The Effects of Early Tracking on Student Performance: Evidence from a Policy Reform in Bavaria*

Marc Piopiunik

Ifo Institute for Economic Research

Poschingerstrasse 5

81679 Munich, Germany

piopiunik@ifo.de

October 14, 2011

Preliminary version. Please do not quote without permission.

Abstract

Numerous studies indicate that tracking students early into different secondary school types increases the dependency of students' performance on their family background. However, little is known about the impact of early tracking on the performance level. This paper studies the effects of a change in tracking age on student performance. The variation in tracking age comes from a policy reform in the German state of Bavaria in 2000. Students in basic schools (Hauptschule) and middle schools (Realschule) were separated at the end of grade 6 before the reform and at the end of grade 4 after the reform; students in the most academic track (Gymnasium) were not affected. To eliminate both state-specific and school-type-specific shocks, a difference-in-differences-in-differences approach is used. This model compares students' performance before and after the reform, between students in Bavaria and students in other German states (where tracking age did not change), and between students in the school types affected by the reform and students in the unaffected school type. Student performance in math, reading, and science comes from the German extensions of PISA 2000, 2003, and 2006. The results indicate that the reform reduced student performance, suggesting that early tracking lowers the performance of both students with lower ability and students with higher ability. With early tracking, there are more students with a very low competency level and fewer students with a high competency level. Finally, the negative effects on student performance seem to persist for several years after the reform went into effect.

JEL Classification: I20, I21, I24

Keywords: Tracking, streaming, student performance, inequality, PISA.

*I thank Francesco Cinnirella, Alexander Danzer, Oliver Falck, Elke Luedemann, and Ludger Woessmann, as well as seminar participants at the Ifo Institute and at the University of Munich for their valuable comments and suggestions.

1 Introduction

The impact of tracking students into different school types is hotly debated in public and an open issue in research (see Betts, 2011; Woessmann, 2009). The debate is particularly fierce in Germany, where almost all states track students into different secondary school types at the end of grade 4 (about age 10), which is much earlier than in other developed countries.[1] Discussions as to whether early tracking is beneficial or detrimental for student performance have increased in Germany since the first results of the Programme for International Student Assessment (PISA) 2000 revealed that German students' competencies in math, reading, and science depend more on family background than they do in most other countries, indicating a strong inequality of educational opportunities in Germany (see OECD, 2001). The fact that Germany ranked only 20th among 31 industrialized countries in the performance assessment of 15-year-old students furthermore suggested that early tracking of students is not instrumental in achieving high performance. While numerous studies indicate that early tracking strengthens the relationship between student performance and family background, there is little evidence on how early tracking influences the performance level (see Betts, 2011).[2]

This paper studies the effects on performance of tracking students into different secondary school types two years earlier in their school career. The variation in tracking age comes from a 2000 policy reform in the German state of Bavaria, where students in the basic school track (*Hauptschule*) and middle school track (*Realschule*) were separated at the end of grade 6 prior to the reform and at the end of grade 4 after the reform. Students in the Gymnasium track, the most academic secondary school type,[3] were not affected by the reform; they attend the four-year primary school along with the students in the other two tracks and are then tracked into Gymnasium both before and after the reform. To estimate the impact of early tracking, student performance before the reform is compared with student performance after the reform. However, Germany-wide shocks that occur simultaneously with the reform might affect student performance, which would confound the reform effects. Therefore, performance is furthermore compared between students in Bavaria and students in other German states, where tracking age did not change. This double comparison of student performance before and after the reform between Bavaria and other German states constitutes a difference-in-differences approach. However, the reform effect might still be confounded by simultaneous state-specific and school-type-specific shocks. To overcome these confounding influences, performance is additionally compared between students in the school types affected by the reform (basic and middle schools) and students in the unaffected school type (Gymnasium). This triple comparison of student performance constitutes the difference-

[1] The median age at tracking in OECD countries is 15 years (see OECD, 2004, p. 262).

[2] Both performance level and inequality of performance have important consequences at the macro level (see Hanushek and Woessmann (2008) for an overview).

[3] The terms school *track* and school *type* are used interchangeably throughout the paper.

in-differences-in-differences approach to estimate the reform effect (see, e.g., Hamermesh and Trejo, 2000, for a triple-differences approach). Finally, student performance in math, reading, and science, provided by the German extension studies (PISA-E) of PISA 2000, 2003, and 2006, is observed for students before and after the reform.[4]

The results indicate that the reform reduced performance and increased dispersion of achievement. Under the early tracking system, there are more students with a very low competency level and fewer students with a high competency level. Furthermore, results suggest that early tracking lowers the performance level of both students with lower ability and students with higher ability. Part of the negative effect is due to increased grade repetition in the affected school tracks, with the consequence that the 15-year-old students tested in the PISA study tend to be in lower grades. Finally, the negative effects on student performance seem to persist for several years after the reform went into effect, indicating that the negative impact is not (only) due to short-term difficulties related to implementation of the reform, such as teacher shortage.

Proponents of tracking argue that tracking increases student performance because teachers teach classrooms comprised of students with more homogeneous abilities, allowing them to adjust their teaching style to the students' ability level. Teachers might also use different pedagogical methods for different groups of students, for example, whole-classroom work, smaller groups, or individual instruction. Furthermore, schools can adjust the curriculum to the students' achievement level or adjust their resources, for example, by hiring teachers with certain qualifications. In contrast, opponents of tracking worry that equality of educational opportunities suffers from early tracking since track placement might be affected by a student's socioeconomic status. They also worry that the tracking decision may be subject to misclassification, which seems particularly likely when decisions are made in early grades, since children are at different stages of their cognitive and noncognitive development. Opponents also argue that both low-performing and high-performing students benefit from interacting with each other: weak students benefit directly from the help of strong students, and strong students benefit from explaining subject matter to weak students as doing so tends to focus and consolidate their knowledge.[5]

To study the effects of age at tracking on student performance, researchers frequently exploit the fact that tracking age differs across countries. To eliminate unobserved time-invariant differences between countries, Hanushek and Woessmann (2006) use both perfor-

[4] Since an official data request to use the PISA-E student-level micro data was refused, the analyses in this paper are based on aggregated performance data published by the German PISA consortium. Note, however, that the published data are representative for each school type within each state and thus vary at the same level as the Bavarian tracking reform, namely at the school-type x state level. Therefore, the reform effects on student performance can be identified with the aggregated data. Due to lack of access to student-level data, it is impossible to investigate the effects of tracking age on the association between student performance and family background.

[5] For a comprehensive discussion of the theoretical effects of tracking on student performance, see Meier and Schütz (2008).

mance data on fourth graders, that is, prior to tracking, and performance data on eighth or ninth graders, that is, after, in some countries, students have already been tracked, in approximately the same years. They find that performance inequality increases from the end of primary education to the end of lower secondary education in countries that track students early. While the authors generally find statistically insignificant effects on performance level, the coefficient estimates suggest higher rather than lower performance levels in countries with later tracking. Since it is not clear whether the performance dispersion necessarily provides information on actual inequality of opportunity, Schütz et al. (2008) investigate a more direct measure of inequality in educational opportunities. They find that student performance depends more strongly on family background the earlier students are tracked into different school types. Woessmann et al. (2009) analyze the relationship between tracking and equality of opportunity with student-level data from the PISA 2003 study. Using the Index of Economic, Social, and Cultural Status (ESCS) as the measure for family background, their results also indicate that student performance is more strongly related to family background when students are tracked early. Ammermüller (2005) finds similar effects based on the international PIRLS and PISA data. Bauer and Riphahn (2006) compare Swiss cantons and find that tracking students early lowers intergenerational mobility. Brunello and Checchi (2007) also exploit the fact that tracking age differs across countries and find that tracking increases the effect of family background on educational outcomes and on earnings in the labor market.[6]

