower Saxony, while there are no disparities between Lower Saxony and the control group regarding age. Our analysis is based on the assumption that changes in the sorting of students prior to and post reform materialise in the treatment and control group in parallel. If we dispense with population weights, however, the sorting effects exhibit certain deviations over time between the treatment and control group. However, as soon as we account for population weights, these trends are much more parallel (see table 8 in the appendix). Furthermore, we will make use of a variable indicating whether the students speak German at home (=1) or a foreign language (=0), thus capturing potential migration backgrounds of students.\textsuperscript{11} We observe this variable for most of the students. However, our sample for Lower Saxony contains fewer students speaking foreign languages compared to the control group prior to and post reform. As the share of students speaking foreign languages at home is not affected by the intervention, we will not make use of this variable as a control. However, we account for it in order to compare effects of students with and without migration backgrounds. The same holds for the variable Parents Abitur which equals 1 if at least one of the parents successfully completed the Gymnasium with an Abitur. We observe slightly fewer students whose parents earned an Abitur in Lower Saxony compared to the control group.

\textsuperscript{10}Note that these contain the raw data without using population weights.

\textsuperscript{11}The questionnaire contains several variables aiming at migration background which are all highly correlated.
Table 4: Descriptive statistics: controls. 9th grade students.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Lower Saxony</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>13,706</td>
<td>0.50</td>
<td>0.5</td>
</tr>
<tr>
<td>2006</td>
<td>19,023</td>
<td>0.50</td>
<td>0.5</td>
</tr>
<tr>
<td>2009</td>
<td>17,213</td>
<td>0.51</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>13,706</td>
<td>189</td>
<td>7</td>
</tr>
<tr>
<td>2006</td>
<td>19,023</td>
<td>188</td>
<td>7</td>
</tr>
<tr>
<td>2009</td>
<td>17,213</td>
<td>188</td>
<td>8</td>
</tr>
<tr>
<td><strong>Lower Track (HS)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>13,706</td>
<td>0.29</td>
<td>0.45</td>
</tr>
<tr>
<td>2006</td>
<td>19,023</td>
<td>0.28</td>
<td>0.45</td>
</tr>
<tr>
<td>2009</td>
<td>17,213</td>
<td>0.27</td>
<td>0.44</td>
</tr>
<tr>
<td><strong>Middle track (RS)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>13,706</td>
<td>0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>2006</td>
<td>19,023</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>2009</td>
<td>17,213</td>
<td>0.34</td>
<td>0.47</td>
</tr>
<tr>
<td><strong>Upper track (GY)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>13,706</td>
<td>0.36</td>
<td>0.48</td>
</tr>
<tr>
<td>2006</td>
<td>19,023</td>
<td>0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>2009</td>
<td>17,213</td>
<td>0.39</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>Home Language</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>12,261</td>
<td>0.92</td>
<td>0.27</td>
</tr>
<tr>
<td>2006</td>
<td>17,133</td>
<td>0.90</td>
<td>0.30</td>
</tr>
<tr>
<td>2009</td>
<td>14,778</td>
<td>0.89</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>Parents Abitur</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>12,508</td>
<td>0.31</td>
<td>0.46</td>
</tr>
<tr>
<td>2006</td>
<td>18,264</td>
<td>0.40</td>
<td>0.49</td>
</tr>
<tr>
<td>2009</td>
<td>7,723</td>
<td>0.37</td>
<td>0.48</td>
</tr>
</tbody>
</table>

In the following section we set out an econometric model in order to shed light on the effect of early school tracking on the distribution of students’ performance.

4.2 Empirical Strategy

In order to ascertain the effect of early school stratification on achievement test scores, we have to disentangle average achievement effects from quantile performance effects. Hence, we complement a difference-in-differences setup with a changes-in-changes framework. In a first step, we lay out and validate the identifying assumptions of the difference-in-differences setup.

4.2.1 Difference-in-Differences

Based on the difference-in-difference approach, we compare the change in average achievement of student cohorts exposed to the intervention in Lower Saxony with the change in average achievement of student cohorts in the control group made up of several German states. In the following we refer to Lower Saxony as the reform group in contrast to the control group. The treatment is defined as being tracked after 4 years of primary school. According to the reform, school tracking in Lower Saxony was prepone from grade 7 to grade 5 in 2004, thus students in the reform group are treated in 2009 and untreated in previous years. In contrast, all control states track their students after 4 years of primary schooling, such that students are treated over the entire sample period. In this respect, our setup departs from standard difference-in-
difference settings leading to an adaptation of the usual common-trend assumption as well, which is described below. In particular, we run the following OLS regression using observations from 2006 and 2009:

\[ Y_{i,t} = \alpha + \gamma g_i + \tau t + \pi I_{i,t} + \beta X_{i,t} + \eta_i \]  

(4)

\( Y_{i,t} \) is the reading score of student \( i \) in time \( t \). \( g_i \) is an indicator that equals one if a student belongs to the reform group, i.e. Lower Saxony. \( I_{i,t} \) is an indicator that equals one whenever the treatment is in place. Thus, it equals one for the control group in both periods and for the reform group in 2009. \( X_{i,t} \) denotes a vector of controls. The average treatment effect \( \pi \) of early school tracking on students’ performance is identified under the following set of assumptions:  

**Single Treatment Assumption**

According to the single treatment assumption, coinciding with the respective reform package in 2004, no further educational reforms were put in place asymmetrically affecting PISA test scores in the treatment and control groups between grade 4 and 9. Coinciding reforms would make it difficult to decompose the effects of different interventions. In light of the single treatment assumption, we carefully studied educational reforms which were put in place in Lower Saxony and the control group defined above. In Lower Saxony no further reforms were imposed, affecting student cohorts between grade 4 and 9 in the time period in question. Although part of the reform package was a shortening of the “Gymnasium” from 9 to 8 years, we are confident that there is no effect of this reform on the students in the 9th grade. Starting in 2006, Lower Saxony requested centralized examinations upon school completion on all tracks. This does not affect our identification strategy either, because students at \( t = 0 \) were already exposed to this reform. However, the additional reform package in 2006 makes it harder to test for common trends, as shown below. Neither were any reforms implemented in the control group, affecting the students differently in 2006 and 2009. Therefore, we are convinced that the single treatment assumption holds in our case.

