Universität Basel
Wirtschaftswissenschaftliche Fakultät

WWZ

December 2017

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

WWZ Working Paper 2017/20

Marcus Roller, Daniel Steinberg

A publication of the Center of Business and Economics (WWZ), University of Basel.

Universität Basel
Peter Merian-Weg 6
4052 Basel, Switzerland
wwz.unibas.ch

Corresponding Author:
Dr. Marcus Roller
Tel: +41 31 631 49 97
Mail: marcus.roller@vwi.unibe.ch

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

Marcus Roller* Daniel Steinberg†

Abstract

The effects of early school stratification on scholastic performance have been subject to controversial debates in educational policy and science. We exploit a unique variation in Lower Saxony, Germany, where performance based tracking was preponed from grade 7 to grade 5 in 2004. We measure the long-run effects of early school stratification on individual PISA test scores along the entire skill distriubution using the changes-in-changes estimator. Our results indicate that preponed school tracking increased test scores at the upper tail and lowered test scores at the lower tail of the skill distribution, compensating each other on average.[1]

Keywords: Analysis of Education, Education and Inequality, Tracking, Government Policy
JEL-Codes: I21, I24, I28

*University of Basel, Department of Economics, Peter Merian-Weg 6, 4002 Basel. E-mail: marcus.roller@unibas.ch and University of Bern, Department of Economics, Schanzeneckstrasse 1, 3001 Bern, Switzerland. E-mail: marcus.roller@vwi.unibe.ch.

†University of Tübingen, Department of Economics, Melanchthonstr. 30, 72074 Tübingen, Germany. E-mail: Daniel.Steinberg@uni-tuebingen.de.

[1]This study does not represent a comprehensive evaluation of the exploited reform in Lower Saxony in 2004. We are grateful for financial support by the WWZ-Förderverein. We are grateful for the data provided by the Research Data Center (FDZ) at the Institute for Educational Quality Improvement (IQB) associated with the Humboldt University of Berlin.

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 effects of school stratification on scholastic achievement is controversially discussed in policy and research across many countries. On the one hand, proponents of integrated learning emphasize positive spillover and peer-group effects in favor of low achievers, and therefore a decline in intergenerational path dependencies regarding educational attainment (see Dustmann, 2004).[2] On the other hand, opponents of integrated learning emphasize 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). We make use of a unique exogenous change in Lower Saxony, Germany, that allows us to identify heterogeneous effects along the entire distribution of individual PISA test scores.

In order to examine whether integrated learning impinges on educational attainment, most of the studies 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. 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 with their roommates, he compares prior SAT scores with subsequent academic

[2]Integrated learning is used in the sense of performance based school tracking rather than integrating pupils with certain disabilities.

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) emphasizes 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, 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, randomized 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 reliable, 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. Our model picks up lots of elements discussed in the literature but combines them to a very flexible model that can easily be adjusted to different school systems.

Empirically, we rely on an quasi-randomized 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 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.[3] 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 heterogeneous effects across the skill distribution. 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. Thus, 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 in the context of a developed country by using micro data and exploiting a quasi-experimental setting. It is up to our knowledge the first study explicitly estimating the heterogeneous effects along the entire distribution of students' achievements.

The paper is organized as follows. Section 2 derives a theoretical framework for the effect of school tracking on the distribution of educational achievement. Section 3 provides empirical evidence, including data, the identification strategy, and results. Section 4 concludes.

[3]The PISA test takes place at the end of the ninth grade.

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)).[4]

Formally, suppose there is a continuum of students and their initial ability θ follows a uniform distribution, $\theta \sim U\left(\underline{\theta}, \bar{\theta}\right)$. Students are tracked into $1, .., J$ 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 θ taught in tier j is deterministic and given by:

$$h\left(\theta, \bar{\theta}_j, \tilde{\theta}_j\right) = f(\theta) + e\left(\theta, \bar{\theta}_j, \tilde{\theta}_j\right) = f(\theta) + s(\theta, \bar{\theta}_j) + c(\theta, \tilde{\theta}_j) \tag{1}$$
$$c = z - k(\theta, \tilde{\theta}_j)$$

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 maximized for the targeted students at z. $k(\theta, \tilde{\theta}_j)$ depends positively on the absolute distance to the median skill $\tilde{\theta}_j$ in class j: $\partial k(\theta, \tilde{\theta}_j)/\partial(|\tilde{\theta}_j - \theta|) > 0$. 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 legitimize 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:

$$\frac{\partial s(\theta, \bar{\theta}_j)}{\partial(\bar{\theta}_j - \theta)} > \frac{1}{2} \frac{\partial k(\theta, \tilde{\theta}_j)}{\partial(|\tilde{\theta}_j - \theta|)} \tag{2}$$

[4] We abstract from possible positive spillovers of social skills through the entire paper, since PISA tests exclusively account for problem solving skills.

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, θ, which is a reasonable assumption in light of our empirical setting discussed below.

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 θ.[5] All students with ability $\theta \in [\underline{\theta}, \bar{\theta}_1]$ are taught in class 1, where θ_1 is defined such that one third of the students have lower ability, $U(\theta_1) = 1/3$. Students with ability $\theta \in (\bar{\theta}_1, \bar{\theta}_2]$ are taught in class 2, where θ_2 is defined such that two thirds of the students have lower ability, $U(\theta_1) = 2/3$. The remaining students with $\theta \in (\bar{\theta}_2, \bar{\theta}]$ are taught in class 3. Table 1 summarizes 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

Effect	Class 1	Class 2	Class 3
Spillovers	$s(\theta, \bar{\theta}_1) < s(\theta, \bar{\theta})$	$s(\theta, \bar{\theta}_2) < s(\theta, \bar{\theta})$	$s(\theta, \bar{\theta}) = s(\theta, \bar{\theta})$
Congruency	$c(\theta, \tilde{\theta}_1) > c(\theta, \tilde{\theta})$	$c(\theta, \tilde{\theta}) = c(\theta, \tilde{\theta})$	$c(\theta, \tilde{\theta}_3) > c(\theta, \tilde{\theta})$
Total	$e(\theta, \bar{\theta}_1, \tilde{\theta}_1) < e(\theta, \bar{\theta}, \tilde{\theta})$	$e(\theta, \bar{\theta}_2, \tilde{\theta}_2) < e(\theta, \bar{\theta}, \tilde{\theta})$	$e(\theta, \bar{\theta}, \tilde{\theta}_3) > e(\theta, \bar{\theta}, \tilde{\theta})$

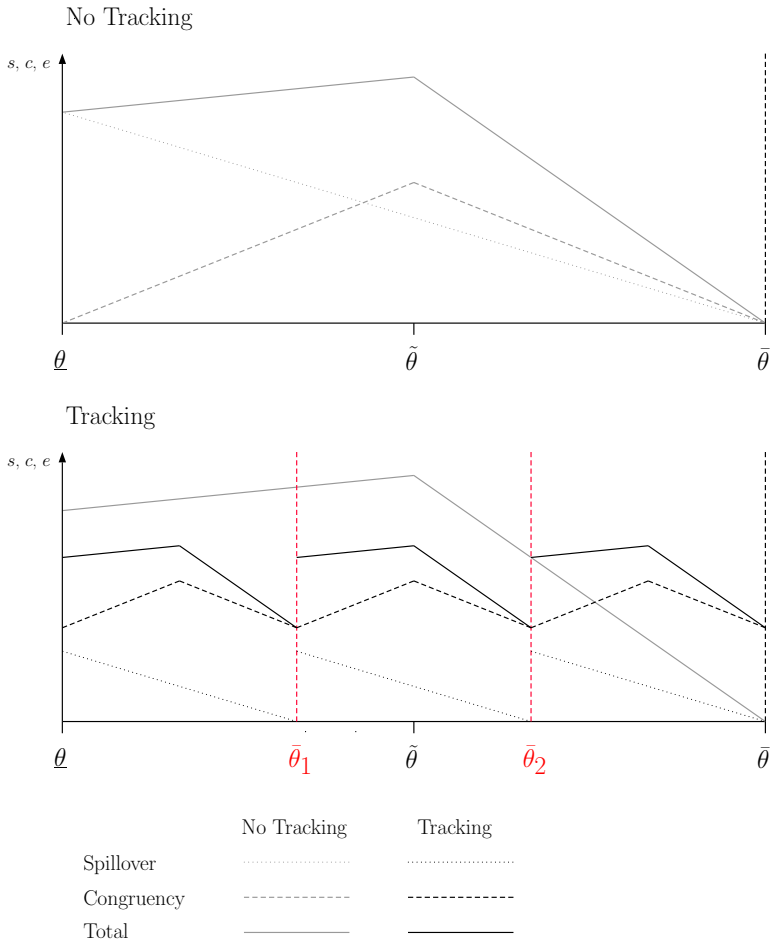
Proposition 1. *Students in the upper track, class 3, are strictly better off in the stratified school system compared to a comprehensive school:* $h(\theta, \bar{\theta}, \tilde{\theta}_3) > h(\theta, \bar{\theta}, \tilde{\theta}) \quad \forall \theta \in (\bar{\theta}_2, \bar{\theta}]$

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, $s(\theta, \bar{\theta}) = s(\theta, \bar{\theta})$. However, the distance to the class median is lower than the distance to the overall median for all ability levels in the upper class, $|\tilde{\theta}_3 - \theta| < |\tilde{\theta} - \theta| \quad \forall \theta \in (\bar{\theta}_2, \bar{\theta}]$ 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: $e(\theta, \bar{\theta}, \tilde{\theta}_3) > e(\theta, \bar{\theta}, \tilde{\theta})$ and $h(\theta, \bar{\theta}, \tilde{\theta}_3) > h(\theta, \bar{\theta}, \tilde{\theta})$.[6] ■

[5] We postulate three classes together containing the entire distribution of students, which ensures that the expenditures of both tracking systems coincide.

[6] Note that the function $f(\theta)$ is independent of tracking. Hence, it is sufficient to compare the combined effect $e = s + c$.