There are very few studies that investigate the effects of tracking on student performance in Germany. One study exploits the fact that two states, Berlin and Brandenburg, track students after six years of primary school, while all other states track students after four years of primary school. Based on a cross-section of aggregated state-level PISA-E data, the results indicate that later tracking is associated with a weaker connection between student performance and socioeconomic background, suggesting that later tracking has a positive effect on equality of opportunity (Woessmann, 2010). Another study exploits the fact that in some states several schools offer an "orientation stage" in the first two grade levels after the four-year primary school, during which students can decide which track is best for them. Mühlenweg (2008) investigates the effect of tracking age on student performance in the state of Hessen by comparing students who are tracked directly after the four-year primary school with students who are tracked at the end of the orientation stage. Controlling for a variety of background characteristics, she finds positive effects of late tracking on reading performance at the end of lower secondary education for students from low socioeconomic backgrounds, but negative effects on students at the top of the performance distribution. Overall, she finds

[6] While tracking into different school types is common practice in Europe, students are often tracked by achievement *within* schools in the United States and Canada. It is very likely that effects on within-school tracking differ from the effects of across-school tracking. Within-school tracking does not change school demographics or the mix of a student's peers, whereas both peers and school characteristics might change considerably with tracking across school types (see Betts, 2011).

no significant association between tracking age and average student performance. Baumert et al. (2009) study student performance in the state of Berlin, where high-performing students are allowed to attend a tracked Gymnasium already after grade 4, while the remaining students stay in comprehensive six-year primary schools. In particular, they investigate whether students who leave the comprehensive six-year primary schools and switch to a Gymnasium after grade four have higher learning gains in math and reading in grades five and six than students who remain in primary school. Using regression methods and propensity score matching, the authors find no negative effects on learning gains of tracking high-ability students later into different school tracks, controlling for initial performance differences.

Cross-sectional evidence on the effects of tracking on student performance is likely hampered by omitted variable bias stemming from unobserved differences between countries or between regions within a country, such as cultural differences, legal differences, or even unnoticed differences between education systems. Findings from cross-country studies are likely to suffer from greater omitted variable bias than studies using within-country variation because of greater unobserved heterogeneity across countries than across regions within a country. Other approaches try to circumvent, or at least mitigate, these omitted variable biases by exploiting changes, typically due to educational reforms, in tracking age over time within a region or country. Using changes over time eliminates any region-specific factors that do not vary over time, rendering studies based on reforms plausibly more convincing. However, there are only few studies to date that exploit changes in tracking policy over time. Meghir and Palme (2005), for example, investigate a major educational reform in Sweden in the 1950s that increased compulsory years of schooling, abolished placement based on academic achievement into an academic and nonacademic track after grade 6, and introduced a nationally unified curriculum. Exploiting the fact that the reforms were implemented in different municipalities at different times, they find that later tracking reduced later inequality in the labor market. Pekkarinen et al. (2009) find similar effects for the Finnish comprehensive school reform which replaced a two-track school system with a nine-year comprehensive school in the 1970s, postponing the tracking into a vocational and an academic track from age 11 to age 16. Galindo-Rueda and Vignoles (2007) study the tracking reform in the United Kingdom where the system of early tracking was gradually replaced by one with comprehensive schools, with both systems coexisting during the 1960s and 1970s. Using a value-added model, which controls for prior performance, and the political affiliation of the electorate as an instrument for early implementation of the new school system (with conservatives switching to comprehensive schools later), they find some evidence of positive effects on student performance in the selective system. However, Pischke and Manning (2006) reanalyze the U.K. reform and demonstrate that it is unlikely to eliminate selection bias due to students choosing what type of school to attend.

The few cross-sectional studies for Germany provide only first descriptive evidence on how tracking affects performance because the time at which states track students is likely endogenous. For example, both student performance and tracking age might be affected by other state-level factors, such as educational preferences or the socio-economic composition of the population. In contrast, the approach of this study overcomes many of the issues arising in cross-sectional studies from unobserved heterogeneity across states because both state-specific and school-type-specific shocks as well as any time-invariant state-specific influences are eliminated. To the best of my knowledge, this is the first study to provide evidence on the effects of tracking age on the level and distribution of student performance in a state with early and rigid tracking, exploiting changes in the tracking age over time.

The remainder of the paper is structured as follows. Section 2 briefly describes the German school system and Section 3 the tracking reform in Bavaria. Section 4 presents the estimation strategy for identifying the impact of the reform on student performance. Section 5 describes the German PISA data and provides descriptive statistics. Section 6 presents results on the effects of early tracking on the level and distribution of student performance, as well as a large set of robustness checks, including evidence on common pre-reform trends. Section 7 concludes.

2 The German School System

In Germany, children start school in the year after they turn six and attend four grades in primary school (*Grundschule*).[7] At about age 10, students are separated into different secondary school types, which differ both by duration and by curriculum. Basic schools (*Hauptschule*) provide basic general education and typically lead to a certificate after grade 9 (in few states after grade 10). Middle schools (*Realschule*) provide a more extensive general education and last six years.[8] High schools (*Gymnasium*), the only secondary school type that exists in all German states, offer the most academic education and typically cover nine grades.[9,10] In the majority of states, comprehensive schools (*Gesamtschule*) exist in addition to the other school types. This school type encompasses all lower and upper secondary education levels and is typically attended only by a small fraction of students. While West German states traditionally have three different school types and East German states two

[7] Because authority and control over education policy in Germany lies with each state (*Bundesland*), the school structure differs somewhat across states. In two states (Berlin and Brandenburg), for example, primary school lasts six years.

[8] Instead of basic and middle schools, East German states and the state of Saarland have integrated schools (often called *Mittelschule* or *Regelschule*), which offer the school-leaving certificates typically obtained in basic and middle schools.

[9] There is no perfect translation for the German school type *Gymnasium*. Sometimes, the British term "grammar school" is used. Note that a U.S. high school is different from a German high school (*Gymnasium*).

[10] In almost all East German states, as well as in Hamburg and Saarland, high schools contain only eight grade levels. All other German states are currently also shortening high school duration from nine to eight years, with Schleswig-Holstein being the last state to complete this reform in 2016.

different school types, all German states offer three school-leaving certificates: those acquired at the end of basic school, middle school, and high school.[11]

Students with a basic school degree typically enter an apprenticeship that combines part-time vocational school and firm-based training. Students with a middle school leaving certificate might do the same apprenticeships, but are also entitled to attend a vocational school that provides a higher education entrance qualification. Specifically, the student can acquire a technical school degree (*Fachhochschulreife*), which qualifies him to attend a polytechnic (*Fachhochschule*). The high school leaving certificate (*Abitur*) is a prerequisite for attending a university or other institution of higher education. Thus, high school is the only type of secondary school that provides direct entry into tertiary education.

The secondary school track decision in Germany is based on teacher recommendations and/or on parents' wishes. At the end of primary school, neither ability tests nor centralized examinations exist that could provide information as to the students' academic potential. Instead, primary school teachers recommend a secondary school type for each student. This recommendation is mostly based on the student's grades in the two major subjects German and math (and sometimes also science). Grade point averages (GPA) for each of these subjects are not only based on the results of written exams taken during the school year, which are devised and graded by the individual teacher, but also depend on other factors, such as students' class participation. The GPA-based school recommendation is binding in some (e.g., Bavaria), but not all states. School authorities usually define a cutoff for the average grade in German and math (and science) that students must achieve to receive a recommendation for a certain secondary school type.[12] In states with a non-binding recommendation, parents are free to choose the child's secondary school track. However, even if the recommendation is not binding, parents, especially those with low education background, might be influenced by the recommendation when making this decision (Dustmann, 2004), leading to a very strong correspondence between teacher recommendation and the secondary school track actually attended. For all states, both those with binding and non-binding recommendations, the secondary school type chosen by the parents coincides with the teacher recommendation 83 percent of the time (cf. Pietsch and Stubbe, 2007).

3 The Tracking Reform in Bavaria

Before the reform, high-performing students attended the Gymnasium after the four-year primary school, while all other students attended the more vocational-oriented basic school (*Hauptschule*). After two years in basic school, the better-performing students attended a

[11] See Lohmar and Eckhardt (2010) for a detailed description of the German school system.