**Common Trend Assumption**

The common trend assumption in our application differs slightly from standard difference-in-difference framework because our control group is treated in both periods while the reform group is treated only in the second period. Formally, let \( g \in \{0,1\} \) denote a regional dummy which equals 1 for the reform group (Lower Saxony) and 0 for a control group made up of all former western German states besides Saarland and Bremen, whilst \( t \in \{0,1\} \) is a time dummy which equals 1 for the student cohorts post reform (2009) and 0 prior to the reform (2006). Note that in our case the control group is exposed to the treatment (primary schooling of

\footnotemark[12]The difference-in-difference approach generally identifies the average treatment effect on the treated. Estimation by OLS imposes homogeneity of the effect across groups.
four years) in both periods (2006-2009) while the treatment group is exposed to the treatment only in 2009. Let $Y^N_{g,t}$ denote the potential outcome of an individual at time $t$ who belongs to group $g$ not exposed to the treatment and $Y^I_{g,t}$ the potential outcome of students exposed to the treatment. The average treatment effect on the treated (ATET) is defined as the expected difference between potential outcomes in the reform group:

$$ATET_t = E\left[Y^I_{1,t} - Y^N_{1,t}\right]$$ \hfill (5)

Naturally, we only observe students in either state, thus either they are tracked after four or six years of schooling. Consequently, we do not measure both outcomes for each student. Rather, we observe the treated outcome in the treatment group at $t = 1$, $E\left[Y^I_{1,1}\right]$, and the untreated outcome at $t = 0$, $E\left[Y^N_{1,0}\right]$. However, the counterfactuals $Y^I_{1,0}$ and $Y^N_{1,1}$ are unknown for the treatment group. A strategy that allows us to estimate the ATET is the difference-in-differences setup. Let us express the $ATET_1$ in the following way:

$$ATET_1 = E\left[Y^I_{1,1} - Y^N_{1,1}\right] = E\left[Y^I_{1,1}\right] - E\left[Y^N_{1,1} - Y^N_{1,0}\right] - E\left[Y^N_{1,0}\right]$$ \hfill (6)

The only unknown part of equation 6 is $E\left[Y^N_{1,1} - Y^N_{1,0}\right]$, which is the expected change in potential untreated outcome over the two periods in the reform group. However, we observe the change in potential treated outcome for the control group. Under the assumption that the expected change in the potential untreated outcome is the same in both groups and the treatment effect is time constant in the control group, we can rewrite equation 6 as follows:

$$ATET_0 = E\left[Y^I_{1,0} - Y^N_{1,0}\right] = E\left[Y^I_{1,1}\right] - E\left[Y^I_{0,1} - Y^I_{0,0}\right] - E\left[Y^N_{1,0}\right]$$ \hfill (7)

Equation 7 only depends on elements that can be estimated by observed outcomes. Thus, our identification additionally rests on the assumption that the treatment effect is time constant in the control group. However, the common trend assumption and the time constant treatment effect assumption cannot be tested separately in our setting. However, both together imply common trends in the untreated potential outcomes in the reform and the treated potential outcomes in the control group. This can be tested by applying placebo difference in difference analyses to previous periods. Yet, finding no effect could also be driven by opposite changes in the treatment effect and the common trends. Descriptively, the panel on the left hand side of figure 7 depicts trends of average achievement scores for ninth graders which are in fact parallel. Although we cannot use the 15-year-olds for our analyses because there is no respective sample in 2009, they help us to analyze the trends as they are available also in 2003. Visually, the trends do not appear parallel for the 15-year-olds from 2000 to 2003. Hence, we conduct a placebo difference-in-differences analysis using each sample.
The results of these placebo difference-in-differences analyses are presented in table 5. They are mainly in line with the figures above; prior to 2009 the results do not depict any major and significant deviation in the achievement trends between the treatment and control group for ninth graders. The significant effects in the 15-year-olds’ sample vanishes when covariates are included. Thus, we are confident that the pretrends are indeed parallel for our main sample, the 9th grade students. Yet, this is still a test for the implied parallel trend which does not allow us to separately address the parallel trend assumption in the untreated outcomes and the time constant effect assumption.

13All estimation results use the provided population weights.
Table 5: Pretrend difference-in-differences

<table>
<thead>
<tr>
<th></th>
<th>9th Grade</th>
<th>15-Year-Old</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Controls</td>
<td>Controls</td>
</tr>
<tr>
<td>Time 2003 x Ref. Group</td>
<td></td>
<td>-3.659***</td>
</tr>
<tr>
<td></td>
<td>(0.902)</td>
<td>(4.328)</td>
</tr>
<tr>
<td>Time 2006 x Ref. Group</td>
<td>-1.737</td>
<td>-5.336</td>
</tr>
<tr>
<td></td>
<td>(18.599)</td>
<td>(4.556)</td>
</tr>
<tr>
<td>Reform-Group</td>
<td>-10.482</td>
<td>-5.211</td>
</tr>
<tr>
<td></td>
<td>(9.525)</td>
<td>(3.438)</td>
</tr>
<tr>
<td>Time 2003</td>
<td>1.377</td>
<td>-2.319</td>
</tr>
<tr>
<td></td>
<td>(5.001)</td>
<td>(2.065)</td>
</tr>
<tr>
<td>Time 2006</td>
<td>12.800***</td>
<td>8.442***</td>
</tr>
<tr>
<td></td>
<td>(4.541)</td>
<td>(1.994)</td>
</tr>
<tr>
<td>Const.</td>
<td>501.439***</td>
<td>646.532***</td>
</tr>
<tr>
<td></td>
<td>(3.292)</td>
<td>(15.800)</td>
</tr>
<tr>
<td>Male</td>
<td>-16.276***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.032)</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-1.161***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.082)</td>
<td></td>
</tr>
<tr>
<td>RS</td>
<td>88.110***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.445)</td>
<td></td>
</tr>
<tr>
<td>GY</td>
<td>150.403***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.395)</td>
<td></td>
</tr>
<tr>
<td>No. Obs.</td>
<td>32,436</td>
<td>32,436</td>
</tr>
<tr>
<td>R2</td>
<td>0.49</td>
<td></td>
</tr>
</tbody>
</table>

Notes: The table reports difference-in-differences estimates for pre-treatment deviations in reading test scores between the reform group and the control group. Reform group: Lower Saxony. Control group: Baden-Württemberg, Bavaria, Hamburg, Hessen, North Rhine-Westphalia, Rhineland-Palatinate and Schleswig-Holstein. All estimations obtained by using population weights. Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01.
Stable Unit Treatment Value Assumption

The stable unit treatment value assumption (SUTVA) states that only one of the potential outcomes is observed: either the potential outcome when treated or the potential outcome when untreated. This rules out cases where the treatment of a subpopulation unleashes spillover effects on the untreated population. In our case this assumption is very likely to hold since all students in Lower-Saxony tested in 2009 are treated and all students in the control group are untreated. But as students were tested with some delay in grade 9, some of them might have moved to another state between the treatment and the test. This could in general generate spillover effects on students in other states. Interstate movements between the treatment and the test would bias our estimates towards zero because both groups would seem to change more equally. In light of the fact that annual in and out migration flows of Lower Saxony within Germany are both lower than 1.5% it is likely that this assumption holds.\textsuperscript{14}

Absence of Anticipation Effects

This assumption is needed to rule out that the students or teachers in Lower Saxony responded prior to the reform, generating effects that are not caused by the treatment but by this prior change in behavior. The new government that implemented the reform came into power in 2003. Thus, there was just one year prior to the reform in which students or teachers could have known of this reform. In 2003, our cohort of interest attended the fourth grade of primary school. Although the curricula were not affected, some might have postponed the treatment by repeating grade 4 voluntarily in 2004/05. Also teachers might have forced more students to repeat. Therefore, figure 8 presents the fraction of repeaters in grade four for the treatment and the control group. According to the figure, we can rule out effects of the school reform on the share of repeaters in the year 2004/05. Additionally, teachers might have changed their effort to prepare students better for the preponed tracking. We cannot entirely rule this issue out but are confident that this is a minor one since we use the first completely treated cohort which already attended grade 4 when the reform was put in place. Thus, the major part of those students’ primary school time was clearly unaffected.