Figure 1: Theoretical effects of school tracking



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:* $h(\theta, \bar{\theta}_2, \tilde{\theta}_2) < h(\theta, \bar{\theta}, \tilde{\theta}) \quad \forall \theta \in (\bar{\theta}_1, \bar{\theta}_2]$

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, $c(\theta, \tilde{\theta}) = c(\theta, \tilde{\theta})$. 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, $\bar{\theta}_3 - \theta < \bar{\theta} - \theta \quad \forall \theta \in (\bar{\theta}_1, \bar{\theta}_2]$, which implies that spillover effects are lower for them, $s(\theta, \bar{\theta}_2) < s(\theta, \bar{\theta})$. In total, students in the middle track are worse off under the tracking system: $e(\theta, \bar{\theta}_2, \tilde{\theta}_2) < e(\theta, \bar{\theta}, \tilde{\theta})$ and $h(\theta, \bar{\theta}_2, \tilde{\theta}_2) < h(\theta, \bar{\theta}, \tilde{\theta})$. ■

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

Proof: The top student in class 1 is worse off because his distance to the class median and the overall median is the same, $|\tilde{\theta}_1 - \bar{\theta}_1| = |\tilde{\theta} - \bar{\theta}_1|$, and hence the congruency is equal, $c(\bar{\theta}_1, \tilde{\theta}_1) =$

$c(\bar{\theta}_1, \tilde{\theta})$. 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, $\bar{\theta}_1 - \bar{\theta}_1 = 0 < \bar{\theta} - \bar{\theta}_1$ which implies that spillover effects are lower, $s(\bar{\theta}_1, \bar{\theta}_1) < s(\bar{\theta}_1, \bar{\theta})$. We can conclude that: $e(\bar{\theta}_1, \bar{\theta}_1, \tilde{\theta}_1) < e(\bar{\theta}_1, \bar{\theta}, \tilde{\theta})$.
The loss of the median class 1 student can be expressed as follows:

$$\begin{aligned}\Delta(\tilde{\theta}_1) &= s(\tilde{\theta}_1, \bar{\theta}) - s(\tilde{\theta}_1, \bar{\theta}_1) - k(\tilde{\theta}_1, \tilde{\theta}) + k(\tilde{\theta}_1, \tilde{\theta}_1) \\ &= s(\bar{\theta}_1, \bar{\theta}) - k(\tilde{\theta}_1, \tilde{\theta}) + \int_{\tilde{\theta}_1}^{\bar{\theta}_1} \frac{\partial s(\theta, \bar{\theta})}{\partial \theta} - \frac{\partial s(\theta, \bar{\theta}_1)}{\partial \theta} d\theta\end{aligned}$$

while $s(\bar{\theta}_1, \bar{\theta}) > k(\tilde{\theta}_1, \tilde{\theta})$ 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: $e(\tilde{\theta}_1, \bar{\theta}_1, \tilde{\theta}_1) < e(\tilde{\theta}_1, \bar{\theta}, \tilde{\theta})$.
For students for which $\tilde{\theta}_1 < \theta < \bar{\theta}_1$ we can formulate the loss in the course of a transition from a comprehensive school to a stratified school system as:

$$\begin{aligned}\Delta(\theta) &= s(\theta, \bar{\theta}) - s(\theta, \bar{\theta}_1) - k(\theta, \tilde{\theta}) + k(\theta, \tilde{\theta}_1) \\ &= s(\bar{\theta}_1, \bar{\theta}) - k(\theta, \tilde{\theta}) + k(\theta, \tilde{\theta}_1) + \int_{\theta}^{\bar{\theta}_1} \frac{\partial s(\theta, \bar{\theta})}{\partial \theta} - \frac{\partial s(\theta, \bar{\theta}_1)}{\partial \theta} d\theta\end{aligned}$$

Since $k(\theta, \tilde{\theta}) < k(\tilde{\theta}_1, \tilde{\theta})$, and all other addends are positive, we can directly conclude that: $e(\theta, \bar{\theta}_1, \tilde{\theta}_1) < e(\theta, \bar{\theta}, \tilde{\theta}) \quad \forall \theta \in (\tilde{\theta}_1, \bar{\theta}_1)$.
For students characterized by $\underline{\theta} \leq \theta < \tilde{\theta}_1$ it holds:

$$\begin{aligned}(\bar{\theta} - \theta) - (\bar{\theta}_1 - \theta) &= \frac{2}{3}(\bar{\theta} - \underline{\theta}) \\ (\tilde{\theta} - \theta) - (\tilde{\theta}_1 - \theta) &= \frac{1}{3}(\bar{\theta} - \underline{\theta})\end{aligned}$$

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: $e(\theta, \bar{\theta}_1, \tilde{\theta}_1) < e(\theta, \bar{\theta}, \tilde{\theta}) \quad \forall \theta \in [\underline{\theta}, \tilde{\theta}_1)$.
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: $e(\theta, \bar{\theta}_1, \tilde{\theta}_1) < e(\theta, \bar{\theta}, \tilde{\theta}) \quad \forall \theta \in [\underline{\theta}, \bar{\theta}_1]$. ■

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 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 remains 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 equalized 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 the theoretical predictions with data based on an exogenous policy change in Lower Saxony, Germany, in 2005. In particular, we analyze the effects of proponed school tracking on the entire distribution of individual PISA test scores. 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) and respective micro data.

3 Evidence

3.1 Empirical Strategy

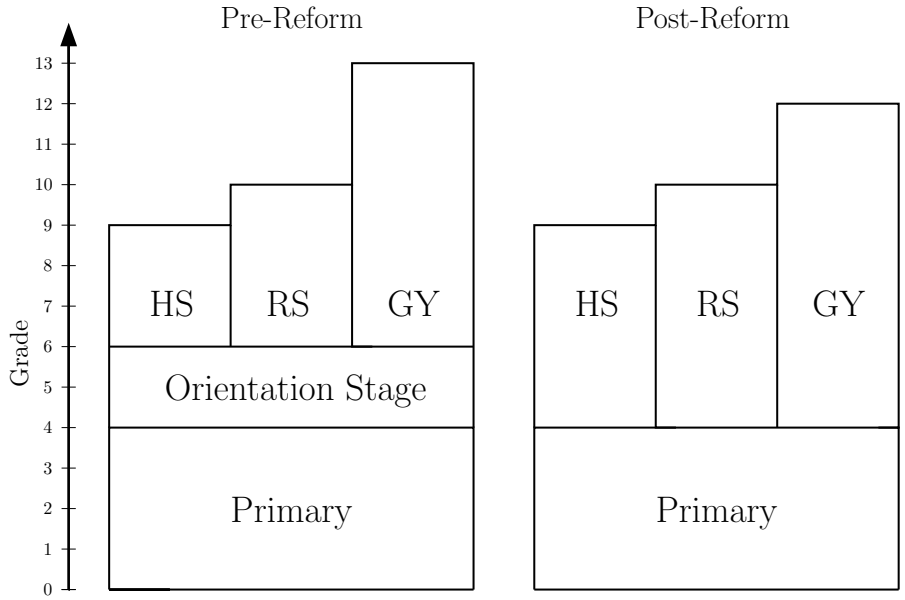
3.1.1 Setting

In Germany, each state is independently in charge of educational policies. Therefore, the educational system in Germany is characterized 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[7], 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 only into two tiers. Complementarily, all states allow comprehensive schools combining the different tracks. Additionally, the timing of tracking differs. Most states track their students after the 4th grade. In Berlin and Brandenburg however, primary school lasts 6 years.
Lower Saxony was a special case 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. Although separate schools in different buildings, the class composition remained the same in grade 5 and 6 during the "Orientierungsstufe", which results in a school system with a 6 year primary school. Figure 2 illustrates the preponed tracking in Lower Saxony graphically.[8]

[7] In some states the recommendation is binding.

[8] 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.

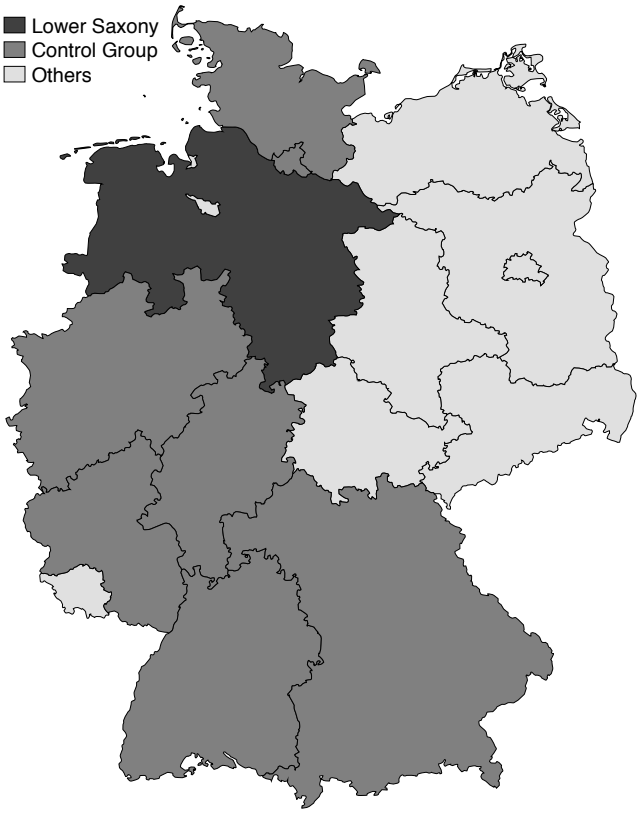
Figure 2: Preponed school tracking in Lower Saxony



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

We can exploit this reform in Lower Saxony for a difference-in-difference strategy while other states serve as a control group. In principle, the control group might be composed of all German states in our difference-in-difference setting, 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 idiosyncracies. First, we need to dispense with Bremen, which had a similar reform in the period of interest. 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 3 visualizes the treatment and control group, respectively.

Figure 3: Treatment and control group



3.1.2 Data

In order to apply this difference-in-difference strategy we need performance measures of students before and after the reform that are comparable across time and across states. As mentioned the German school system is federal, thus relying on students' grades is not a valid identification strategy. We instead use individual PISA test scores of ninth graders in order to analyze the effect of preponed school tracking on the distribution of students' achievement.[9] The international PISA-test takes place every three years starting as of 2000. 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, generating representative samples even on a state level. These tests serve as a foundation for regional comparisons within Germany, commonly known as PISA-E-tests. However, PISA-E-tests were phased out in 2006 and replaced by IQB-State-Comparison-Tests.[10]. Both tests measure the same underlying skills

[9]In Germany, PISA-2000 was designed as a national research program by the German PISA-Konsortium (Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, Manfred Weiß). The lead was with Professor Dr. Jürgen Baumert, Max-Planck-Institut für Bildungsforschung, Berlin. Results of the primary research are among others published in Baumert et al. (2001, 2002, 2003). The research tools are documented in Kunter et al. (2002). We are grateful to the German PISA-Konsortium and the Forschungsdatenzentrum (FDZ) associated with the Humboldt University of Berlin for their approval and support of this secondary analysis.

[10]The IQB National assessment study (IQB-Ländervergleich) was conducted by the Institute for Educational Quality Improvement.

and rely on similar test exercises.
In each period except for 2006 there are two samples available: students at the end of grade 9 and 15-year-old students. Table 2 gives an overview of the data sets used along with the respective sample sizes.