[12] School grades in Germany range from 1 (very good) to 6 (fail). Students in Baden-Württemberg and Saxony, for example, need an average grade of 2.5 in German and math to receive a recommendation for high school. In Bavaria, for example, students need an average grade of 2.0 (see Kropf et al., 2010, for more details).

middle school (*Realschule*), which lasted four years. Because student performance differed strongly at the beginning of middle school (partly because students came from different basic schools), the Bavarian parliament, in April 2000, decided to institute six-year middle schools (*sechsstufige Realschule*), to be attended immediately after primary school and containing grades 5 to 10. The expectation was that the strongly differing performance levels of students upon entering middle school would be evened out in grade 5, with the ultimate aim of raising performance levels in grades 7 to 10. Under the reform, basic school and middle school students are separated at the end of fourth grade. Before the reform, basic and (future) middle school students studied together until the end of sixth grade since classrooms were not formed on the basis of students' abilities in grades five and six. Importantly, the reform did not affect Gymnasium students, who always started Gymnasium immediately after the four-year primary school. Furthermore, the reform did not change the grade point averages required to attend a Gymnasium or middle school.[13]

State-wide implementation of the six-year middle schools did not occur within one school year. A few private six-year middle schools already existed in the 1970s. Several middle schools added a fifth and sixth grade until 1992 when the Bavarian Ministry of Education began a pilot project to test the functioning of six-year middle school.[14] Under the new schooling law of April 2000, middle schools were given several years to implement the reform and add a fifth and sixth grade. State-wide implementation of the six-year middle schools was complete by school year 2003/2004.[15] Because basic school students in grades 5 and 6 were guaranteed to start middle school in seventh grade when the reform was just introduced, 2004/2005 was the last school year in which students started attending a four-year middle school (cf. Bavarian State Ministry of Education, 2008, p. 80). Since the implementation of six-year middle schools extended over a longer period, four-year and six-year middle schools

[13] Details about the reform as reported here were provided by employees of the Bavarian Ministry of Education working in different departments responsible for basic schools, middle schools, and pilot projects.

[14] In principle, all middle schools could apply for the pilot project. However, the entity paying the material costs of the middle school (*Sachaufwandsträger*), the municipality for state-run schools and, usually, the church for private schools, had to approve the application. The costs involved could be substantial. For example, in some cases, the need for additional classrooms meant building projects were necessary. Especially, participating middle schools could only participate in the project if they were able to hire extra teachers. In the end, the Bavarian school supervisory board decided which middle schools would participate in the pilot project.

[15] During the school year 1995/1996, when most of the students who were tested in PISA 2000 attended grade 5, only 41 out of 326 middle schools contained six grade levels. Since students could attend either 4 or 6 grade levels in these middle schools, numbers on students attending a fifth or sixth grade in middle schools are even smaller: while 2,500 students attended grade level 5 (only 1,000 students grade 6), about 35,000 students attended each of the higher grade levels 7 to 10. The respective figures for the school year 1998/1999, when PISA 2003 students attended the fifth grade level are as follows: 65 out of 326 middle schools offered six grade levels; while about 35,000 students attended a seventh grade, there were about 5,000 students in each of the grade levels 5 and 6. The numbers are considerably different for the school year 2001/2002, when most of the PISA 2006 students were in fifth grade: now 224 out of 334 middle schools offered six grade levels. While about 32,000 students attended grade level 7 in middle schools, there were already about 25,000 students in the fifth grade level of a middle school (cf. Bavarian Statistical Office (1996) and respective volumes).

co-existed for several years. This means that some basic and middle school students in the pre-reform period PISA-E 2000 and 2003 were already tracked after grade 4, while some students in the post-reform period PISA-E 2006 continued to be tracked after grade 6 (see Table 1).

Potential Reform Channels

The Bavarian tracking reform might have affected student performance through several channels, probably the most important of which is peer effects (for an extensive literature review, see Sacerdote (2011)). Prior to the reform, the lower-performing basic school students interacted with the better-performing (future) middle school students until the end of sixth grade; the reform changed the peer groups for both basic and middle school students in grades 5 and 6.[16] Sacerdote (2011, p. 250) notes that "many researchers and teachers have argued that peer composition is as important a determinant of student outcomes as other widely cited inputs including teacher quality, class size, and parental involvement." This means that the change in peer groups might have substantial effects on student performance in both basic and middle school, with peer effects broadly defined as encompassing any impact that classmates' background, behavior, or outcomes have on own outcome. For example, a student may learn directly from her classmates; teachers may teach at a higher level or faster pace if classmates have higher abilities; or classmates' high achievement might motivate a student to work harder. Indeed, researchers have estimated modestly sized and statistically significant peer effects on student performance, with evidence from nonlinear models of large peer effects where high-ability students benefit from the presence of other high-ability students and are hurt by low-performing peers. Achievement for students in the middle of the performance distribution tends to be less affected by peer composition (see Sacerdote, 2011).

Another channel through which the reform might affect the performance of middle school students is a change in the curriculum. Because middle school students were subject to different curricula in grades 5 and 6 after the reform, the curricula of grade levels 7 to 10 also had to be completely revised. Even though middle school curricula is more demanding content than basic school curricula, it is not entirely certain that the new middle school curricula increased student performance. Because the curricula in basic school did not change, there is no curriculum effect on the performance of basic school students. A further channel, again only affecting middle school students, has to do with teaching personnel. After the reform, middle school students in grades 5 and 6 are taught by middle school teachers instead of by basic school teachers. Teaching personnel also changed somewhat in the higher grade levels because the reform required that middle schools hire

[16] If the reform affected the composition of basic and middle school students—which cannot be investigated with the aggregated data—then peer groups also changed at higher grade levels.

numerous new teachers. Whether middle school teachers in general, and the new teachers in particular, are better able to induce student learning than basic school teachers is unclear. Another potential way the reform might have affected student performance is through school resources. In both the basic and middle school track, school inputs such as class size might have changed as fewer students attended basic schools and more students attended middle schools. Indeed, official statistics reveal that average class size in the middle school track increased slightly, while average class size in the basic school track decreased slightly after the reform (see Kultusministerkonferenz, 2007). Whether class size increased or decreased in the individual middle school depended on how quickly the school was able to hire new teachers.[17] Because research is ambiguous with respect to the effect of class size on student performance (Hanushek, 2002), it is not likely that the rather small changes in class size influenced student performance.[18]

In sum, the direction of the reform effect on student performance in the basic and middle school tracks is not clear a priori. Concerning the possibly most important channel, peer effects, the literature suggests that performance of basic school students should fall after the reform because students with higher ability now attend middle schools instead of basic schools in grades 5 and 6. The direction of the peer effect on the performance of middle school students is less clear. On the one hand, performance level might increase in middle schools because students with lower ability no longer interact in grades 5 and 6. On the other hand, performance level might fall because existing evidence suggests that test scores are higher in classrooms containing a more heterogeneous mix of students (Vigdor and Nechyba, 2007).[19]

4 Identification Strategy

To identify the causal effect of early tracking on student performance, this paper exploits variation in tracking age over time in the German state of Bavaria. The Bavarian reform affected basic and middle school students, but it did not affect Gymnasium students in Bavaria or students in any school type in other states. Instead of estimating the reform

[17] The reform created high demand for middle school teachers, which led to a shortage. The shortage was especially severe for the science subjects, French, and sports, whereas the subjects of German and math were less affected. Middle schools with a teacher shortage compensated for it by requiring additional hours of work from the extant teachers, and very few lessons had to be canceled. In sum, the teacher shortage likely had no discernible effect on German and math performance of middle school students, and at most a moderate effect in the science subjects.

[18] Indeed, including average class size at the school-type level in robustness checks leaves the reform estimates unchanged (see Table 8).