\textsuperscript{14}Source: Statistische Ämter des Bundes und der Länder
**Figure 8:** Share of repetitions in grade 4

Notes: Share of repeaters as percentages of total students attending grade 4. Source: Federal Statistical Office.

*Exogeneity of Controls Assumption*

There are two reasons why one wants to include controls to the OLS estimation. First, if covariates asymmetrically affecting the outcome in the reform and control group, and thereby undermining the common-trend assumption. These could also be exogenous composition effects. Second, they can help getting more efficient estimators of the effect of interest. Yet, in both cases the set of covariates must be exogenous. Thus, the covariates themselves must be unaffected by the reform. In our baseline specification, we do not add any control variables. For robustness checks we run a specification with the following control variables: an indicator for being male, an indicator, for being in the middle track (RS), an indicator for being in the upper track (GY), and the age of students in months. While these variables are mainly exogenous for the individual, they might be endogenous considering the reform due to composition effects. The reform might have changed the composition of students attending different tracks which could affect all controls mentioned. Since we do not use any controls in the baseline specification, these results will not depend on this assumption.

*Common Support Assumption*

This assumption states that the support of the control variables must be the same in the reform and the control group. Since all school types exist in the respective groups and there is no specific age cut-off between both groups, we can be sure that this assumption holds in our
After discussing the identifying assumptions of the difference-in-differences setup, we proceed with the changes-in-changes framework.

4.2.2 Changes-in-Changes

The disadvantage of the difference-in-differences approach is that it measures the average treatment effect (of the treated). Thus, it exclusively captures an average of the achievement effect over the treatment group. However, as shown in the theoretical section, we expect a positive effect of the earlier tracking on high-skilled students, while we expect the effect for low-skilled students to be lower or even negative. In an extreme scenario, if students in the upper tail of the skill distribution are positively affected and students in the lower tail of the skill distribution are impaired through earlier school tracking, these effects might fully compensate for each other. The changes-in-changes approach proposed by Athey and Imbens (2006) allows us to estimate such heterogenous effects in a difference-in-difference framework. In particular, the changes-in-changes approach estimates the effect at each percentile of the skill distribution. The effect at the p-th percentile is defined as the difference between the p-th percentile of the potential outcome distribution of the treated and the potential outcome distribution of the untreated:

\[
\Delta_{g,t}^{CiC}(p) = F_{I_{g,t}}^{-1}(p) - F_{N_{g,t}}^{-1}(p) \tag{8}
\]

While we observe the post-reform distribution of the potential treated outcome in Lower Saxony, \(F_{Y_{1,1}}\), the post-reform distribution of the potential untreated outcome, \(F_{Y_{1,0}}\), is unknown. Following Athey and Imbens (2006), we can estimate the counterfactual distribution, and therefore the quantile treatment effect in the following way:

\[
\Delta_{1,1}^{CiC}(p) = F_{Y_{1,1}}^{-1}(p) - F_{Y_{1,0}}^{-1} \left( F_{Y_{1,0}}^{-1}(p) \right) \tag{9}
\]

Deriving this counterfactual distribution involves three steps: Firstly, we make use of the observed pre-reform distributional outcome of the reform group in order to calculate the p-th percentile, \(F_{Y_{1,1}}^{-1}(p)\). Secondly, we determine the percentile \(p'\) corresponding with this score, using the observed distribution of the pre-reform period of the control group. Thirdly, we calculate \(F_{Y_{1,0}}^{-1}(p')\), which is the \(p'\)-th percentile of the post-reform period distribution of the control group. Figure 9 illustrates this estimation strategy graphically.
Figure 9: Changes-in-changes identification

Notes: This figure illustrates the iterative identification strategy of the changes-in-changes approach.

Note that unlike Athey and Imbens (2006), we posit that the change in the distribution of the potential treated outcome of the control group is the same as the change in the distribution of the potential untreated outcome in the reform group. However, the basic procedure proposed by Athey and Imbens (2006) dispenses with confounding factors as they base their framework on marginal distributions rather than conditional distributions on $X$. Melly and Santangelo (2015) complementarily build upon the changes-in-changes method and allow for covariates as well. Again, we have to control for covariates asymmetrically impinging on the outcomes in the treatment and control groups and we might want to control for further relevant covariates for the sake of efficiency.

As the estimator proposed by Melly and Santangelo (2015) has not been applied frequently, we briefly outline the estimation strategy. First, we estimate the conditional quantile regression processes, making use of the estimator proposed by Koenker and Bassett (1978) for the reform
and control group in the two periods using the observed outcome variable along with covariates:

\[
\hat{\beta}_{g,t}(u) = \arg \min_{b \in \mathbb{R}^{k+1}} \sum_{i : G_i = g, T_i = t} (u - 1(Y_i \leq X_i'b))(Y_i - X_i'b)
\]

We wind up with conditional quantile functions for all four samples, which can be transformed into the respective estimator in the following way:

\[
\hat{F}_{Y_{g,t,x}}(y) = \int_0^1 1(x'\hat{\beta}_{g,t}(u) \leq y) \, du
\]

where \(1(\cdot)\) is an indicator that equals 1 when the included expression is true. In practice, it is not feasible to estimate the conditional distributions for all quantiles, which is why we also estimate the conditional distributions for a finite mesh containing all natural number percentiles. The estimator for the unconditional distribution is derived by integrating over the group specific distribution of covariates \(F_{X,1,1}(x)\):

\[
\hat{F}_{Y_{g,t}} = \int \hat{F}_{Y_{g,t,x}} \, dF_{X,g,t}(x)
\]

The unconditional quantile treatment effect is given by:

\[
\hat{\Delta}^{QTE}_{1,1}(p) = \hat{F}_{Y_{1,1}^{-1}(p)} - \hat{F}_{Y_{1,1}^{-1}(p)}
\]

where \(\hat{F}_{Y_{1,1}^{-1}(p)}\) is not only the observed sample counterpart but also an integral over the estimated conditional distributions based on observations from Lower-Saxony in 2006. Although the results of the underlying quantile regressions might be interesting in order to predict the effect of a reform for different states, interstate disparities in the effects might not only originate from differences in the socioeconomic structure, but might also be due to differences in school systems. Further, we cannot be sure of measuring a treatment effect of the entire population. We will therefore only report conditional results for dichotomous covariates in an effort to identify the main underlying channels of our results.