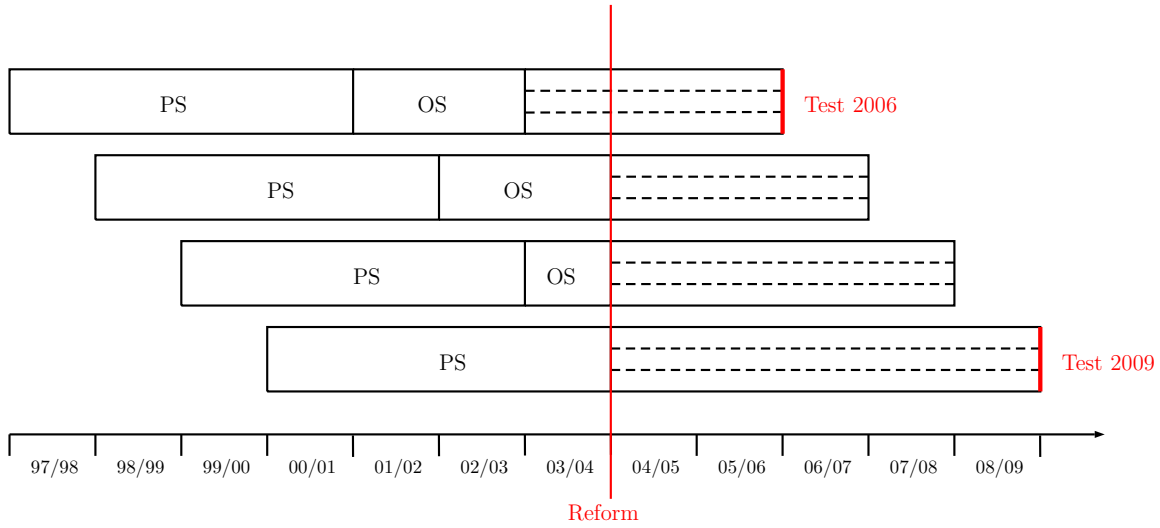
Table 2: Overview data sets

Year	Test	Sample	Observations	Source
2000	PISA-E	15-year-old	35,584	Baumert et al. (2009)
		9th grade	34,754	Baumert et al. (2009)
2003	PISA-E	15-year-old	46,185	Prenzel et al. (2007)
2006	PISA-E	15-year-old	39,573	Prenzel et al. (2010)
		9th grade	39,216	Prenzel et al. (2010)
2009	IQB-Test	9th grade	39,663	Köller et al. (2011)
				Sachse et al. (2012)

Notes: The data was provided by the research data center (Forschungsdatenzentrum) at the Institute for Educational Quality Improvement (Institut zur Qualitätsentwicklung im Bildungswesen).

Figure 4 shows how we can exploit this data for our identification. The first student cohort which was totally exposed to the intervention in Lower Saxony was tested in 2009 and started school in 2000. The previous cohort starting school in 1999 was partially treated because the OS was phased out directly after grade five while earlier cohorts are totally untreated. The last cohort which is totally untreated is the 1997 cohort which was tested in 2006. The treated 2009 test cohort and the untreated 2006 test cohort serve as the basis for our difference in difference strategy which is described in more detail below. As the focus in the IQB test in 2009 was put on reading, we have to rely on reading test scores.

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.

The reading test scores we make use of in our empirical strategy below are provided as plausible values, i.e. we observe five plausible values for each student in the sample which demands us to repeat each analysis five times. The reported results are therefore averages over the five results with adjusted standard errors (von Davier et al., 2009). Formally, the standard errors for the mean of the 5 results need to be adjusted according to the following formula:

$$SE(\hat{\mu}) = \left(\left[1 + \frac{1}{K} \right] \left[\frac{1}{K-1} \sum_{i=1}^{K} (\hat{\mu}_i - \hat{\mu})^2 \right] + \frac{1}{K} \sum_{i=1}^{K} \widehat{Var(\hat{\mu}_i)} \right)^{0.5} \tag{3}$$

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: $\hat{\mu} = \sum_{i=1}^{K} \mu_i$.
Along with the performance measures the tests collected a lot of background information on the students. We construct serveral variables from this information we make use of in the analyses. Table 3 reports descriptive statistics for all these covariates.[11] With respect to covariates, we account for the age in months along with a dummy variable, which is 1 for male students an 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 materialize 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

[11] Note that these contain the raw data without using population weights.

and control group. However, as soon as we account for population weights, these trends are much more parallel (see table 7 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.[12] 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.

Table 3: Descriptive statistics: controls. 9th grade students.

		All			Lower Saxony			Control Group		
	Year	No. Obs.	Mean	Std. Dev.	No. Obs.	Mean	Std. Dev.	No. Obs.	Mean	Std. Dev.
Male	2000	13,706	0.50	0.5	1,675	0.48	0.5	12,031	0.50	0.5
	2006	19,023	0.50	0.5	1,894	0.49	0.5	17,129	0.50	0.5
	2009	17,213	0.51	0.5	1,766	0.52	0.5	15,447	0.51	0.5
Age	2000	13,706	189	7	1,675	188	7	12,031	189	7
	2006	19,023	188	7	1,894	188	7	17,129	188	7
	2009	17,213	188	8	1,766	188	8	15,447	188	8
Lower Track (HS)	2000	13,706	0.29	0.45	1,675	0.29	0.45	12,031	0.29	0.46
	2006	19,023	0.28	0.45	1,894	0.29	0.46	17,129	0.27	0.45
	2009	17,213	0.27	0.44	1,766	0.23	0.42	15,447	0.27	0.45
Middle track (RS)	2000	13,706	0.35	0.48	1,675	0.34	0.48	12,031	0.35	0.48
	2006	19,023	0.37	0.48	1,894	0.37	0.48	17,129	0.37	0.48
	2009	17,213	0.34	0.47	1,766	0.36	0.48	15,447	0.34	0.47
Upper track (GY)	2000	13,706	0.36	0.48	1,675	0.37	0.48	12,031	0.36	0.48
	2006	19,023	0.35	0.48	1,894	0.33	0.47	17,129	0.35	0.48
	2009	17,213	0.39	0.49	1,766	0.40	0.49	15,447	0.39	0.49
Home Language	2000	12,261	0.92	0.27	1,543	0.95	0.22	10,718	0.92	0.28
	2006	17,133	0.90	0.30	1,730	0.93	0.25	15,403	0.90	0.30
	2009	14,778	0.89	0.31	1,655	0.94	0.23	13,123	0.89	0.31
Parents Abitur	2000	12,508	0.31	0.46	1,518	0.30	0.46	10,990	0.32	0.46
	2006	18,264	0.40	0.49	1,810	0.36	0.48	16,454	0.41	0.49
	2009	7,723	0.37	0.48	725	0.33	0.47	6,998	0.38	0.49

In the following section, we set out the econometric strategy more formally.

3.1.3 Difference-in-Difference

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

[12] The questionnaire contains several variables aiming at migration background which are all highly correlated.

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 preponed 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:

$$Y_{i,t} = \alpha + \gamma g_i + \tau t + \pi I_{i,t} + \beta X_{i,t} + \eta_i \tag{4}$$

$Y_{i,t}$ is the reading test score of student i in time t. In particular, we measure PISA-E test scores prior to 2009 and IQB test scores post 2006. Our identification strategy allows for transitions in test exercises as long as these tests measure the same sort of skills and the transition materializes in parallel in the treatment and control group ((Melly and Santangelo, 2015)). 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 π of early school tracking on students' performance is identified under the following set of assumptions:[13]

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,

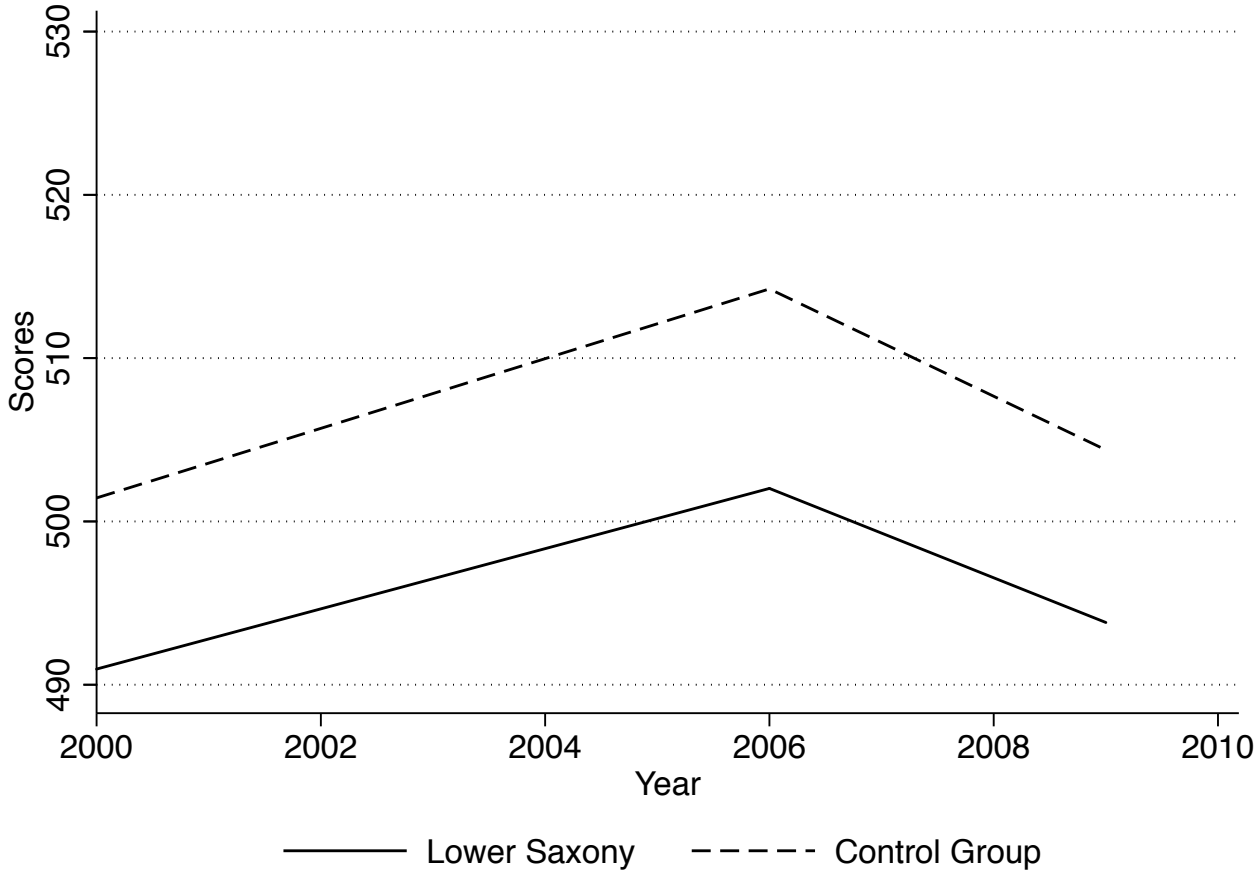
[13]The difference-in-difference approach generally identifies the average treatment effect on the treated. Estimation by OLS imposes homogeneity of the effect across groups.

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. Thus, our identification additionally rests on the assumption that the treatment effect is time constant in the control group. This cannot be tested separately. 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. Figure 5 depicts trends of average achievement scores for ninth graders which are in fact parallel from 2000 to 2006.

Figure 5: Pre-reform trends



Notes: Average test scores in reading for 9th grade students in Lower Saxony and the control group.

Placebo difference-in-differences analyses which are presented in table 6 in the appendix confirm this result. They are mainly in line with the figure 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.[14] 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 composed of 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.

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 sub-population 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.[15]

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. This was not the case (see figure 23 in the appendix). 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.

[14] All estimation results use the provided population weights.