[19] Two issues of nonrandom selection might bias the reform effects in models with student-level data. First, the time at which a middle school implemented the fifth and sixth grades might be correlated with student performance. Second, since four-year and six-year middle schools co-existed for several years, placement of students into one or the other might be correlated with student ability. However, these two possible selection biases are not an issue in this study, since the aggregated performance outcomes at the school-type level and at the state level always combine the performance of students in four-year and six-year middle schools.

effects in a difference-in-differences model, comparing student performance in Bavaria and control states before and after the reform, the Bavarian reform allows us to estimate the effects of early tracking on student performance in a difference-in-differences-in-differences model. Besides the before-after and Bavaria-control states comparisons, which only eliminate Germany-wide shocks, the triple-differences approach additionally compares performance between students in the school types affected by the reform (non-Gymnasium tracks) and students in the unaffected school type (Gymnasium), thereby also eliminating any state-specific and school-type-specific shocks.

The difference-in-differences strategy is based on the strong assumption that there are no state-specific shocks on student performance; however, the identification assumption that underlies the triple-differences model is less restrictive: the reform effect is identified if the performance difference between non-Gymnasium school tracks and Gymnasium would have developed identically over time in Bavaria and control states in absence of the reform. This assumption would be violated, for example, if the control states that performed poorly in PISA-E 2003 (pre-reform) put special pressure on the low-performing non-Gymnasium school types to boost performance.[20] Identification of the reform effects would also be hampered if other significant educational reforms were implemented between the pre-reform and post-reform period in Bavaria or in the control states. Especially changes in the basic and middle school track in Bavaria might confound the effects of the tracking reform. Indeed, a new type of classes (*M-Klassen*) was introduced in the basic track in Bavaria in the school year 1999/2000. These classes provide basic track students the opportunity to acquire a middle school degree after grade 10. To prepare students for the higher degree, the curriculum of these classes is more demanding.[21] Therefore, this new type of classes might have increased the performance of basic track students.[22] Another important reform in Bavaria reduced the length of Gymnasium from nine to eight grade levels. But since the first cohort graduates from the eight-year Gymnasium (G8) only in the school year 2010/2011, this reform did not affect the Bavarian Gymnasium students that were tested in PISA-E 2003 or 2006. A Germany-wide education reform was implemented by the Secretariat of the Standing Conference of the Ministers of Education (*Kultusministerkonferenz*) which decided to introduce new educational standards that define general educational goals and specify competencies that students of a certain grade level should have acquired. Because these new standards were implemented in all German states at the same time, this reform

[20] A robustness check shows that results are similar when estimated with a smaller group of control states that performed similarly to Bavaria in PISA-E 2003. As these states belonged to the best-performing states in Germany, non-Gymnasium schools are unlikely to have been subject to much pressure from education ministries to increase student performance.

[21] A voluntary tenth grade in basic schools has existed since the school year 1994/1995. However, special classes that prepared students for the middle school degree did not exist.

[22] Note, however, that this new class type already existed during the pre-reform period PISA-E 2003 and that the share of ninth-grade basic school students attending this type increased only slightly between the school year 2002/2003 (16.3 percent) and 2005/2006 (18.1 percent).

does not confound the Bavarian reform effect since Germany-wide shocks are eliminated in the triple-differences model.[23] Concerning educational reforms in the control states, there were some between 2003 and 2006 and these are addressed in robustness tests below.

All models in this study pool the three PISA subjects of math, reading, and science such that all results are based on the same subjects. Numerous robustness checks show that excluding the subjects that are not directly comparable across the PISA waves, math and/or science, yields quite similar results. Because performance outcomes are not published at the school-type level (except Gymnasium) in PISA-E 2000, the triple-differences model includes only the PISA-E waves 2003 and 2006. Specifically, the following model is estimated:

$$\begin{aligned} y_{kist} = {} & \alpha + \beta reform + \gamma_1 Bavaria + \gamma_2 PISA2006 + \Gamma_3 schoolType \\ & + \gamma_4 Bavaria * PISA2006 + \gamma_5 nonGym * Bavaria + \gamma_6 nonGym * PISA2006 \\ & + \Gamma_7 X_{ist} + \epsilon_{kist} \end{aligned} \tag{1}$$

where y_{kist} is the outcome y in subject k, school type i, state s, and PISA-E wave t. We are interested in β, the coefficient on the reform indicator, which equals 1 for basic and middle school track in Bavaria for PISA-E 2006, and equals 0 otherwise. *nonGym* is a binary indicator that equals 1 for all non-Gymnasium school tracks and equals 0 for Gymnasium. Due to the limited statistical degrees of freedom as a result of the limited number of aggregated observations, the main specifications do not include control variables. Since the baseline model does not contain any covariates at the student level, regressions with aggregated outcomes at the school-type level produce coefficients identical to those obtained using the underlying student-level data, given that observations are weighted accordingly (see Angrist and Pischke, 2009, Chapter 3.1.2). Only in robustness checks, are several control variables, X_{ist}, aggregated at the school-type level, such as the share of migrants and average class size, added. Comprehensive schools (*Gesamtschulen*) are excluded from the estimation sample because they offer all types of secondary school degrees. Therefore, it is not clear whether comprehensive schools should be considered as a treated or untreated school track. However, results are quite similar if comprehensive schools are included and either assigned to the treatment or the control group. In additional robustness checks, the outcomes of all non-Gymnasium school types within a state are aggregated, since considering outcomes of basic and middle school students separately might lead to wrong inferences if the reform affects the student composition in these tracks (see Section 6.3).

To ensure that the treatment group (non-Gymnasium tracks) receives the same total weight as the control group (Gymnasium) in each state, observations are weighted. Each non-Gymnasium school track is weighted by the share of 15-year-old students attending the respective school track as a fraction of all non-Gymnasium students in the state. Aside from

[23] Educational standards were implemented in the subjects German, math, and first foreign language for the basic school degree (grade level 9) in the school year 2005/2006 and for the middle school degree (grade level 10) one year earlier. See also Section 6.6.

the Gymnasium track, the only school type that exists in all German states, the number of non-Gymnasium tracks in a state varies between one and three (including comprehensive schools).[24] The weighting scheme implies that the total weight of all non-Gymnasium school tracks equals 1 in each state; similarly, the Gymnasium track in each state is also assigned weight 1.[25]

Estimation Method for Outcomes at State Level

A second type of "triple-differences" model uses performance data aggregated at the state level, such as percentiles and the shares of low- and high-performing students. While the low-performance measures, such as the 10th percentile, are largely determined by students in the non-Gymnasium tracks, the high-performance measures (e.g., 90th percentile) are basically only determined by Gymnasium students.[26] Instead of pooling the low-performance and high-performance measure of each state and estimating a "real" triple-differences model as for the school-type level outcomes, the reform effects of the state-level outcomes are estimated in a difference-in-differences model in which the dependent variable equals the difference between the low-performance and high-performance measure. This model has to be used since variables at the state level, e.g., the share of special education students, are included in robustness checks as control variables. State-level covariates, however, can affect only the coefficients on variables that also vary at the state level. In contrast, control variables at the state level *cannot* affect the estimated coefficients on variables that vary within states, which is the case for the reform indicator that would equal 1 for the low-performance measure in Bavaria in 2006 but would equal 0 for the low-performance measure. Note that if the specification does not contain control variables, the difference-in-differences models with differenced outcomes yield reform effects that are identical to those in the triple-differences model.

The following model is estimated for performance outcomes at the state level:

$$y_{kst}^{low} - y_{kst}^{high} = \alpha + \beta reform + \gamma_1 Bavaria + \Gamma_2 PISAdummies + \Gamma_3 X_{st} + \epsilon_{kst} \qquad (2)$$

where y_{kst}^{low} is the low performance measure y in subject k, state s, and PISA wave t. The reform indicator equals 1 for Bavaria in PISA-E 2006 and equals 0 otherwise. In robustness checks, the share of special education students and the share of Gymnasium students in a state, X_{st}, are added to the model. Since outcomes are not reported separately for each school

[24] While West German states traditionally have three school tracks (basic school, middle school, and Gymnasium), most East German states have only two tracks (integrated school and Gymnasium). Comprehensive schools exist in several German states, typically in addition to the other tracks.