The changes-in-changes method builds upon assumptions similar to those of the difference-in-difference method. Melly and Santangelo (2015) point out four additional assumptions: The functional form of the potential outcome, the strict monotonicity of this function, the time invariance assumption and the identifiability assumption. In the following we will discuss these assumptions in light of our setting:

*Potential Outcome*
The potential treated outcome can be expressed as a function of covariates, time and a random variable $U$:

$$Y_{g,t,x}^N = h(X,T,U)$$

while the functional form of $h(\cdot)$ is not restrictive and $U$ the vector of unobserved errors.

**Strict Monotonicity**

The monotonicity assumption states that $h(\cdot)$ is strictly increasing in $U$ for both time periods and all $x \in X$. Neither of the two assumptions can be tested.

**Time Invariance**

The distribution of $U$ is independent of the time period, given the group and the covariates: $U \perp T|G,X$. This assumption is the counterpart to the common trend assumption in the difference-in-differences approach. Since we have more than the two periods necessary for identification, we can test for this assumption. The following figures display placebo changes-in-changes estimates for two subsamples and different transition periods: whilst for ninth graders, we exclusively rely on the transition from 2000 to 2006 due to missing data in 2003, for 15-year-old students we complementarily perform the changes-in-changes procedure for the transition periods from 2000 to 2003, 2003 to 2006 and 2000 to 2006.

**Figure 10:** Changes-in-changes: pre-reform period results

(a) 9th grade students 2000/2006  
(b) 15 year-old students 2000/2006

Notes: Panel (a): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 9th grade students, 2000/2006, 33,073 observations. Panel (b): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 15-year-old students, 2000/2006, 30,174 observations.

Figure 10 visualizes pre-reform changes-in-changes estimates for ninth graders in the panel on the left hand side and for 15 years old students in the panel on the right hand side for the
period 2000-2006. The estimates consistently reveal an increase in the test scores for the lower tail of the achievement distribution, albeit an insignificant one. We decompose the period 2000 to 2006 into the transition from 2000 to 2003 on the left and the transition from 2003 to 2006 on the right of figure 11. Unlike the analysis below in which we exclusively rely on data for ninth graders, the sample of 15-year-old students allows us to shed light on test scores in 2003 as well. Apparently, the placebo changes-in-changes estimates indicate no major distributional effects between 2000 and 2003 but significant distributional achievement effects between 2003 and 2006; namely, the scores saw a significant increase for the lower tail of the achievement distribution.

**Figure 11:** Changes-in-changes: pre-reform period results 15 year-old students  
(a) 15 year-old students 2000/2003  
(b) 15 year-old students 2003/2006

**Notes:** Panel (a): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 15-year-old students, 2000/2003, 32,080 observations. Panel (b): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 15-year-old students, 2003/2006, 35,274 observations.

Hence, the pretrend changes-in-changes estimates above suggest that an additional reform was put in place prior to 2006, setting the stage for a catch-up in the achievement at the bottom of the distribution. In fact, centralized final examinations were imposed in the lower track, i.e. the Hauptschule, in 2006. These centralized examinations raised the pressure on both students and teachers. The latter have to keep in mind overall teaching standards and curricula as the exams are formulated by the state government, while the former encounter further pressure as well since they have to compete with their peers in the whole state. In order to check whether this reform drives the findings above, we rely on North Rhine-Westphalia to serve as a control group since it implemented centralized examinations in 2006 in the lower track as well. With North Rhine-Westphalia serving as a control group, we get the changes-in-changes estimates displayed in figure 12 and 13 for all possible transitions with respect to ninth graders and 15-year-old students, respectively.
**Figure 12:** Changes-in-changes: pre-reform period results with NRW as control group.

(a) 9th grade students 2000/2006  
(b) 15 year-old students 2000/2006

Notes: Panel (a): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 9th grade students using only North Rhine-Westphalia as control group, 2000/2006, 7,894 observations. Panel (b): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 15-year-old students using only North Rhine-Westphalia as control group, 2000/2006, 6,625 observations.

**Figure 13:** Changes-in-changes: pre-reform period results 15 year-old students with NRW

(a) 15 year-old students 2000/2003  
(b) 15 year-old students 2003/2006

Notes: Panel (a): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 15-year-old students using only North Rhine-Westphalia as control group, 2000/2003, 8,052 observations. Panel (b): Quantile treatment effects of placebo early school tracking on reading scores (plausible values 1-5) for 15-year-old students using only North Rhine-Westphalia as control group, 2003/2006, 7,625 observations.

Conspicuously, the distributional effects becoming apparent in the pretreatment periods undergo a decline or even vanish for ninth graders making up our sample when North Rhine-Westphalia is used as the control group. Hence, these heterogenous effects in fact originate
from the implementation of centralized examinations. Correspondingly, we provide pretreat-
ment changes-in-changes estimates for all German states in the appendix (see figures 27 to 30).
The results are consistent with an increase in the test scores for the lower tail due to the
introduction of the central exit exams. Note that the negative effect along almost the entire
income distribution for Rhineland-Palatinate is caused by an increase in the fraction of students
attending comprehensive schools. The effect therefore reflects a selection of an overall smaller
fraction that attends regular tracks which become part of our sample. The extension of the
comprehensive school system in Rhineland Palatinate was completed by 2006 and hence does
not undermine our results.

Identifiability
The support of the observed distribution of $U$ of the treatment group must be a subset of the
support of the distribution of $U$ of the control group: $U_{1,0,x} \subset U_{0,0,x}$ $\forall x \in X$. This assump-
tion implies that $Y_{1,0,x}^{N} \subset Y_{0,0,x}^{I}$ $\forall x \in X$, i.e. that the support of the pre-reform distributional
outcomes in the treatment group must be a subset of the pre-reform distributional outcome in
the control group. If this is not the case, we can still estimate quantile treatment effects for all
quantiles for which the condition holds. The identification is not undermined by the fact that
the effects cannot be identified for all quantiles. In our application, the condition holds for the
percentiles we consider, which range from the 1st up to the 99th percentile.

In the following section, we contrast the results of the difference-in-differences and the changes-
in-changes strategy.

4.3 Results
4.3.1 Difference-in-Differences
In the theoretical section, we predicted unambiguously positive effects of preponed stratification
in the upper tail of the achievement distribution, while the effects in the lower tail depend on the
relative sizes of the congruency and spillover effects. Without specifying these sizes exactly, we
do not know whether the positive effects in the upper tail are larger or smaller than potentially
negative effects in the lower tail. Hence, the predictions for the average achievement effect
are ambiguous. Consequently, we first perform difference-in-differences analyses as reported in
Table 6. While the specification in column 1 makes use of plausible values 1-5, specifications in
columns 2-3 are exclusively based on plausible value 1. In the baseline specification (column
1), the point estimate of the treatment effect is 1.6, though not significant at any conventional
level of significance. The same holds for the estimations exclusively based on plausible value
1 (column 2). Column 3 reports the results of the estimation including covariates. We do not
find significant estimates of the treatment effect either. Thus, we conclude that there is no
effect of preponed school stratification on average student performance. This might be the case
either because there is no effect at all at any part of the distribution or the effects compensate for each other at the mean.