[15] Source: Statistische Ämter des Bundes und der Länder

3.1.4 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 heterogeneous 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:

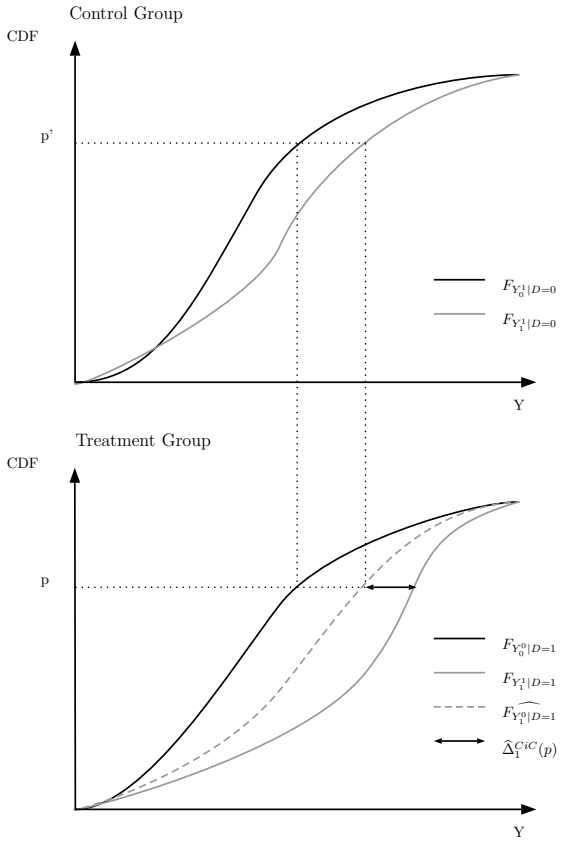
$$\Delta_{g,t}^{CiC}(p) = F_{Y_{g,t}^I}^{-1}(p) - F_{Y_{g,t}^N}^{-1}(p) \tag{5}$$

Where $F_{Y_{g,t}^I}$ denotes the potential treated outcome distribution of group g at time t and $F_{Y_{g,t}^N}$ the potential untreated outcome distribution. While we observe the post-reform distribution of the potential treated outcome in Lower Saxony, $F_{Y_{1,1}^I}$, the post-reform distribution of the potential untreated outcome, $F_{Y_{1,1}^N}$, is unknown. Following Athey and Imbens (2006), we can estimate the counterfactual distribution, and therefore the quantile treatment effect in the following way:

$$\Delta_{1,1}^{CiC}(p) = F_{Y_{1,1}^I}^{-1}(p) - F_{Y_{0,1}^I}^{-1}\left(F_{Y_{0,0}^I}\left(F_{Y_{1,0}^N}^{-1}(p)\right)\right) \tag{6}$$

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, $F_{Y_{1,0}^N}^{-1}(p)$. 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 $F_{Y_{0,1}^I}^{-1}(p')$, which is the p'-th percentile of the post-reform period distribution of the control group. Figure 6 illustrates this estimation strategy graphically.

Figure 6: Changes-in-changes identification



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

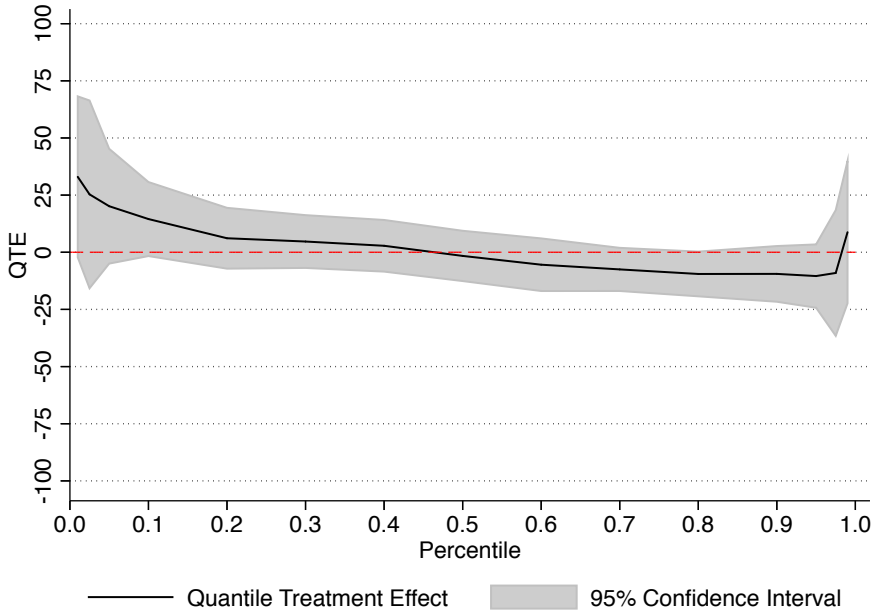
Note that unlike Athey and Imbens (2006), we posit that the change in the distribution of the potential treated outcome of the control group is the same as the change in the distribution of the potential untreated outcome in the reform group.

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. We will only make minor use of this estimator and therefore refer to the original paper for a detailed description.

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: First, the potential treated outcome can be expressed as a function of covariates, time and a random variable U: $Y_{g,t,x}^N = h(X, T, U)$. Second, this function is strictly increasing in U. Third, 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. Otherwise, the iterative procedure fails to identify

the treatment effect at the respective percentiles. Forth, 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 the test cohorts of 2000 and 2006, we can test for this assumption.

Figure 7: Changes-in-changes: pre-reform period results



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 9th grade students using only North Rhine-Westphalia as control group, 2000/2006, 7,894 observations.

Figure 7 visualizes pre-reform changes-in-changes estimates. The estimates reveal an increase in the test scores for the lower tail of the achievement distribution, albeit an insignificant on. This increase is likely due to the introduction of central exit exams for the students in the lowest track at the end of grade nine in 2006.[16]

Thus, our setting allows us to apply a standard difference-in-difference strategy as well as a changes-in-changes estimator that will give us estimates at each percentile of the performance distribution. The following section will present and discuss the results.

[16]North Rhine-Westphalia exhibits a similar reform. Unreported results of cross-checks with this state serving as a control group are in line with our argument.

3.2 Results

3.2.1 Difference-in-Difference

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 4. 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 4: Results difference-in-differences

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

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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

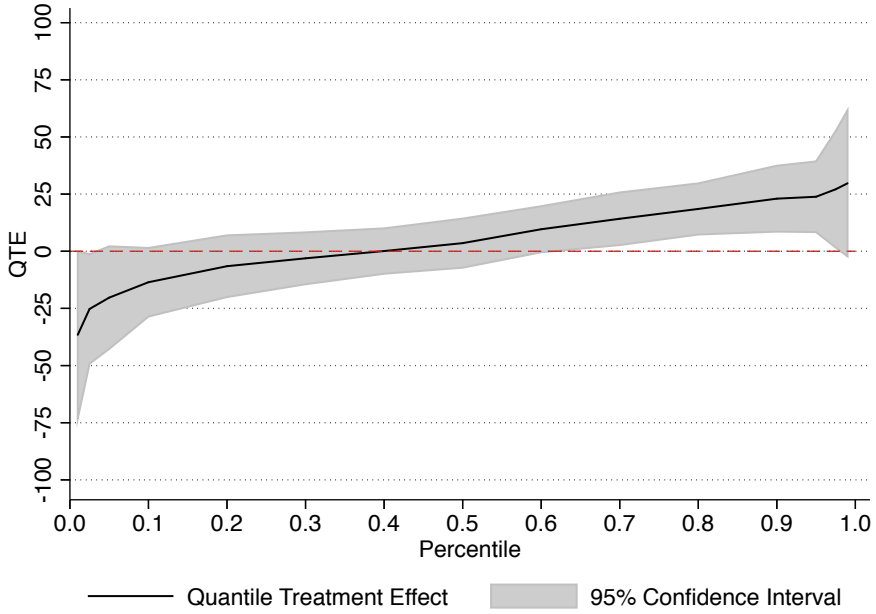
The consistency of the results through specifications with and without covariates in columns 2 and 3 suggests that the results are not driven by possible confounding factors like a different threshold of stratification. Thus, we might only want to include covariates to enhance efficiency.

3.2.2 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 8 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[17]. 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.

Figure 8: Changes-in-changes results without controls



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,

[17] All standard errors in this section are bootstrapped standard errors from 500 draws.

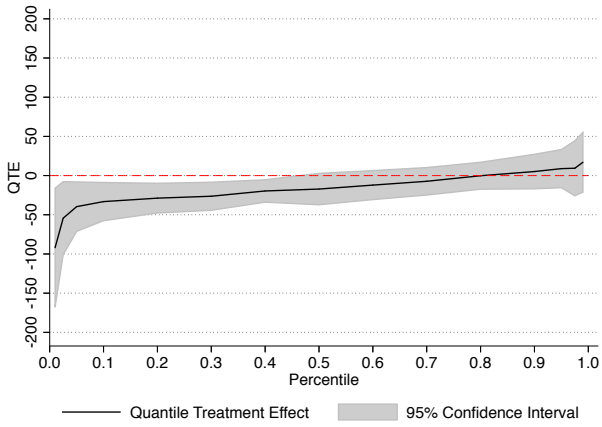
the increase of about 25 units for the top students means an increase of about a quarter standard deviation. The results suggest that the insignificant difference-in-differences estimates are exactly due to the compensation of the positive and negative effects in the upper and lower tail of the distribution. Hence, exclusively relying on a standard difference-in-difference strategy in our context would dispense with the quantile effects revealed above.

In order to 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 7 in the appendix indicates[18]. In line with the theoretical predictions, according to figure 9, 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.

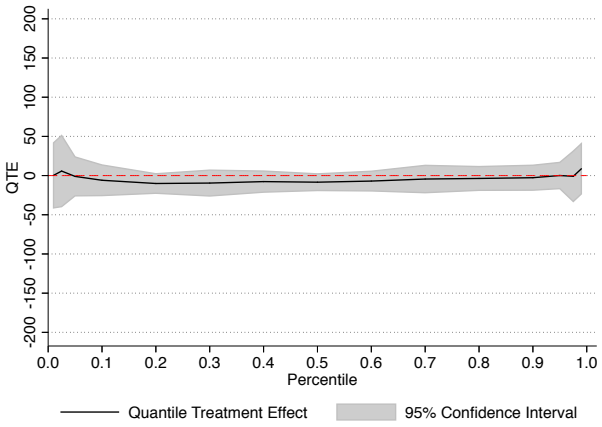
[18]Our data only allow us to separate the students according to their current track. We basically assume that the students did not change tracks since grade 5. This seems reasonable because changes from one track to another are quite rare in Germany.