[25] Alternatively weighting each school type with the share of 15-year-old students attending this type within a state yields very similar results. Results are also similar if each school track is weighted by the inverse of the number of different school tracks in a state.

[26] Note that the low-performance measures are also affected by the performance of students in special education and vocational schools. This issue will be addressed below.

type in PISA-E 2000 (except for Gymnasium), the pre-reform trends are only estimated with the difference-in-differences model and the PISA-E waves 2000 and 2003. In the pre-reform trend model, the reform indicator is replaced by the dummy variable *Bavaria 2003*, which equals 1 for Bavaria in PISA-E 2003 and 0 otherwise. In all models of this paper, standard errors are clustered at the state level, allowing for correlations of the error terms within states and over time.[27]

Theoretically, the control group for Bavaria can include all German states. However, a control state should not have experienced significant educational reforms that might have affected student performance. Several states introduced central exit exams in the basic track, middle track, or Gymnasium track between 2003 and 2006, that is, between the PISA-E pre-reform and post-reform period.[28] So as not to confound the effect of the Bavarian tracking reform with the effects of central exams introduction, these states are not part of the control group because central exams are frequently found to increase student performance (see, e.g., Jürges et al., 2005). Interestingly, all states that introduced central exams performed quite poorly in PISA-E 2000 and 2003. Therefore, these states might have introduced central exams precisely because of low performance levels. Among the seven excluded states are the two German states, Brandenburg and Berlin, that track students after six years of primary school, implying that all states in the control group track students into different school types after four years of primary school.

5 German PISA Data

Data on student performance are from the German extension studies (PISA-E) of the Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development (see Baumert et al., 2002; Prenzel et al., 2005, 2008, for details on the three PISA-E waves). PISA tested representative samples of 15-year-old students in math, reading, and science literacy in 2000, 2003, and 2006. The tests emphasize understanding and flexible and context-specific application of knowledge, and contain both multiple-choice and open-answer questions (see, for example, Prenzel et al., 2005, p. 15). The German extensions, PISA-E, used the same tests as the international PISA, but increased the samples to ensure representative sampling within each school type in each of the 16 German states.[29] Since the official data request to use the PISA-E student-level data was refused, aggregated performance outcomes published by the German PISA consortium

[27] Clustering standard errors at the school-type x state level yields quite similar results. This is not surprising as outcomes are not at the student level but already aggregated at the school-type x state level.

[28] These states are Brandenburg in 2003 and 2005, Hessen in 2004, Hamburg and Mecklenburg-Western Pomerania in 2005, and Berlin, Bremen, and Lower Saxony in 2006.

[29] PISA-E 2003, for example, tested a total of 44,580 students in 1,487 schools, with state samples ranging from 1,618 students in Saxony-Anhalt to 4,904 students in Hamburg. Sample sizes are similar for the other two PISA-E surveys.

are used instead (see Baumert et al., 2002; Prenzel et al., 2005, 2008). Performance is reported for 15-year-old students in PISA-E 2003 and 2006, which are representative of each school type within each of the 16 German states (see Prenzel et al., 2005, pp. 14 and 386 for PISA-E 2003).[30] In contrast, PISA-E 2000 reports performance measures for both 15-year-old students and ninth graders at the state level, and performance for ninth grade Gymnasium students, but does not report separate performance outcomes for the other school types, such as basic and middle schools.[31] In all three PISA-E waves, performance measures (of 15-year-old students) at the state level encompass the performance of students in general education schools as well as in special education and vocational schools.

The official PISA-E publications do not contain information on the share of basic and middle school students affected by the new tracking regime in Bavaria. However, PISA-E 2003 and 2006 provide the shares of students within each school type and state attending grades 7 to 11. Combining the grade level distributions of 15-year-old PISA-E students with official Bavarian statistics on the numbers of students transiting from primary schools to six-year middle schools or from basic schools to four-year middle schools in each school year (Bavarian State Ministry of Education, 2008, p. 114), one can compute the share of basic and middle school students in Bavaria for each PISA-E wave that is expected to be affected by the new tracking regime.[32] In the first PISA-E, wave 2000, only a small fraction, 9.2 percent of basic school students and 8.5 percent of middle school students, were tracked after grade 4. Note that these students were tracked early not because of the reform (which took place in April 2000), but because of the few early-adopting middle schools and middle schools participating in the pilot project that started in 1992.[33] Until PISA-E 2003, these shares increased slightly to 26.2 percent for basic school and 25.4 percent for middle school students. Because much less than half the students were affected by the new tracking regime,

[30] To be precise, student performance is reported for each general education school type that is attended by at least 5 percent of the 15-year-old students in a state. Thus, performance is not reported separately for special education students or students in vocational schools. Because the testing process is quite stressful for students from special education schools, the PISA-E consortium decided not to increase the sample of this student group. Instead, the average performance level of German special education students in the international PISA was used and added to the performance level of each German state in PISA-E according to the share of 15-year-old special education students in that state (see Prenzel et al., 2008, p. 381). This procedure is based on the assumption that competency levels of special education students do not differ between states.

[31] State-level performance was not published for Berlin and Hamburg in PISA-E 2000 due to low participation rates. Only Gymnasium performance was reported.

[32] For PISA-E 2000, the grade level distributions within each school type are imputed with those reported in PISA-E 2003. Note that even with individual student-level data (PISA-E 2000 being an exception), it is not possible to identify whether a specific student was tracked after grade 4 or after grade 6.

[33] The shares of middle school students affected by the new tracking regime are computed straightforwardly based on the official Bavarian statistics. In contrast, there are no numbers on basic school students in these statistics, as this group does not switch school type after starting basic school in grade 5. Hence, I assume that in each school year, the share of basic school students affected by the new regime is identical to the share of middle school students affected by the new regime. The small differences between basic and middle school shares affected by the reform is due to small differences in the grade distribution of 15-year-old students tested in PISA-E.

the PISA-E waves 2000 and 2003 are considered as pre-reform periods. In contrast, a large majority of students went through the new tracking regime in PISA-E 2006 (77.4 percent in the basic school track and 74.7 percent in the middle school track), making the PISA-E 2006 wave the post-reform period. Because the share of affected students did not increase from 0 to 100 percent between PISA-E 2003 and PISA-E 2006, the true reform effect will be underestimated.

To estimate the impact of the tracking reform requires performance data that are comparable over time. However, the possibility of comparing PISA performance across the three waves differs between the three subjects. Reading performance is comparable across all three PISA waves 2000, 2003, and 2006, but math performance is comparable only between 2003 and 2006 (see Prenzel et al., 2005, p. 73). Science performance is comparable between 2000 and 2003, but is not comparable between 2003 and 2006 due to changes in the concept of the tests (see Prenzel et al., 2008, p. 54f,149,383).[34] Even though not all subjects are comparable across all PISA waves, each specification pools all three subjects to be consistent across models. However, if the noncomparability of a PISA subject across two waves is the same for Bavarian basic and middle school students as it is for Bavarian Gymnasium students or students in the control states, including noncomparable subjects yields unbiased estimates. This assumption thus requires that the type of noncomparability be uncorrelated with the reform indicator in the empirical models. Indeed, robustness checks show that the reform effects are almost identical (results not shown) if noncomparable subjects are excluded, suggesting that this assumption holds in practice.

Table 1 presents summary statistics of outcomes at the state level. The statistics are reported separately for Bavaria and control states (average values) for each PISA-E wave. As all specifications pool the subjects math, reading, and science, the first rows report performance measures averaged across the three subjects. Since the PISA scaling reflects the difficulty of test items, different competency levels can be deduced on the basis of item difficulty and characterized (see OECD, 2002, Chapter 16).[35] Students below or at competency level 1 are individuals who will find it difficult to continue learning later in life and have trouble finding an apprenticeship position. Therefore, these students are considered "students at risk." In contrast, students who reach the highest competency level are considered top-performers.