Table 6: Results difference-in-differences

<table>
<thead>
<tr>
<th></th>
<th>[1]</th>
<th>[2]</th>
<th>[3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatm. Effect</td>
<td>1.649</td>
<td>3.569</td>
<td>2.677</td>
</tr>
<tr>
<td></td>
<td>(12.840)</td>
<td>(12.554)</td>
<td>(4.699)</td>
</tr>
<tr>
<td>Reform-Group</td>
<td>-12.219</td>
<td>-12.897</td>
<td>-10.505***</td>
</tr>
<tr>
<td></td>
<td>(8.224)</td>
<td>(8.010)</td>
<td>(2.968)</td>
</tr>
<tr>
<td>Time</td>
<td>-9.858**</td>
<td>-10.263**</td>
<td>-16.099***</td>
</tr>
<tr>
<td></td>
<td>(4.187)</td>
<td>(4.152)</td>
<td>(1.886)</td>
</tr>
<tr>
<td>Const.</td>
<td>514.239***</td>
<td>514.600***</td>
<td>663.509***</td>
</tr>
<tr>
<td></td>
<td>(3.131)</td>
<td>(3.089)</td>
<td>(14.436)</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td>-14.834***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.025)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td>-1.176***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.076)</td>
</tr>
<tr>
<td>RS</td>
<td></td>
<td></td>
<td>79.198***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2.400)</td>
</tr>
<tr>
<td>GY</td>
<td></td>
<td></td>
<td>141.071***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2.339)</td>
</tr>
<tr>
<td>No. Obs.</td>
<td>36,236</td>
<td>36,236</td>
<td>36,236</td>
</tr>
<tr>
<td>R2</td>
<td>0.00</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Dep. Var.</td>
<td>PV1-5</td>
<td>PV1</td>
<td>PV1</td>
</tr>
</tbody>
</table>

Notes: 9th grade students. Dependent variable: Reading achievement. Column [1] uses all 5 plausible values for reading. Columns [2] and [3] use only the first plausible value for reading. All estimations conducted using population weights. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The consistency of the results through specifications with and without covariates in columns 2 and 3 suggests that the results are not driven by possible confounding factors like a different threshold of stratification. As pointed out previously, the role of covariates is twofold. First, they allow us to preclude confoundedness, and second, they improve the efficiency of our estimates. According to column 4 in table 6, older students on average perform worse on the test. This effect might arise from having to repeat a class, increasing the age of these test participants in grade 9. As grade repeaters are negatively selected, it is the selectivity rather than the age which negatively impinges on test scores. Students whose parents earned a Gymnasium diploma, the Abitur, perform better on average as well, which is in line with
the prediction that educational attainment is at least partially intergenerationally transmitted. Whether educational attainment is transmitted through initial ability, peer effects at home or even self-selection effects into the upper track is analysed more thoroughly in the next section. Furthermore, boys perform worse on average than girls in the Pisa and IQB test tasks. In general, focusing on reading test scores might privilege females who show stronger competencies in languages ((Marks, 2008)). Overall, the direction of the effects of the covariates point in the expected direction. Since the treatment effect remains insignificant although the standard error decreased significantly from column 2 to 3, we conclude that the covariates considered are unlikely to be confounding but efficiency enhancing.

It is worth mentioning that the fuzziness of school tracking is not affected in the course of the reform in light of table 9 in the appendix. The table presents the treatment effects of preponed tracking on being in the respective track for 9th grade students and tests for differences in this effect for covariates. Apparently, the reform did not significantly affect the composition of students in the different tracks.

We will make further use of these findings below, when we condition our changes-in-changes framework on gender and migration histories.

4.3.2 Changes-in-Changes

The results of the differences-in-differences model in the previous section depicted insignificant effects of preponed school tracking on average student performances. If students in the upper tail of the skill distribution are positively affected and students in the lower tail of the skill distribution are impaired through earlier school tracking, these effects might fully compensate for each other. In order to decompose the effects of preponed tracking on an individual level, we henceforth make use of a changes-in-changes model. With respect to the changes-in-changes framework, we base our analysis on two specifications, dispensing with covariates in the sense of Athey and Imbens (2006) and accounting for covariates in the sense of Melly and Santangelo (2015).

Figure 14 below depicts the results of the changes-in-changes estimations without covariates. Apparently, preponed tracking elicits a decline in reading comprehension in the lower tail of the achievement distribution. However, only estimates for percentiles lower than 5 are significant at the 5% level\textsuperscript{15}. Conversely, for students above the 50th percentile, we find positive effects of preponed tracking that are significant at the 5% level for students between the 50th and 97th percentile.

\textsuperscript{15}All standard errors in this section are bootstrapped standard errors from 500 draws.
Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students, 2006/2009, 36,236 observations.

These results are totally in line with the predictions we raised in the theoretical section. Students in the upper tail do not experience a change in positive peer group spillover effects in the course of a transition from a comprehensive school to a stratified system, as the upper track comprises the best students prior to and post of the intervention while they experience a positive congruency effect. Thus, the overall effect of the reform should be positive. Students in the lower tail experience strictly negative peer-group effects due to the loss of the spillovers but also experience positive congruency effects. In light of our model, the negative effect due to the loss of spillovers seems to outweigh the positive congruency effect for students in the lower tail. Note that the effect we estimate is in 2009 test units, which does not make a difference in our case since the scores are normalised with a mean of 500 and a standard deviation of 100. Thus, the increase of about 25 units for the top students means an increase of about a quarter standard deviation. The results suggest that the insignificant difference-in-differences estimates are exactly due to the compensation of the positive and negative effects in the upper and lower tail of the distribution. Hence, exclusively relying on a standard difference-in-difference strategy in our context would dispense with the quantile effects revealed above.

In order to test whether our results are caused by changes in the treatment group and not by suspicious changes in the control group, we conduct several placebo tests. In a first step, we randomly assign all students in our sample to a federal state. We should not get any significant results since the actual treatment assignment should be random with respect to the artificial treatment. Figure 15 reports the results of the changes-in-changes estimates for this placebo
The estimates are not significantly different from zero at the 5% level except for the 1st percentile. However, this finding might still be accidental, considering a type 1 error probability of 5%. If there is a systematic effect causing these findings in the control group as well, we should be able to detect it by considering each state of the initial control group as treated and using the rest of the initial control group as a final control group in a placebo changes-in-changes analysis. Figures 31 and 32 in the appendix show that this is not systematically the case across the control group states. Only for Hamburg we find positive results for the 1st percentile. All other estimates are insignificant. As pointed out previously, conditional on common pretreatment trends, the results have to be insensitive with respect to shifts in the composition of the control group. In order to test this main assumption, we report results of changes-in-changes estimates separately using each of the control states as the control group. Figures 33 and 34 in the appendix visualise these results. We consistently find the same pattern as in the baseline results for all single states. Some of the results are not significant, which is obviously due to the limited sample sizes. Overall, the placebo tests make us confident that the effects we find are not driven by suspicious effects in the control group.