Figure 9: Changes-in-changes results by school form

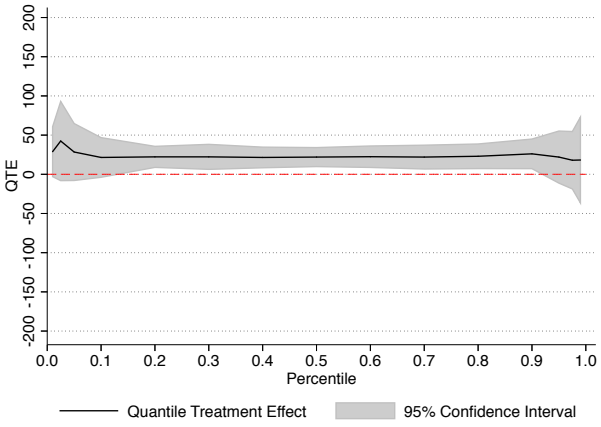
(a) Lower track (HS)



(b) Middle track (RS)



(c) Upper track (GY)



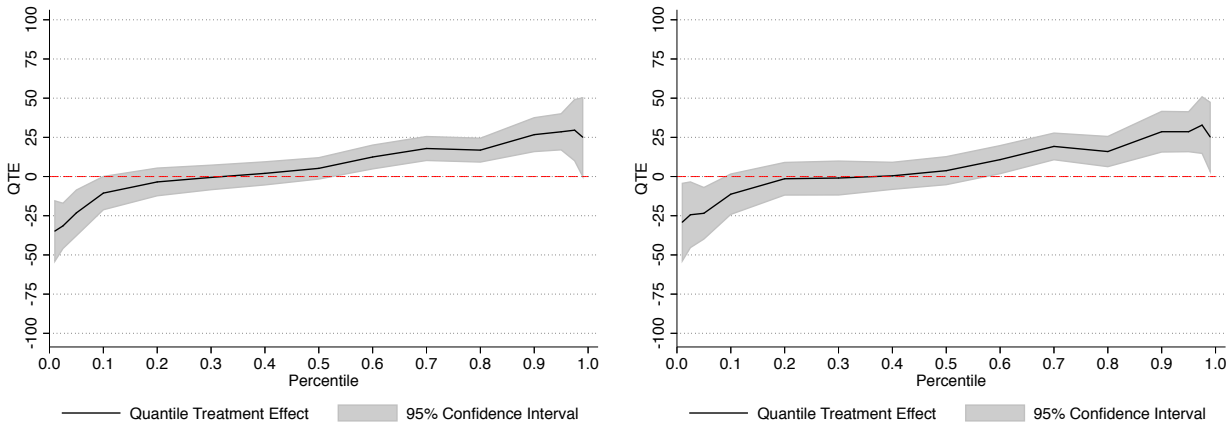
Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by school form, 2006/2009, Observations: (a) 9,872 (b) 12,941 (c) 13,423.

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 10. In line with the difference-in-differences estimates in the 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 10 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.

Figure 10: Changes-in-changes results with controls

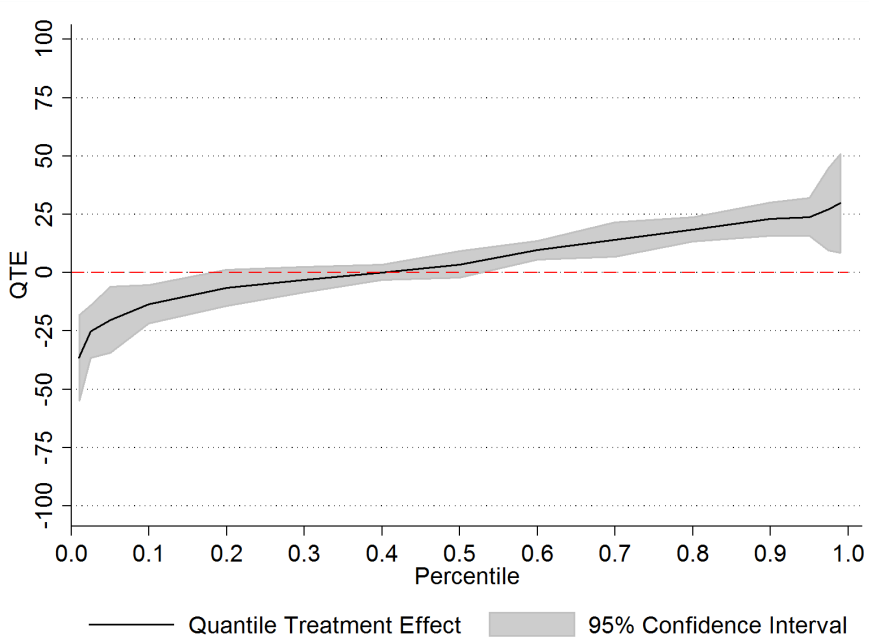
(a) PV1 with controls **(b)** PV1 without controls



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.

In section 3.1.4, we showed before that there were no effects observed prior to our reform. In order to rule out that our results are in fact driven by changes in the distributions of a state in the control group, we repeated our analysis and bootstrapped on the state-level. The resulting standard errors do not give necessarily valid standard errors for the original analysis due to the low number of states but indicate the sensibility of our results with respect to changes in the control group. Figure 11 presents the results.

Figure 11: Changes-in-changes results - state bootstrap



Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students, 2006/2009, 36,236 observations. Standard errors are generated by bootstrapping on the state-level.

In the following section, we discuss sorting and selection issues as part of our empirical setup.

3.2.3 Sorting and Selection

Apart from being the result of peer-effects, our findings might be caused by sorting effects, i.e. the shorter duration of primary schooling might simply change the sorting of students into the different tracks asymmetrically conditional on observables. In order to test for asymmetric sorting effects, we perform difference-in-difference analyses where the likelihood of attending the respective school type serves as an outcome variable. Table 5 reports the results which allow us to analyze potential heterogeneity conditional on observables. The overall relative sizes of the school types did not significantly change as indicated in the first row. A decomposition of this overall effect with respect to various covariates suggests that the relative probability of attending a school type did neither change heterogeneously for girls and boys nor conditional on parents' education. Exclusively the likelihood of attending the lower track seems to have changed for German speaking students and for non-German speakers differently. Yet, testing for joint composition changes in all schools leads to the conclusion of no asymmetric composition effects for different language backgrounds. Thus, we can are confident that the reform did not asymmetrically change the sorting of students into the different tracks conditional on covariates.

Table 5: Sorting Tests

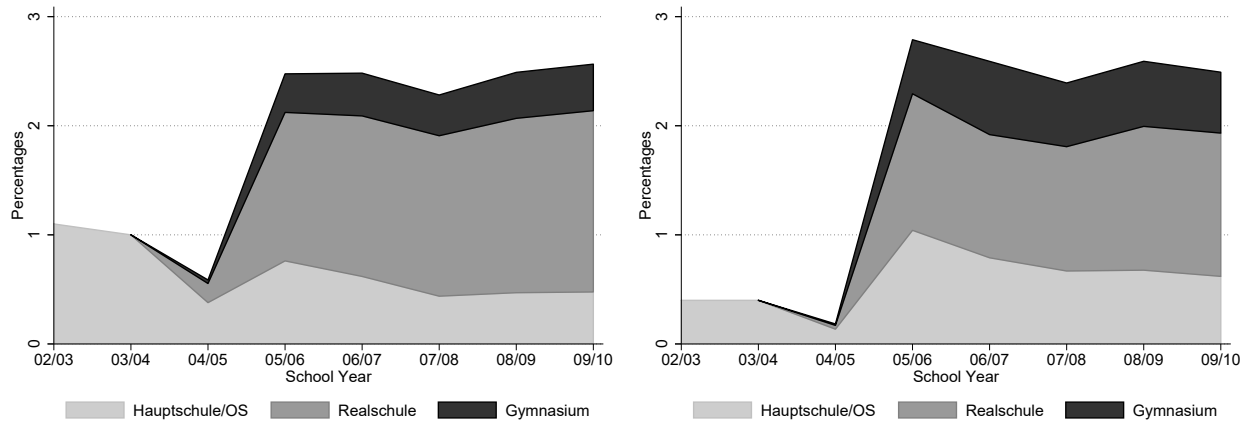
	HS	RS	GY
All	-0.0092	-0.0068	0.016
	(0.0799)	(0.0973)	(0.0953)
Male	0.0052	-0.0543	0.0491
	(0.042)	(0.046)	(0.0474)
Home Language	-0.2062*	0.0852	0.121
	(0.1091)	(0.1089)	(0.093)
Parents Abitur	0.017	-0.0594	0.0423
	(0.0747)	(0.0807)	(0.0704)

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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Besides the sorting conditional on covariates, the reform might have changed the sorting conditional on students' true ability which is unobservable. This would be the case if the precision of the tracking declined due to the shorter assessment period. We cannot test for this issue based on our data since we do not have panel data on students' performance. However, if any it would bias our results towards zero. In particular, the more imprecise the tracking is, the more randomly the students are grouped into classes. In an extreme case, this mimics the case without tracking, thus the reform could not have been effective at all.
An additional issue that could undermine some of our results is an increase in grade repetition rates. 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 in figure 12 suggests that grade repetition rates in grade 5 and 6 in fact experienced a strong increase right after the reform in school year 2005/2006, respectively.

Figure 12: Fraction of repeaters

(a) Grade 5 **(b)** Grade 6



Notes: 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' 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.

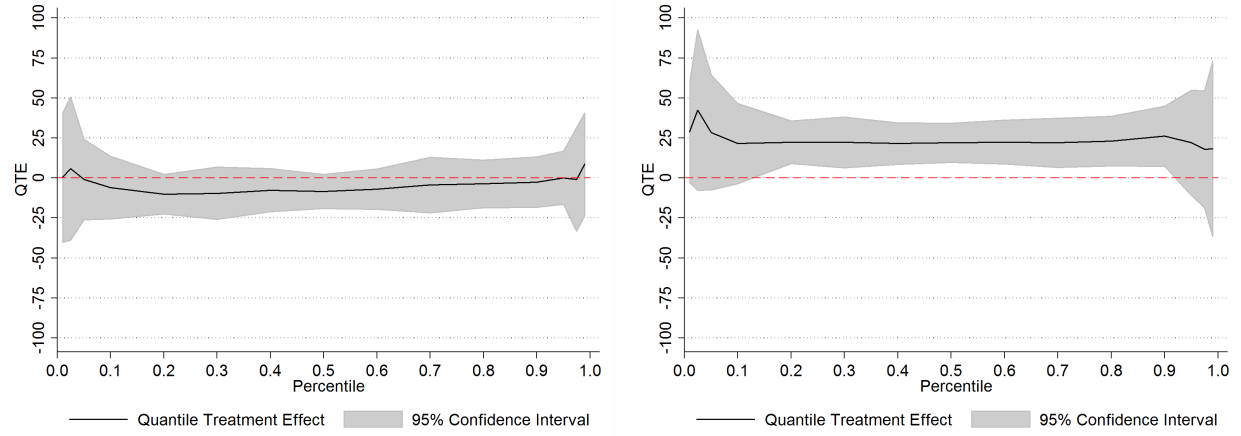
In the long run this is part of the reform and therefore of own interest as a channel of effect. In the short run however, these additional repeaters are only partially in the 2009 cohort and are therefore missing in the performance distribution. Since the rate already increases immediately for the sixth 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.