Table 2 presents summary statistics of performance measures and control variables at the school-type level. Following the idea of the triple-differences model, statistics are presented

[34] Note that PISA had a special focus on reading literacy in 2000, with about half the testing time devoted to this subject. Math was the focus in PISA 2003 and science in PISA 2006. The international PISA scale was standardized to have a mean of 500 and a standard deviation of 100 for all subjects in PISA 2000 and whenever a subject was the focus (except for a standard deviation of 95 for science in PISA 2006).

[35] Formally, competency levels are defined such that students belonging to a given level can solve the items assigned to that level with a probability of 62 percent. Tasks related to higher competency levels will be solved with a much smaller probability and tasks related to a lower competency level will be solved with a much higher probability.

separately by treatment status of the school type (non-Gymnasium vs. Gymnasium), by treatment status of the state (Bavaria vs. control states), and by reform period (pre-reform 2003 vs. post-reform 2006). The non-Gymnasium statistics are based on weighted averages across all secondary school types in a state for which information was reported in the official PISA publications, excluding the Gymnasium track and comprehensive schools.[36] The table shows the evolution of student performance between PISA-E 2003 and PISA-E 2006, again averaged across the three subjects of math, reading, and science. Furthermore, important background information at the school-type level, such as the fraction of 15-year-old students attending the school tracks and the share of students with migration background (at least one parent born in a foreign country), is also provided by the official PISA-E publications.[37]

6 Results

This section first shows the development of student performance in Bavaria and control states graphically, then presents results on the reform effects on student performance. The following subsection provides evidence of similar pre-reform trends and presents numerous robustness checks. Then, grade repetition as a potential channel for the negative reform effects is investigated. Finally, results on differential effects for basic and middle school students and evidence of the persistence of reform effects are presented.

6.1 The Effects of Early Tracking on Student Performance

Figure 1 presents the development of students' reading performance in non-Gymnasium and Gymnasium schools in Bavaria and the control states across the three PISA-E waves. While the average reading performance of Gymnasium students developed similarly in Bavaria and control states both before and after the reform, the reading performance of non-Gymnasium students developed similarly only before the reform, but decreased in Bavaria after the reform. Figure 2 presents a very similar pattern for the shares of low- and high-performing students across all PISA subjects math, reading, and science. It shows that the shares of very low-performing and high-performing students developed similarly in Bavaria and control states before the reform, that is, between PISA-E 2000 and 2003. After the reform, the share of very low-performing students increased considerably in Bavaria, by about 3 percentage points relative to the control group, with the share of top-performing students falling slightly by less than 1 percentage point.[38] Aside from a small number of special education students and some students in vocational schools, students who achieve at most competency level

[36] See the table notes on how statistics are aggregated across non-Gymnasium school tracks.

[37] Average class size within each school track comes from Kultusministerkonferenz (2007).

[38] Note that the figure plots residuals from regressions of the low-performer shares and high-performer shares, respectively, on survey year dummies, thus eliminating Germany-wide changes between the individual PISA-E waves.

1 attend a non-Gymnasium school track, while students achieving the highest competency level mainly attend a Gymnasium. Therefore, Figure 2 again suggests that the tracking reform in Bavaria led to an increase in low-performing students, while not affecting the high-performing students in the Gymnasium track.[39]

Difference-in-Differences Models with Outcomes at State Level

The reform effects on student performance are first studied with mainly state-level outcomes in difference-in-differences models in which the dependent variable is the difference between a low-performance and a high-performance measure, reflecting the performance of students in the non-Gymnasium tracks and Gymnasium track, respectively (see Equation (5.2)). We are interested in the coefficients on the reform indicator, which equals 1 for Bavaria in 2006 and equals 0 otherwise. The negative and statistically significant coefficients for the performance gaps measured by differences between performance percentiles in Table 3 indicate that the performance of low-performers declined in Bavaria relative to the control group, keeping the level of high-performing students constant (see Columns (1) and (2)). To get an idea of the magnitude of the effect size, note that the competency increase in one school year is on average about 40 points in PISA 2003 (cf. Prenzel et al., 2005) and 25 to 30 points in PISA 2006 (see Prenzel et al., 2008, p. 59). Hence, the effect of 11 PISA points is nontrivial, reflecting the performance increase of between a quarter and half a school year. As already suggested by Figure 2, the reform increased the share of very low-performing students by 5.3 percentage points (see Column 3). This is a sizeable effect, considering that in Bavaria 13.9 percent of all 15-year-old students were "students at risk" in PISA-E 2003 (this figure was 21.0 percent, on average, in the control states). The reform effect on mean reading performance is about 15 PISA points, which roughly equals the increase in competencies gained during half a school year (see Column 4).

Note that while the state-level outcomes in Columns (1) to (3) are identical across all three PISA-E waves, reading performance in Column (4) is constructed differently in PISA-E 2000 than in PISA-E 2003 and 2006 due to data availability. Gymnasium performance refers to 15-year-old students in PISA-E 2003 and 2006, but refers to ninth graders in PISA-E 2000 because it is not reported for 15-year-old students (cf. Baumert et al., 2002). Assuming that the grade level attendance of 15-year-old Gymnasium students developed similarly in Bavaria and the control group between PISA-E 2000 (not reported) and PISA-E 2003, the performance development between PISA-E 2000 (ninth-grade students) and 2003 (15-year-old students) is quite comparable. Note that the share of Gymnasium students

[39] The performance measures of Figure 1 are used as dependent variables in Column (4) of Tables 3 and 4 to quantify the reform effect and in Tables 6 and 7 to study the pre-reform trends. The performance measures of Figure 2 are used as dependent variables in the same tables in Column (3). The performance measures are described in detail in the next section. Note that there are slight differences in the measures between PISA-E 2000 and the following two waves.

attending a ninth grade (59.6 percent) is quite close to the respective share in the control group (61.3 percent) in PISA-E 2003, as is the average grade attended in Gymnasium in Bavaria (9.19) and the control group (9.27) (see Table 2).[40] Non-Gymnasium average reading performance in all waves is computed on the basis of the mean reading performance of all 15-year-old students in the state, the mean reading performance in Gymnasium, and the share of Gymnasium students in the state, with the latter two measures referring to ninth graders in PISA-E 2000 and to 15-year-old students in PISA-E 2003 and 2006.[41]

One issue that might bias the reform coefficient is that the low-performance measures at the state level not only reflect the performance of basic and middle school students, the treatment group, but also that of special education students. These students typically attend special education schools either since the beginning of compulsory schooling (at age six) or directly after primary school. Hence, special education students, who perform very poorly in the PISA tests, typically never attend a basic or middle school and therefore are not affected by the tracking reform in Bavaria. This means that the reform coefficients might be biased if the share of special education students developed differently in Bavaria than in the control group between the PISA-E waves. Similarly, the top-performance measures, mainly reflecting the performance of Gymnasium students, the vast majority of whom never attended a basic or middle school, might bias the reform estimates if the share of Gymnasium students changed differently in Bavaria and the control group.[42] However, controlling for the share of special education students and the share of Gymnasium students changes the reform effects only marginally (see Table 4). Another issue is that the models pool all three PISA-E subjects (math, reading, and science), although only reading is comparable across all three PISA-E waves 2000, 2003 and 2006. Using students' reading literacy only yields similar, even slightly stronger, results. Furthermore, reform effects are also very similar if either the first pre-reform period, PISA-E 2000, or the second pre-reform period, PISA-E 2003, are excluded from the analysis.[43]

[40] That the grade level distribution of 15-year-old Gymnasium students in Bavaria and the control group was similar in PISA-E 2000 (not reported) and 2003 is suggested by the very similar development of the grade level distribution in Bavaria and the control states between PISA-E 2003 and PISA-E 2006.

[41] Note that the state mean of reading performance is, analogously to mean Gymnasium performance, also available for ninth grade students in PISA-E 2000. However, this state-level measure does not include the reading performance of students in special education schools, thus making this measure less comparable to the state-level performance in 2003 and 2006, which do include the performance of special education students.

[42] The share of vocational school students in a state is not included as an additional control variable because 15-year-old vocational school students previously attended either a basic or a middle school, thus rendering the share of vocational school students an endogenous regressor with respect to the Bavarian reform.