**Figure 15:** Placebo changes-in-changes

*Notes:* Placebo quantile treatment effects of randomly assigned early school tracking on reading scores (plausible values 1-5) for 9th grade students, 2006/2009, 36,236 observations.

In order to examine whether reverse effects in the tails are in fact driven by the lower track and the upper track in line with our theoretical predictions, we decompose the changes-in-changes setup for the lower track (HS), the middle track (RS) and the upper track (GY). In fact, the identification rests on the additional identifying assumption that the sorting effects are unaffected by the preponed tracking, i.e. if there is a change it must evolve in the same way
as in the control group. This is the case, as table 8 in the appendix indicates. In line with the theoretical predictions, according to figure 16, students in the upper track are unambiguously better off while the bottom 50 percent of the lower track are significantly worse off. Students in the middle track are rather insensitive to the reform. These findings are in line with our model. In light of our setup, we would expect negative results if students in the middle track had experienced significant spillover effects from top students in the orientation stage. This seems not to be the case, which suggests that the spillover effects are extremely non-linear.

**Figure 16:** Changes-in-changes results by school form  
(a) Lower track (HS)  
(b) Middle track (RS)  
(c) Upper track (GY)

*Notes:* Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by school form, 2006/2009, Observations:  
(a) 9,872  
(b) 12,941  
(c) 13,423.

In order to gain further insights and for the sake of efficiency, we complementarily make use of plausible value 1 and provide changes-in-changes estimates under consideration of covariates visualised in the left panel of figure 17. In line with the difference-in-differences estimates in the

\footnote{Our data only allow us to separate the students according to their current track. We basically assume that the students did not change tracks since grade 5. This seems reasonable because changes from one track to another are quite rare in Germany.}
previous section, these covariates entail gender disparities through a dummy, which is 1 for male students and 0 otherwise, as well as the paternal educational degree through a dummy, which is 1 if at least one of the parents earned the Abitur and 0 otherwise. However, the efficiency in our case also depends on the use of plausible values. Therefore, the panel on the right side of figure 17 complementarily depicts the baseline result exclusively making use of plausible value 1. Obviously, relying on plausible value 1 leads to minor gains in terms of efficiency. Qualitatively, controlling for covariates is essentially neutral with respect to distributional effects. This is not surprising because the tracking policy did not change differently in Lower Saxony than in the control group and all other controls are unlikely to react to the reform. Since the smaller confidence bands are caused by only using plausible value 1 and there seem to be no confounding factors, the remainder of the estimations are conducted without controls and all plausible values. This enables us to further investigate whether different students react differently to the reform. Particularly, we conduct the estimates separately for males and females, students who speak a foreign language at home and students who speak German at home as well as for students whose parents have an Abitur and those who do not.

**Figure 17:** Changes-in-changes results with controls

(a) PV1 with controls

(b) PV1 without controls

Notes: Panel (a): Quantile treatment effects of early school tracking on reading scores (plausible value 1 only) with controls (male, age, rs, gy) for 9th grade students, 2006/2009, 36,236 observations. Panel (b): Quantile treatment effects of early school tracking on reading scores (plausible value 1 only) without controls for 9th grade students, 2006/2009, 36,236 observations.

With respect to gender disparities, as the focus was laid on reading rather than math and science test exercises in 2012, the test might privilege women which, in combination with sorting effects, might set the stage for gender disparities as a consequence of preponed school tracking. In contrast to the previous changes-in-changes setup, in figure 18 we account for male (left-hand side) and female students (right-hand side) in the reform and control groups, respectively. The results suggest that women respond less sensitively to the reform compared to their male peers at both tails of the distribution. Although the point estimates differ, these differences
are insignificant. The heterogenous results have three possible explanations. First, comparable male and female students react differently to the reform. Second, the initial distribution is different for girls and boys, which would lead to different effects in light of our model. Third, male students are overrepresented in the lower and upper tracks.

**Figure 18:** Changes-in-changes results by gender

(a) Males  
(b) Females

*Notes:* Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by gender, 2006/2009, Observations: (a) 18,333 (b) 17,903.
Concerning point 2, the upper left panel of figure 19 shows the difference in the distributions of test results between male and female students prior to the reform\textsuperscript{17}. The difference estimates are derived by running quantile regressions on a constant and the male indicator for the respective percentiles. Apparently, women excel in reading comprehension such that female test scores first order stochastically dominate those of the males. If this was also true for the initial distributions in grade 4, it could explain the more negative findings for males in the lower tail, who lose more spillovers according to our model. However, it cannot explain the findings in the upper tail of the distribution. Hence, we focus on the third explanation. Indeed, males are overrepresented in the lower track but underrepresented in the upper track. We therefore account for gender disparities in the distributional effects separately for all three tracks shown in figure 20. The panels depict the estimates for all students together as well as male and female students separately for the respective track, though no confidence bands are shown due to the

\textsuperscript{17}Ideally, we would want to compare initial distributions, i.e. achievement distributions at the end of grade 4 for both groups. However, there is no data on these distributions for our test students.
insignificant differences mentioned above. The results for the lower track still differ greatly, which might be due to the first order stochastic dominance that also exists in the lower track (see upper right panel of figure 19) for girls although they are underrepresented. For the other tracks the results are very similar except for the tails of the distribution in the upper track. There is no explanation in line with our model for these findings except that male students in the upper track in fact experience different gains from congruency of teaching than do female students.

**Figure 20:** Changes-in-changes results by gender and school form

(a) Lower track (HS) (b) Middle track (RS) (c) Upper track (GY)

Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by gender and school form, 2006/2009. Observations: (a) Male 5,547; Female 4,325 (b) Male 6,392; Female 6,549 (c) Male 6,394; Female 7,029.

Students with migration backgrounds disproportionately sort into the the lower track and their initial German language proficiency is lower on average compared to their native peers, independently of the reform.\(^\text{18}\) Due to lower initial reading comprehension combined with

\(^{18}\)Lower initial proficiency might lead to a catching-up effect as well.
disproportional sorting effects into the lower track, we expect migrants to be particularly worse off in the course of the reform compared to natives. Therefore, we repeated our analysis for students who speak German at home and students who speak a foreign language at home separately. The results are reported in figure 21. Conspicuously, the performance effects of students speaking foreign languages at home are magnified in both tails of the distribution in the course of the reform. This in fact points to severe sorting effects of migrants into the lower track. Nevertheless, the number of observations is very small, leading to wide confidence bands and insignificant estimates.

**Figure 21**: Changes-in-changes results by language at home  
(a) Language at home German  
(b) Language at home not German

*Notes:* Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by language spoken at home, 2006/2009, Observations: (a) 28,662 (b) 3,249.