Since there is no constructed change for the lowest track our results for the lowest track are not affected by this issue. In order to check by how much the other tracks are affected we can artificially add those students back at the lower end of the distribution in the respective track.[19] While we cannot get reliable estimates for the lowest percentiles, especially the highest percentiles of the respective tracks should not be affected any more. In light of figure 13 the differences to our baseline results are negligible. Thus, we conclude that our results do not originate from these missing students.

[19]If the repeaters are not worse or better on average, our results would hold anyhow. Assuming that they are rather adversely selected is arguably reasonable.

Figure 13: Changes-in-changes results accounting for repeaters

(a) Middle Track (RS) **(b)** 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 with artificial observations to account for missing repeaters, 2006/2009, Observations: (a) 12,965 (b) 13,429.

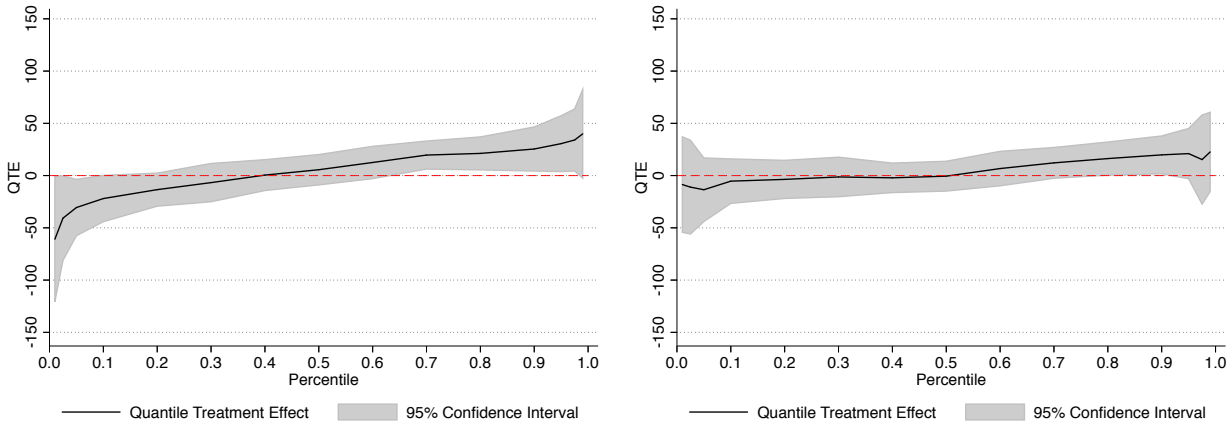
3.2.4 Heterogeneity

The results presented so far apply to the distribution of the performance of all students. This aggregate result is the consequence of potentially heterogeneous responses of different groups. Therefore, we repeated the analysis separately for males and females, natives and foreigners, and students with parents with and without high-school degree.

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 14 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 heterogeneous 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 14: Changes-in-changes results by gender

(a) Males **(b)** Females

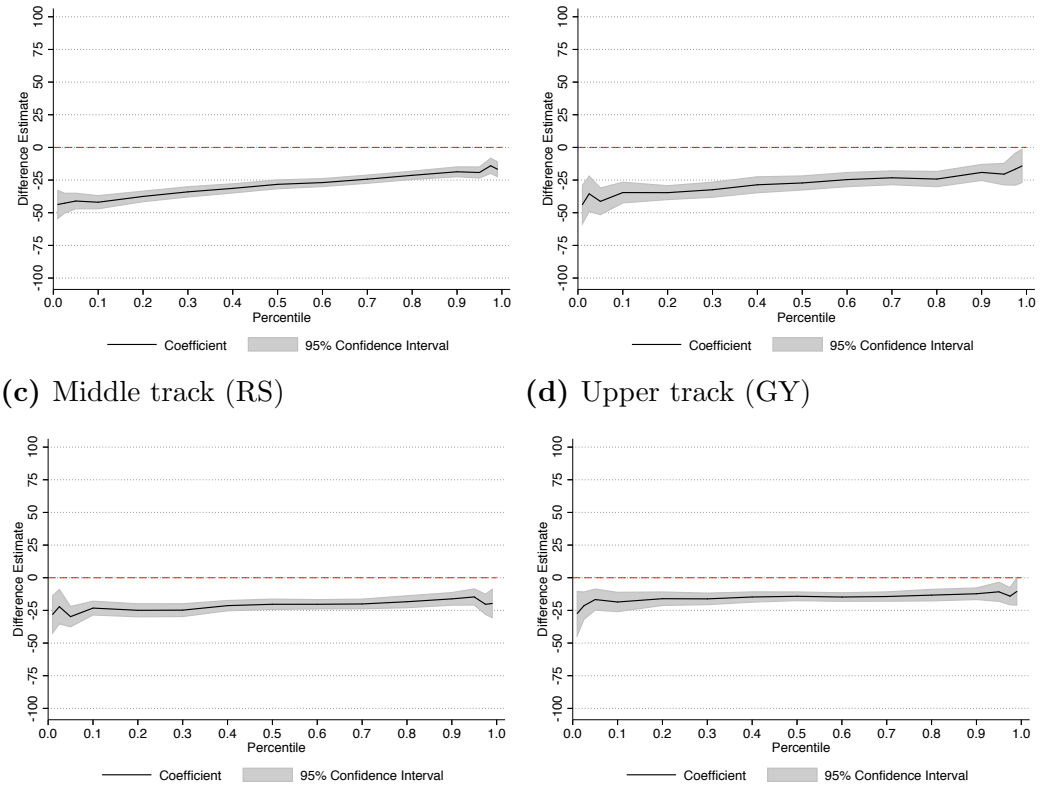


Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by gender, 2006/2009, Observations: (a) 18,333 (b) 17,903.

Figure 15: Distributional differences by gender

(a) All **(b)** Lower track (HS)

(c) Middle track (RS) **(d)** Upper track (GY)



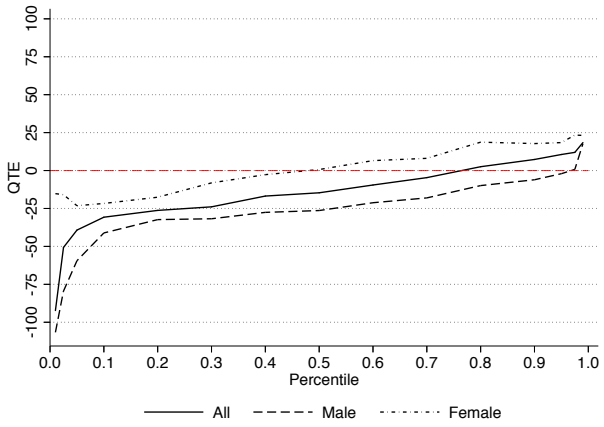
Notes: Differences in percentiles of the distributions of reading scores (plausible values 1-5) of males and females by school form. Results from quantile regressions with a constant and the male dummy. Female are the reference category. 9th grade students, 2006. Observations: (a) 19,023 (b) 5,241 (c) 7,072 (d) 6,710.

Concerning point 2, the upper left panel of figure 15 shows the difference in the distributions of test results between male and female students prior to the reform[20]. 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 under-represented in the upper track. We therefore account for gender disparities in the distributional effects separately for all three tracks shown in figure 16. 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 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 15) for girls although they are under represented. 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.

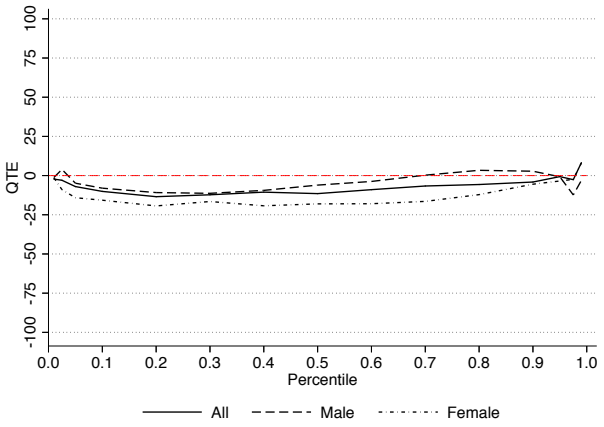
[20]Ideally, we would want to compare initial distributions, i.e. achievement distributions at the end of grade 4 for both groups. However, there is no data on these distributions for our test students.

Figure 16: Changes-in-changes results by gender and school form

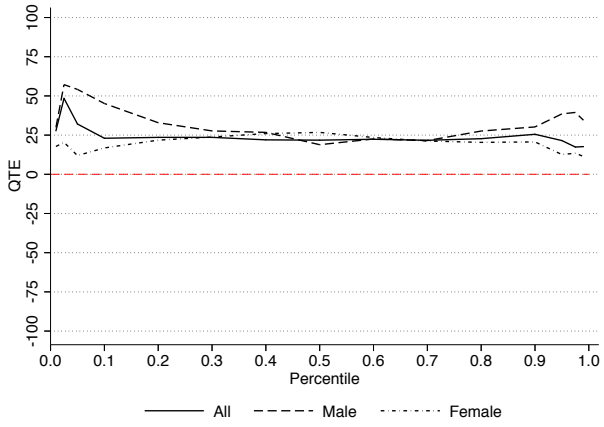
(a) Lower track (HS)



(b) Middle track (RS)



(c) Upper track (GY)



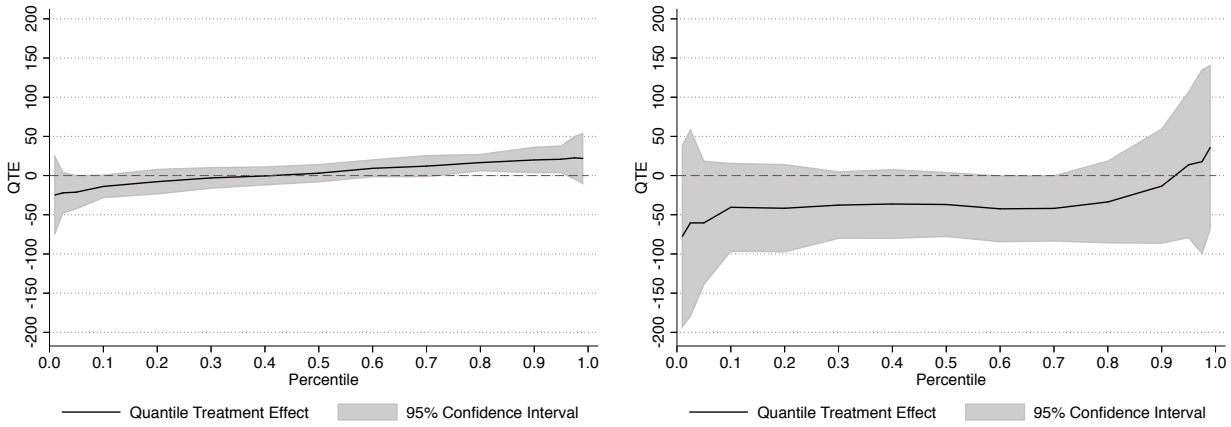
Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by gender and school form, 2006/2009. Observations: (a) Male 5,547; Female 4,325 (b) Male 6,392; Female 6,549 (c) Male 6,394; Female 7,029.