[43] The reform coefficient becomes smaller and statistically insignificant only in the model with the performance gap between the 10th and 90th percentile if the PISA-E 2003 wave is excluded.

Triple-Differences Models with Outcomes at School-Type Level

The following models estimate the impact of the tracking reform on student performance with performance outcomes at the school-type level (see Equation (5.1)). In these models, all school types that typically lead to either a basic school degree (*Hauptschulabschluss*) or middle school degree (*mittlere Reife*) are considered as *treated* school type, since this student group corresponds to the students in basic and middle schools in Bavaria who were affected by the tracking reform.[44] The *control* school type consists of the Gymnasium track in each state because Gymnasium students in Bavaria were not affected by the reform. As noted above, comprehensive schools (*Gesamtschulen*) are excluded from the estimation sample because comprehensive schools contain all types of secondary school tracks.

Table 5 presents the reform effects on student performance with outcomes aggregated within each school type. Similar to the results of the state-level model (see Column (4) in Table 3), the reform has a negative and statistically significant effect on mean performance (see Column (1) in Table 5). The coefficient on the reform dummy, which equals 1 for basic and middle schools in Bavaria for PISA-E 2006 and equals 0 otherwise, indicates that the reform lowered mean student performance across math, reading, and science by 11.5 PISA points. Column (2) shows that the reform also widened performance dispersion. To the extent student performance is correlated with family background, this suggests that equality of educational opportunities is lower when students are tracked early. Column (3) corroborates the state-level results that the share of very low-performing students is substantially larger when students are tracked two years earlier, and that fewer students achieve a high performance level under early tracking (Column 4). The coefficients on the other covariates have the expected signs. The coefficient on Bavaria shows that Bavaria outperforms the control states with respect to each performance measure. The coefficient on PISA-E 2006 corroborates the well-known finding that German students performed better in PISA-E 2006 than in the previous PISA-E waves (cf. Prenzel et al., 2008). The coefficients on the dummy indicating non-Gymnasium school tracks in Bavaria show, as expected, that students in the non-Gymnasium school tracks perform better in Bavaria than in other states. The coefficients on the interaction between PISA-E 2006 and non-Gymnasium tracks are inconclusive, and the coefficients on the interaction between Bavaria and PISA-E 2006 are rather small for all outcomes.[45]

Causal interpretation of the reform effect in the triple-differences model relies on the assumption that the development of the performance difference between non-Gymnasium

[44] In addition to basic and middle schools, the treated group also consists of integrated schools (called *Mittelschule* or *Regelschule*), which offer both basic and middle school degrees and typically are found in East German states.

[45] The reform effects are quite similar if science, the subject not comparable across PISA waves 2003 and 2006, is excluded from the estimation sample. Furthermore, results are very similar if comprehensive schools are included and either assigned to the treated or the untreated school type. Finally, results are robust to excluding integrated schools that lead to either a basic or middle school degree.

school types and Gymnasium would have been the same in Bavaria and the control states in the absence of the tracking reform. However, this assumption might not hold if the control states that performed considerably worse than Bavaria in PISA-E 2003 exerted more pressure on non-Gymnasium school types than on the Gymnasium schools in an effort to improve student performance. Indeed, this is likely the case, since Gymnasium students outperform non-Gymnasium students in each state. Therefore, political pressure on and small changes in the low-performing non-Gymnasium tracks in the control states might have led to a stronger increase in student performance in the non-Gymnasium tracks than in the Gymnasium track in the control states. Such a development would induce reform coefficients indicating that performance in the non-Gymnasium tracks in Bavaria worsened. To test whether the estimated coefficients are driven by such a catching-up process by the non-Gymnasium tracks in the low-performing control states, I use only those three states as control states that, next to Bavaria, belonged to the best-performing states in the pre-reform period PISA-E 2003.[46] Using only these three high-performing states as the control group leads to results very similar to those in Table 5, with the magnitude of the effects being even slightly larger for mean performance and for the share of high-performing students.[47] This finding suggests that the estimated coefficients on the reform indicator are due to a negative effect of the Bavarian tracking reform on student performance and not to a catching-up process in the non-Gymnasium tracks in the control states.

Given that the state-level performance measures encompass the performance of students in special education and vocational schools, it is reassuring that the reform effects estimated with performance outcomes at the state level are very similar to the effects based on school-type level outcomes. First, the reform effect on mean reading performance is -15 PISA points in the model in which non-Gymnasium performance includes special education and vocational school students (see Column (4) in Table 3) and -16.7 PISA points in a school-type level model in which non-Gymnasium performance excludes the performance of special education and vocational school students (not shown in table; see Column (1) in Table 5 for the respective effect across all subjects). Second, the effect on the share of very low-performing students is 5.3 percentage points with state-level outcomes (see Column (3) in Table 3) and is 5.6 percentage points in a model in which the share of low-performers only includes basic and middle school students.[48] This finding suggests that models with outcomes

[46] To be precise, I use the average performance across all three PISA-E subjects (math, reading, and science) to determine the best-performing states in PISA-E 2003. Right after Bavaria at the very top, are Saxony, Baden-Württemberg, and Thuringia.

[47] Due to the much smaller sample sizes, standard errors increase considerably, which reduces the significance levels of the estimated coefficients.

[48] Note that the dependent variable in the state-level model actually is the difference between the share of students achieving at most competency level 1 and the share of students in a state achieving the highest competency level. However, using also the top-performance measure for the dependent variable should not affect the state-level results because the share of students achieving the highest competency level developed very similarly in Bavaria and the control states, especially between PISA-E 2003 and 2006 (see Table 1).

at the school-type level that ignore the performance of vocational school students who were also affected by the reform yield unbiased reform effects.

6.2 Pre-Reform Trends

This section provides analytical evidence of similar pre-reform trends, that is, between PISA-E 2000 and PISA-E 2003, when the majority of basic and middle school students in Bavaria attended school under the old tracking regime (see Table 1). Because performance results were not published at the school-type level in PISA-E 2000 (except for Gymnasium), pre-reform trends will be analyzed mainly with state-level outcomes in difference-in-differences models in which the dependent variable is the difference between a low-performance and a high-performance measure (see Equation (5.2)).

Table 6 shows that the performance gap between low- and high-performing students developed quite similarly in Bavaria and control states before the reform, as indicated by the statistically insignificant interaction term *Bavaria 2003* between Bavaria and PISA-E wave 2003. The coefficient is very small for the difference between the 5th and 95th performance percentile (Column 1), for the difference between the share of students achieving at most competency level 1 and the share of students achieving the highest competency level (Column 3), and for the difference between the mean reading performance of non-Gymnasium and Gymnasium students (Column 4). Only the coefficient on the performance gap between the 10th and 90th percentile is somewhat larger, though statistically insignificant (Column 2). Note that the outcomes in Columns (1)-(3) are pooled across all three PISA-E subjects (math, reading, and science). The results are quite similar if math, the subject that is not comparable between PISA-E 2000 and 2003, is excluded from the estimation sample. Again, a different development in the shares of special education and Gymnasium students in Bavaria and in control states might bias the effects. Therefore, the models of Table 7 include both student shares within a state, finding that the pre-reform trends are unaffected by these control variables.

6.3 Robustness Checks

This section shows that the reform effects estimated with the triple-differences models on the basis of school-type-level outcomes are robust to including control variables at the school-type level, such as PISA-E participation rates or the share of migrants, robust to aggregating performance measures across all non-Gymnasium school types, and robust to excluding any one of the eight control states from the estimation sample.