In order to verify whether the results are in fact driven by migrants predominantly sorting into the lower track, we decompose the achievement effects for the language spoken at home for all tracks in figure 22. Consistently, the results for the lower track suggest that most of the observed differences indeed originate from disproportional sorting effects of migrants into the lower track. The extremely similar findings in the lower track are surprising because in all three tracks migrants perform worse over the entire distribution, as shown in figure 23. Conversely, in the middle track, the achievement of students who speak a foreign language at home decreases between the 5th and 80th percentile. This might indicate that this group is particularly sensitive to the loss of spillovers. The migration-background students at the top of the distribution react more positively than their native-speaker counterparts. Card and Giuliano (2016) also find these strong positive effects for high-skilled migrants. Together with the surprisingly similar findings for the students in the lower track, this leads to the conclusion that students with a migration background are also particularly sensitive to the congruency of teaching.
Figure 22: Changes-in-changes results by language at home and school form
(a) Lower track (HS)  (b) Middle track (RS)  (c) Upper track (GY)

Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by language at home and school form, 2006/2009. Observations: (a) German 6,598; Not German 1,475 (b) German 10,402; Not German 1,060 (c) German 11,662; Not German 714.
Figure 23: Distributional differences by language at home

(a) All
(b) Lower track (HS)
(c) Middle track (RS)
(d) Upper track (GY)

Notes: Differences in percentiles of the distributions of reading scores (plausible values 1-5) of students who speak German at home and those that do not speak German at home by school form. Results from quantile regressions with a constant and the dummy speaking German at home. Foreign language speakers at home are the reference category. 9th grade students, 2006.

Observations: (a) 17,133 (b) 4,420 (c) 6,394 (d) 6,319.
Figure 24: Changes-in-changes results by parents’ education

(a) Parents with Abitur

(b) Parents without Abitur

Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by parents having an Abitur or not, 2006/2009, Observations: (a) 10,234 (b) 15,753.

Performing the changes-in-changes procedure separately for parents with and without the Abitur, reveals the distributional effects displayed in figure 24. Apparently, the results show that students whose parents earned an Abitur almost all experience positive achievement effects as a result of the reform. However, since students whose parents completed the upper track are more likely to complete the upper track as well, it is not possible to disentangle the effects of initial ability, positive peer effects at home and disproportionate sorting effects into the Gymnasium based on this figure. Hence, again we decompose the analysis for the three tracks and obtain figure 25.
Figure 25: Changes-in-changes results by parents’ education and school form

(a) Lower track (HS)

(b) Middle track (RS)

(c) Upper track (GY)

Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by parents’ education and school form, 2006/2009. Observations: (a) with Abitur 1,265; without Abitur 5,709 (b) with Abitur 2,939; without Abitur 6,313 (c) with Abitur 6,030; without Abitur 3,731.

In figure 25, we compare the change in reading comprehension of ninth graders between 2006 and 2009 in the treatment and control group over the entire distribution. Obviously, the positive effect on the achievement of students with parents holding an Abitur are mainly due to sorting effects into the upper track. The above-median students with parents with Abitur in the middle track react positively to the reform. This cannot be explained by absolute differences in the performance in the middle track, as panel (c) of figure 26 demonstrates.
Figure 26: Distributional differences by parents’ education

(a) All

(b) Lower track (HS)

(c) Middle track (RS)

(d) Upper track (GY)

Notes: Differences in percentiles of the distributions of reading scores (plausible values 1-5) of students with parents with Abitur and students with parents without Abitur by school form. Results from quantile regressions with a constant and the dummy for parents having an Abitur. Students with parents without Abitur are the reference category. 9th grade students, 2006. Observations: (a) 18,264 (b) 4,834 (c) 6,843 (d) 6,587.

To conclude with respect to group-specific effects, we find stronger changes in the distribution of test scores for male students than for female students, which can partly be explained by disproportional sorting of male students into the different tracks and partly by the first order stochastic dominance of the performance of female students. Furthermore, the performance of students not speaking German at home changes more negatively for almost all percentiles. This can almost entirely be explained by the disproportional sorting of these students into the lower track. Furthermore, our results suggest that students with migration backgrounds are more sensitive to the congruency of teaching and spillover effects from peers. Additionally, we find more positive effects for students whose parents earned an Abitur. We show that this finding is almost entirely due to a disproportional sorting of these students into the upper track. Hence, group specific distributional effects seem to be caused by different treatments in the tracks and disproportionate representation of these subgroups in the tracks rather than differing responses of these subgroups to the treatment.
5 Conclusion

At the beginning, we raised the question whether school tracking has a persistent effect on the distribution of students’ performance. We were particularly interested in heterogenous effects that may occur in the lower and upper tail of the performance distribution. In order to tackle our research question, we combined a theoretical analysis with an empirical investigation.

Theoretically, we set out a simple model of human capital development, contrasting peer group spillover and teaching congruency effects. According to the latter, a lower variance of skills within the classroom is more conducive to optimization in the lower as well as the upper track. According to the former, low achievers experience positive spillover effects if they are grouped with high achievers. Hence, in the course of preponed school tracking, we expect ambiguous achievement effects in the lower track and unambiguously positive effects in the upper track.

Empirically, we relied on a differences-in-differences setup in order to isolate average effects and on a changes-in-changes setup in order to account for distributional effects of the school reform. In line with the theoretical predictions, we find negative achievement effects in the lower tail of the achievement distribution and positive effects in the upper tail of the distribution. As the effects compensate for each other, average achievement is not affected in the course of preponed school tracking. Further, we find stronger effects for males than for females. We can show that gender disparities at the lower tail are driven by the overrepresentation of males in the lower track and a first-order stochastic dominance of females’ performances. The selection into the lower track can also explain the major part of the difference between students with and without a migration background. However, there is some evidence that students with migration backgrounds might react more strongly to the reform, which suggests that they are particularly sensitive to teaching congruency and peer spillover effects.