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.[21] Due to lower initial reading comprehension combined with 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 17. 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.

[21] Lower initial proficiency might lead to a catching-up effect as well.

Figure 17: Changes-in-changes results by language at home

(a) Language at home German **(b)** Language at home not German

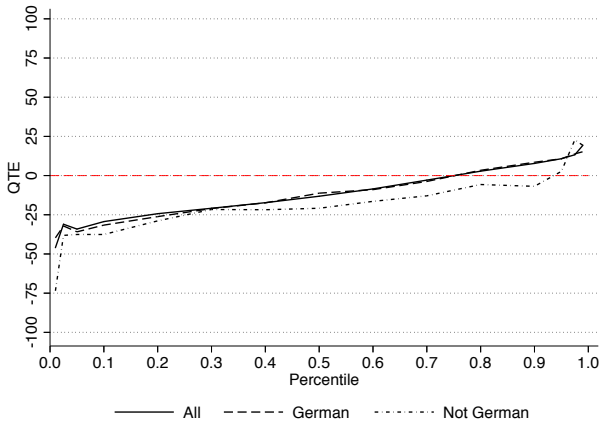


Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by language spoken at home, 2006/2009, Observations: (a) 28,662 (b) 3,249.

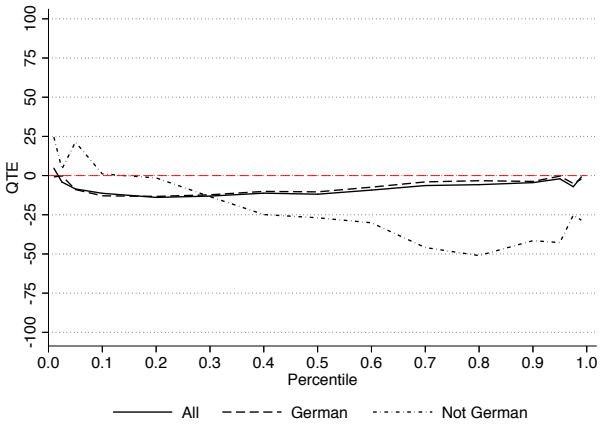
In order to verify whether the results are in fact driven by migrants predominantly sorting into the lower track, we decompose the achievement effects for the language spoken at home for all tracks in figure 18. Consistently, the results for the lower track suggest that most of the observed differences 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 19. 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 18: Changes-in-changes results by language at home and school form

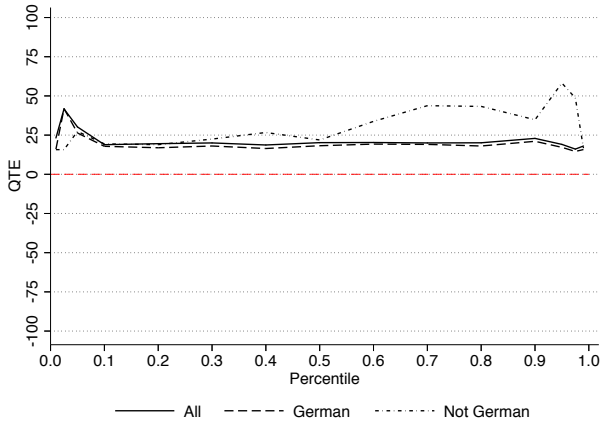
(a) Lower track (HS)



(b) Middle track (RS)

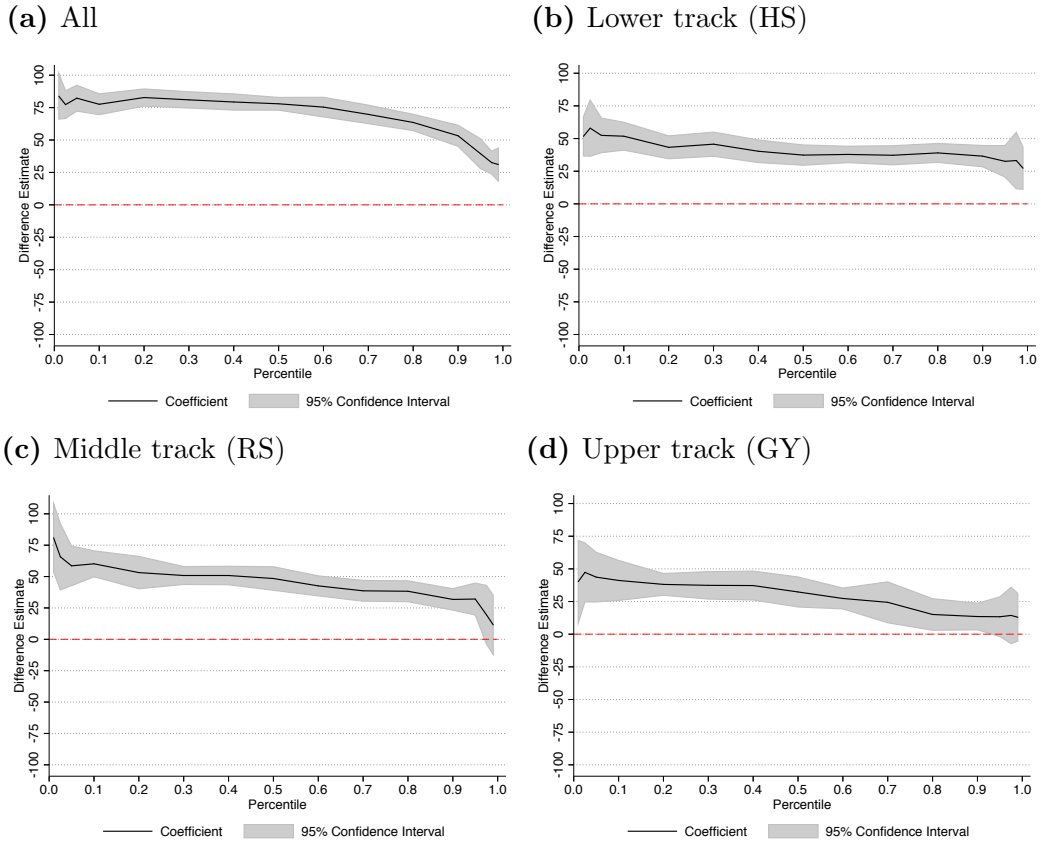


(c) Upper track (GY)



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.

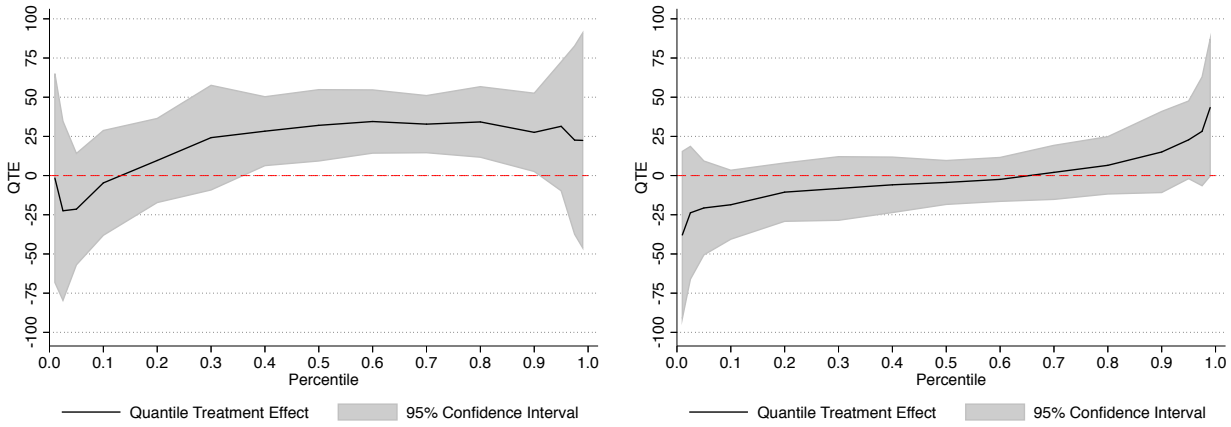
Figure 19: Distributional differences by language at home



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 20: Changes-in-changes results by parents' education

(a) Parents with Abitur **(b)** Parents without Abitur

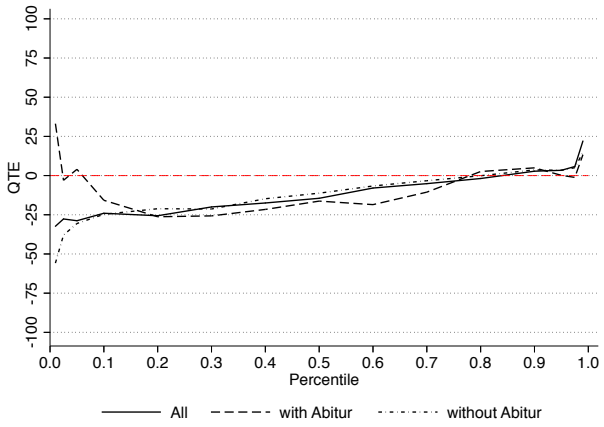


Notes: Quantile treatment effects of early school tracking on reading scores (plausible values 1-5) for 9th grade students by parents having an Abitur or not, 2006/2009, Observations: (a) 10,234 (b) 15,753.

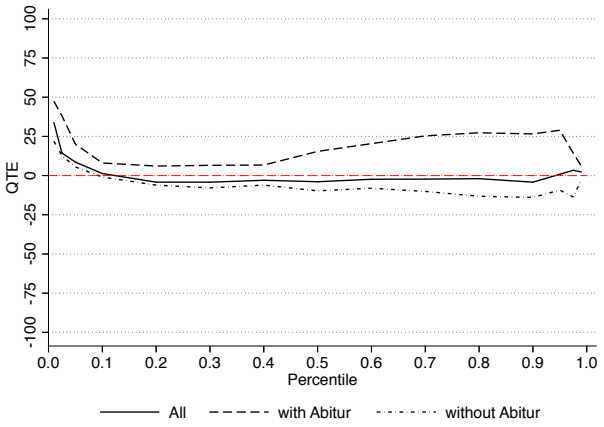
Performing the changes-in-changes procedure separately for parents with and without the Abitur, reveals the distributional effects displayed in figure 20. Apparently, the results show that students whose parents earned an Abitur almost all experience positive achievement effects as a result of the reform. 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 21.