The achievement effects of stratified schooling has attracted considerable attention among scientists and politicians for many decades. This paper emphasises the role of heterogeneity of the effects along the skill distribution. Our results suggest that politicians encounter a trade-off between optimizing high and low skilled students’ achievement through the timing of school tracking.
References


Appendix

A  Comprehensive Statistics

Table 7: Descriptive Statistics: controls. 15-years-old students.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Year</th>
<th>No. Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Year</th>
<th>No. Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Year</th>
<th>No. Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>2000</td>
<td>13784</td>
<td>0.48</td>
<td>0.6</td>
<td>2003</td>
<td>18351</td>
<td>0.5</td>
<td>0.5</td>
<td>2006</td>
<td>15939</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>18351</td>
<td>0.5</td>
<td>0.5</td>
<td>2003</td>
<td>18351</td>
<td>0.5</td>
<td>0.5</td>
<td>2006</td>
<td>15939</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>15939</td>
<td>0.5</td>
<td>0.5</td>
<td>2006</td>
<td>15939</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>2000</td>
<td>13784</td>
<td>189</td>
<td>4</td>
<td>2003</td>
<td>18351</td>
<td>189</td>
<td>3</td>
<td>2006</td>
<td>15939</td>
<td>190</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>18351</td>
<td>189</td>
<td>3</td>
<td>2003</td>
<td>18351</td>
<td>189</td>
<td>3</td>
<td>2006</td>
<td>15939</td>
<td>190</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>15939</td>
<td>190</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower track (HS)</td>
<td>2000</td>
<td>13784</td>
<td>0.29</td>
<td>0.45</td>
<td>2003</td>
<td>18351</td>
<td>0.29</td>
<td>0.46</td>
<td>2006</td>
<td>15939</td>
<td>0.27</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>18351</td>
<td>0.29</td>
<td>0.46</td>
<td>2003</td>
<td>18351</td>
<td>0.29</td>
<td>0.46</td>
<td>2006</td>
<td>15939</td>
<td>0.27</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>15939</td>
<td>0.27</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle track (RS)</td>
<td>2000</td>
<td>13784</td>
<td>0.35</td>
<td>0.48</td>
<td>2003</td>
<td>18351</td>
<td>0.34</td>
<td>0.47</td>
<td>2006</td>
<td>15939</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>18351</td>
<td>0.34</td>
<td>0.47</td>
<td>2003</td>
<td>18351</td>
<td>0.34</td>
<td>0.47</td>
<td>2006</td>
<td>15939</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>15939</td>
<td>0.37</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper track (GY)</td>
<td>2000</td>
<td>13784</td>
<td>0.37</td>
<td>0.48</td>
<td>2003</td>
<td>18351</td>
<td>0.37</td>
<td>0.48</td>
<td>2006</td>
<td>15939</td>
<td>0.36</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>18351</td>
<td>0.37</td>
<td>0.48</td>
<td>2003</td>
<td>18351</td>
<td>0.37</td>
<td>0.48</td>
<td>2006</td>
<td>15939</td>
<td>0.36</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>15939</td>
<td>0.36</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home Language</td>
<td>2000</td>
<td>12258</td>
<td>0.91</td>
<td>0.28</td>
<td>2003</td>
<td>16252</td>
<td>0.86</td>
<td>0.35</td>
<td>2006</td>
<td>14322</td>
<td>0.89</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>16252</td>
<td>0.86</td>
<td>0.35</td>
<td>2003</td>
<td>16252</td>
<td>0.86</td>
<td>0.35</td>
<td>2006</td>
<td>14322</td>
<td>0.89</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>14322</td>
<td>0.89</td>
<td>0.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parents Abitur</td>
<td>2000</td>
<td>12529</td>
<td>0.32</td>
<td>0.47</td>
<td>2003</td>
<td>16311</td>
<td>0.38</td>
<td>0.48</td>
<td>2006</td>
<td>15258</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>16311</td>
<td>0.38</td>
<td>0.48</td>
<td>2003</td>
<td>16311</td>
<td>0.38</td>
<td>0.48</td>
<td>2006</td>
<td>15258</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>15258</td>
<td>0.41</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Fraction of students in respective track with weights. 9th grade students.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower Saxony</td>
<td>HS</td>
<td>0.29</td>
<td>0.24</td>
<td>-0.05</td>
<td>0.29</td>
<td>0.24</td>
<td>-0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>0.39</td>
<td>0.38</td>
<td>-0.01</td>
<td>0.39</td>
<td>0.39</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GY</td>
<td>0.33</td>
<td>0.38</td>
<td>0.06</td>
<td>0.33</td>
<td>0.37</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>HS</td>
<td>0.29</td>
<td>0.25</td>
<td>-0.04</td>
<td>0.33</td>
<td>0.29</td>
<td>-0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RS</td>
<td>0.33</td>
<td>0.33</td>
<td>0</td>
<td>0.33</td>
<td>0.34</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GY</td>
<td>0.38</td>
<td>0.42</td>
<td>0.04</td>
<td>0.34</td>
<td>0.36</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Source: Federal Statistical Office.
B Sorting Tests

Table 9: Sorting Tests

<table>
<thead>
<tr>
<th></th>
<th>HS</th>
<th>RS</th>
<th>GY</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>-0.0092</td>
<td>-0.0068</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>(0.0799)</td>
<td>(0.0973)</td>
<td>(0.0953)</td>
</tr>
<tr>
<td>Male</td>
<td>0.0052</td>
<td>-0.0543</td>
<td>0.0491</td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
<td>(0.046)</td>
<td>(0.0474)</td>
</tr>
<tr>
<td>Home Language</td>
<td>-0.2062*</td>
<td>0.0852</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td>(0.1091)</td>
<td>(0.1089)</td>
<td>(0.093)</td>
</tr>
<tr>
<td>Parents Abitur</td>
<td>0.017</td>
<td>-0.0594</td>
<td>0.0423</td>
</tr>
<tr>
<td></td>
<td>(0.0747)</td>
<td>(0.0807)</td>
<td>(0.0704)</td>
</tr>
</tbody>
</table>

Notes: This table presents the treatment effects on being in the lower (middle/upper) track in the first row for 9th grade students. The following rows test for differences in this effect for males/females (Male), students that speak German at home and not (Home Language) and for students whose parents have a highschool degree and whose do not (Parents Abitur). All estimations conducted using population weights. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 

54
C State-level Pretrend Tests

Figure 27: Changes-in-changes: pre-reform period results separate states

Rhineland-Palatinate

Schleswig-Holstein

Hamburg

Bavaria

Baden-Württemberg

Hessia

Figure 28: Changes-in-changes: pre-reform period results separate states 2

Figure 29: Changes-in-changes: pre-reform period results separate states

Figure 30: Changes-in-changes: pre-reform period results separate states

Figure 31: Changes-in-changes: placebo results separate states 1

9th grade students, 2006/2009. Reform group: respective state; control group: other states of control group.
Figure 32: Changes-in-changes: placebo results separate states 2

9th grade students, 2006/2009. Reform group: respective state; control group: other states of control group.
Figure 33: Changes-in-changes: results separate states 1

Figure 34: Changes-in-changes: results separate states 2

The Center for Regional Economic Development (CRED) is an interdisciplinary hub for the scientific analysis of questions of regional economic development. The Center encompasses an association of scientists dedicated to examining regional development from an economic, geographic and business perspective.

Contact of the authors:

Marcus Roller
University of Bern
Schanzeneckstrasse 1
P.O.Box
CH-3001 Bern
Telephone: +41 31 631 49 97
Email: marcus.roller@vwi.unibe.ch

Daniel Steinberg
University of Tübingen
Melanchthonstr. 30
D-72074 Tübingen. E-mail:
Telephone: +49 7071 29 77068
Email: Daniel.Steinberg@uni-tuebingen.de.

This paper can be downloaded at:

http://www.cred.unibe.ch/forschung/publikationen/cred_research_papers/index_ger.html