Figure 21: Changes-in-changes results by parents' education and school form

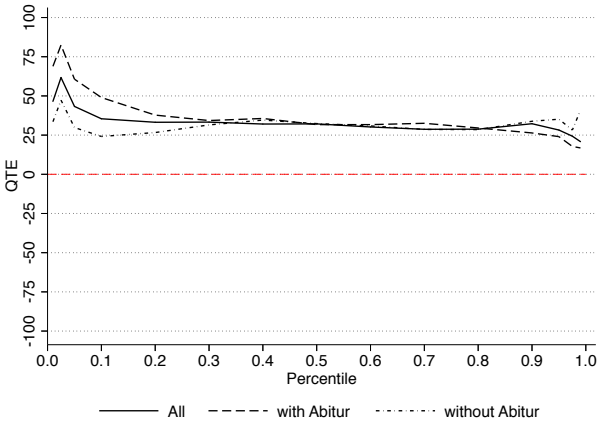
(a) Lower track (HS)



(b) Middle track (RS)



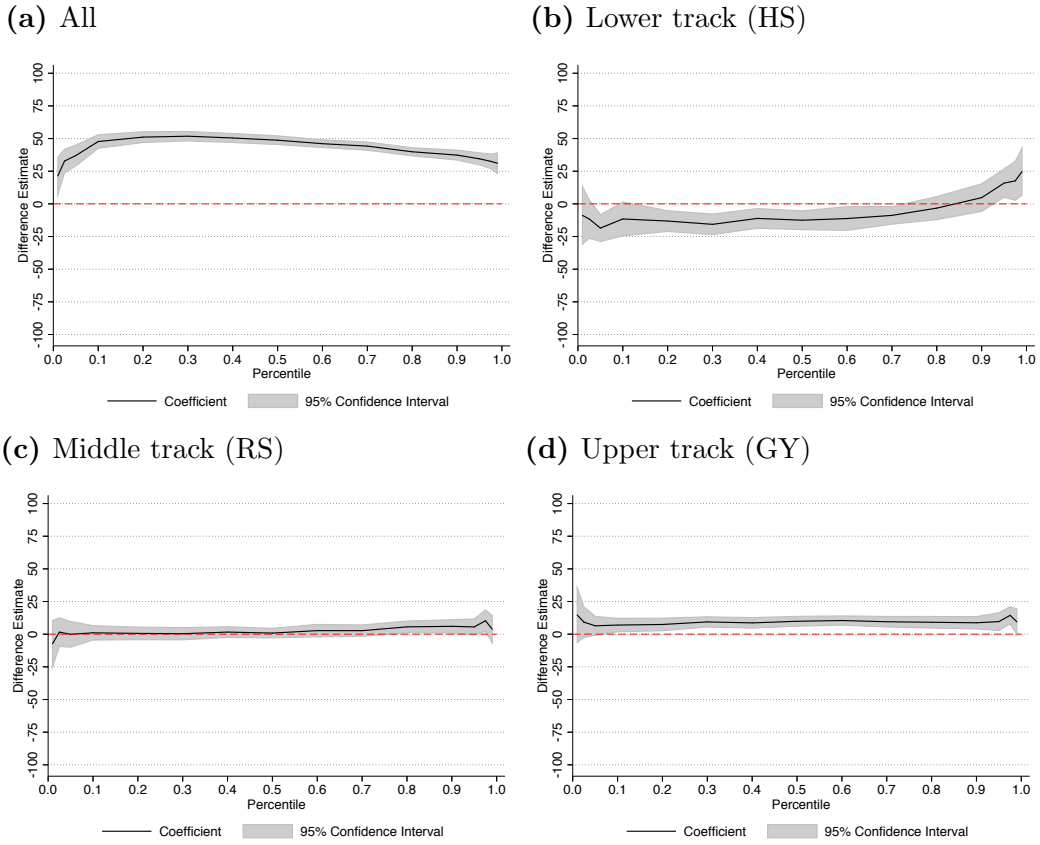
(c) Upper track (GY)



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 21, 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 22 demonstrates.

Figure 22: Distributional differences by parents' education



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.

4 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 heterogeneous 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 over representation 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

Aakvik, A. (2003). Estimating the employment effects of education for disabled workers in Norway. *Empirical Economics*, 28(3):515–533.

Ariga, K., Brunello, G., Iwahashi, R., and Rocco, L. (2005). Why is the timing of school tracking so heterogeneous? IZA Discussion Paper No. 1854, Bonn.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Schümer, G., Stanat, P., Tillmann, K.-J., and Weiß, M. (2002). *Pisa 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J., and Weiß, M. (2003). *Pisa 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*. Opladen: Leske + Budrich.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J., and Weiß, M. (2009). *Programme for International Student Assessment 2000 (PISA 2000)*. Version: 1. IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Dataset. http://doi.org/10.5159/IQB_PISA_2000_v1.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., and Weiß, M. (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Brunello, G., Giannini, M., and Ariga, K. (2004). The optimal timing of school tracking. IZA Discussion Paper No. 995, Bonn.

Card, D. and Giuliano, L. (2016). Can tracking raise the test scores of high-ability minority students? NBER Working Paper No. 22104, Cambridge, MA.

Duflo, E., Dupas, P., and Kremera, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *The American Economic Review*, 101(5):1739–1774.

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

Entorf, H. and Lauk, M. (2008). Peer effects, social multipliers and migrants at school: An international comparison. *Journal of Ethnic and Migration Studies*, 34(4):633–654.

Hall, C. (2012). The effects of reducing tracking in upper secondary school - Evidence from a large-scale pilot scheme. *Journal of Human Resources*, 47(1):237–269.

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

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation. NBER Working Paper No. 7867, Cambridge, MA.

Köller, O., Knigge, M., and Tesch, B. (2011). *IQB Ländervergleich Sprachen 2008/2009 [IQB National Assessment Study on Languages 2008/2009](IQB-LV 2008-9).* Version: 1. IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Dataset. http://doi.org/10.5159/IQB_LV_2008_v1.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K., and Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente.* Berlin: Max-Planck-Institiut für Bildungsforschung.

Lavy, V., Paserman, M. D., and Schlosser, A. (2012). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal*, 122(559):208–237.

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

Melly, B. and Santangelo, G. (2015). The changes-in-changes model with covariates. Mimeo, Department of Economics, University of Bern.

Mühlenweg, A. M. (2008). Educational effects of alternative secondary school tracking regimes in germany. *Schmollers Jahrbuch: Journal of Applied Social Science Studies/Zeitschrift für Wirtschafts-und Sozialwissenschaften*, 128(3):351–379.

OECD, P. (2004). Learning for tomorrow's world: First results from PISA 2003. Paris.

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

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., and Pekrun, R. (2010). *Programme for International Student Assessment 2006 (PISA 2006).* Version: 1. IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Dataset. http://doi.org/10.5159/IQB_PISA_2006_v1.

Prenzel, M., Baumert, J., Blum, W., Lehmann, W., Leutner, R., Neubrand, D., Pekrun, M., Rolff, R., Rost, H.-G., and Schiefele, U. (2007). *Programme for International Student Assessment 2003 (PISA 2003).* Version: 1. IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Dataset. http://doi.org/10.5159/IQB_PISA_2003_v1.

Sachse, K., Kretschmann, J., Kocaj, A., Köller, O., Knigge, M., and Tesch, B. (2012). *IQB Ländervergleich Sprachen 2008/2009. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente.* Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. DOI: 10.20386/HUB-42659.

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

von Davier, M., Gonzalez, E., and Mislevy, R. (2009). What are Plausible Values and why are they useful. *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*, 2:9–36.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics*, 85(1):9–23.

Appendix

Table 6: Pretrend difference-in-differences

	9th Grade		15-Year-Old	
	No Controls	Controls	No Controls	Controls
Time 2003 x Ref. Group			-3.659***	-5.317
			(0.902)	(4.328)
Time 2006 x Ref. Group	-1.737	-5.336	-5.797	-5.751
	(12.599)	(4.556)	(13.850)	(4.814)
Reform-Group	-10.482	-5.211	-14.384	-4.592
	(9.525)	(3.438)	(10.183)	(3.005)
Time 2003			1.377	-2.319
			(5.001)	(2.065)
Time 2006	12.800***	8.442***	11.936**	3.309
	(4.541)	(1.994)	(5.020)	(2.096)
Const.	501.439***	646.532***	504.181***	22.901
	(3.292)	(15.800)	(0.321)	(21.172)
Male		-16.276***		-16.329***
		(1.032)		(2.058)
Age		-1.161***		2.034***
		(0.082)		(0.112)
RS		88.110***		109.161***
		(2.445)		(2.180)
GY		150.403***		182.004***
		(2.395)		(2.007)
No. Obs.	32,436	32,436	47,780	47,780
R2		0.49		0.51

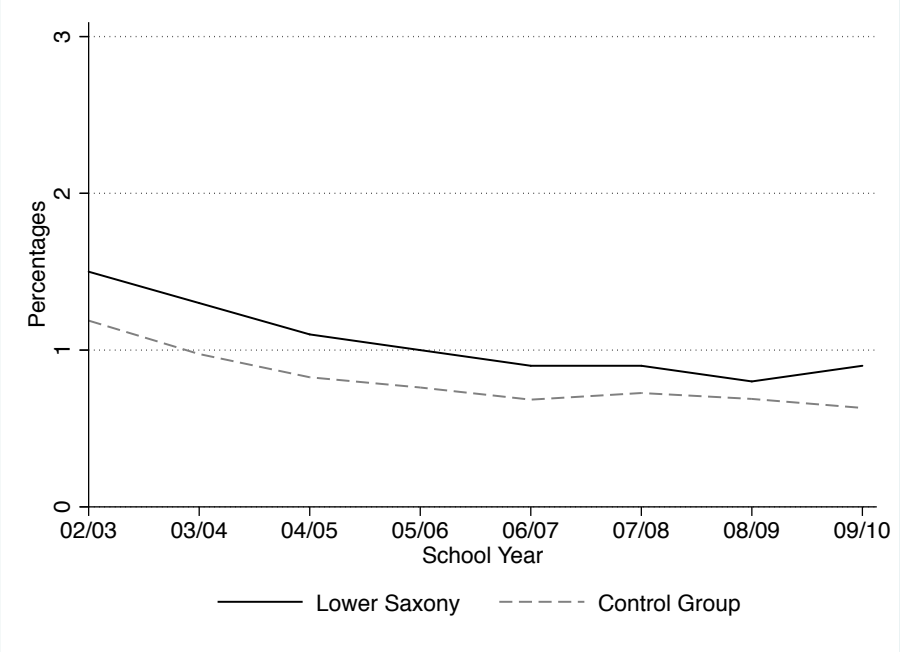
Notes: The table reports difference-in-differences estimates for pre-treatment deviations in reading test scores between the reform group and the control group. Reform group: Lower Saxony. Control group: Baden-Württemberg, Bavaria, Hamburg, Hessen, North Rhine-Westphalia, Rhineland-Palatinate and Schleswig-Holstein.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

Table 7: Fraction of students in respective track with weights. 9th grade students.

		Sample			Official Statistics[a]		
Sample	Schoolform	2006	2009	Change	2006	2009	Change
	HS	0.29	0.24	-0.05	0.29	0.24	-0.05
Lower Saxony	RS	0.39	0.38	-0.01	0.39	0.39	0.00
	GY	0.33	0.38	0.06	0.33	0.37	0.04
	HS	0.29	0.25	-0.04	0.33	0.29	-0.04
Control	RS	0.33	0.33	0	0.33	0.34	0.01
	GY	0.38	0.42	0.04	0.34	0.36	0.02

a) Source: Federal Statistical Office.

Figure 23: Share of repetitions in grade 4



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