Several factors at the school-type level might affect the reform estimates. One of the most obvious is student composition of basic, middle, and Gymnasium tracks, which might influence the reform coefficient if the track composition changed differently in Bavaria than in the control states. The tracking reform might have particularly affected the share of middle school students in Bavaria. As parents generally found the six-year middle school quite attractive, well-performing students might have attended a middle school instead of Gymnasium after the reform, which possibly increased mean performance in the middle school track and decreased the mean performance of Gymnasium students. Another potential confounding factor is the share of migrants in a school track as migrants are generally found to have performance levels than non-migrants (cf. Prenzel et al., 2008). Furthermore, differential changes in PISA-E participation rates might affect the reform estimates if the likelihood of taking the test is correlated with the student's performance, especially with lower-performing students being more likely not to participate (see Chapter 1.4 in Prenzel et al. (2008) and Chapter 2.4 in Prenzel et al. (2008) on nonparticipation rates in PISA-E 2003 and 2006). Finally, average class size, a measure of classroom resources, might affect student performance.[49] Table 8 includes all these factors at the school-track level. The reform coefficients barely change when these school-track level variables are taken into account, indicating that the negative reform effects are not mediated through any of these channels.[50]

To this point, the triple-differences models have included student performance of each non-Gymnasium school track separately. However, the effects could be seriously biased in models with separate school type outcomes if the reform changes the student composition of basic and middle school tracks in Bavaria. Suppose, for example, that the introduction of the six-year middle schools increased the attractiveness of middle schools such that better-performing students decided to attend a middle school instead of a basic school. In this case, mean performance will necessarily fall in the basic school track. However, mean performance may even decrease in the middle school track if the newly attracted students perform on average worse than the average middle school student (before the reform). In this scenario, mean performance falls in both basic and middle school tracks, and the reform coefficient would be negative even if the reform had no negative effect whatsoever on student performance. To check whether such changes in student composition in Bavarian basic and middle schools affect the reform estimates, I compute one performance measure for all non-Gymnasium students by aggregating the outcomes of all non-Gymnasium school types within each state. For the aggregation, each school track is weighted by the share of 15-year-old students attending the respective school track as a fraction of all students attending any non-Gymnasium school track.[51]

Table 9 presents results from triple-differences models with the performance measure aggregated across all non-Gymnasium school types. The coefficients on the reform dummy are very similar to those models in which the performance of the non-Gymnasium tracks

[49] Woessmann (2010), however, finds that a state's average class size is not statistically significantly related to student performance.

[50] Average class size in grades *5-6* refers to two years prior to the PISA-E school year, while average class size in grades *7-9/10* refers to the PISA-E school year.

[51] While the (weighted) aggregation of the other performance outcomes is straightforward, the aggregated standard deviation across all non-Gymnasium school tracks is computed using the formula in Fahrmeir et al. (2003), p. 72.

enters separately, indicating that compositional changes do not affect the reform estimates. The proportionally largest, though still small, coefficient difference is that for the standard deviation, which is plausible since the aggregated performance variance takes into account both the variance within the individual school tracks and the difference in mean performance between the non-Gymnasium types. Thus, the effect on the standard deviation theoretically could differ substantially between the aggregated non-Gymnasium outcome and the separate non-Gymnasium outcomes.[52] The other, very small, coefficient differences across models with aggregated and non-aggregated outcomes are due to differences in the weighting scheme of the observations. In sum, the estimated reform effects are very similar regardless of whether the performance outcomes of the non-Gymnasium tracks are aggregated or not.

The key identification issue concerns whether the coefficients on the reform indicator do, indeed, reflect the effects of the Bavarian tracking reform or, rather, changes in the control states. To analyze whether the reform effects are driven by one of the control states, I conduct several robustness checks by reestimating the triple-differences models with and without controls (see Tables 5 and 8), with each control state excluded individually. The results show that the reform effects are very similar when any one of the eight control states is excluded from the estimation sample.[53] Only three coefficients become statistically insignificant in the specification with control variables if Rhineland-Palatinate is excluded; however, these coefficients are not statistically different from those of the model with all control states. These findings increase confidence that the estimated reform coefficients are driven by the tracking reform in Bavaria and not by severe performance changes in one of the control states.

6.4 Reform Effects on Grade Repetition

Another way the tracking reform might lower performance level is increased grade repetition by students affected by the reform. This is a potentially important channel since the target population of PISA is a certain age group (15-year-old students) and not a certain grade level (e.g., ninth graders). The 15-year-old students participating in PISA-E basically attended four different grade levels, ranging from grade 7 to grade 10.[54] There are two reasons why students of the same age might be in different grades. First, they might have started primary school one year later or one year earlier than officially mandated by their date of birth.[55]

[52] Note that the coefficient is statistically insignificant because the standard error increased substantially. The reform coefficient would be statistically significant with the standard error from the model with separate non-Gymnasium outcomes.

[53] Results are available from the author upon request.

[54] Very few 15-year-olds who participated in PISA-E attended grade 6 or grade 11. Therefore, the share of sixth graders and the share of 11th graders are ignored in the following analysis.

[55] With late enrollment being much more common than early enrollment, the share of late-enrolled students is substantial and varies considerably across school types. In Bavaria, for example, 16.9 percent of the basic track students in PISA-E 2003 enrolled late in primary school, but only 9.0 percent of middle school students, and only 5.5 percent of students in Gymnasium (see Prenzel et al., 2005, p. 176).

Second, students might need to repeat a grade if their academic achievement, as measured by grades on a report card, is too low. Like late school enrollment, the probability that a student repeats a grade is higher in the non-Gymnasium tracks.

The importance of increased grade repetition for the negative effect of the reform on student performance is investigated with the shares of students in different grade levels, which are reported separately for each school track in each state.[56] Thus, the grade repetition channel is investigated indirectly; it cannot be analyzed directly because the share of students that repeated a grade throughout their school career is reported only for PISA-E 2003, not for PISA-E 2006. The analysis based on the grade level distribution identifies the effects of grade repetition if the share of early- or late-enrolled students did not change due to the reform, which is plausible. That the share of early- or late-enrolled students changed differently in Bavaria than in the control states is especially unlikely since the average grade attended by Gymnasium students changed very similarly in Bavaria and the control states between PISA-E 2003 and PISA-E 2006. Based on the triple-differences model with control variables, Table 10 additionally includes the share of students that attend grade 7, grade 8, and grade 10, respectively, within each school type in each state.[57] The reform effects become considerably smaller once the grade distribution of the tested students is taken into account. The effect on mean performance, for example, decreases by half (Column (1)). The effect becomes statistically insignificant, partially due to the low statistical power of the small sample with numerous control variables. Similarly, the effect on the share of low-performing students, while still in the same direction, is much smaller and statistically insignificant (Column 3). This suggests that the reform especially induced more 15-year-old students to attend very low grade levels (i.e., grade 7 or 8), since it is especially the students in low grade levels who are likely to achieve only a low competency level. In contrast, the reform effect on the dispersion of performance is unaffected by controlling for grade level attendance (Column 2). Similarly, the negative reform effect on the share of high-performing students is unaffected (Column 4).[58] Finally, note that the coefficient signs on the grade level shares are as expected, with the share of students attending a low grade (e.g., seventh grade) related to lower performance levels and the share of students attending a high grade (e.g., tenth grade) related to higher performance levels.

Table 11 presents direct evidence that the reform affected the grade level attendance of 15-year-old students. Column (1) shows that the average grade attended by 15-year-old students decreased by 0.09 grade levels in Bavarian basic and middle schools. Column (2)

[56] Table 2 shows that 15-year-old students in Gymnasium tend to attend higher grade levels than 15-year-old students in the non-Gymnasium tracks, both in Bavaria and in the control group. Furthermore, the statistics reveal that the students attended on average higher grade levels in the PISA-E 2006 survey than in the 2003 survey, both in Bavaria and in the control states in both the non-Gymnasium tracks and Gymnasium.

[57] The share of students attending a ninth grade, the most common grade level of 15-year-old students within each school track, is omitted.

[58] Note that the enormously increased standard error renders the coefficient statistically insignificant, whereas the magnitude of the coefficient remains virtually unchanged.

excludes Bavarian middle schools and finds the same effect size for basic track students, while Column (3) indicates that the decrease of average grade level was somewhat smaller in the middle school track. While the effects on average grade level attended are moderate, the reform could have especially increased the share of students attending a low grade, as suggested by the much smaller reform effect on the share of very low-performing students (see Column (3) in Table 10). Because the majority of students within each school track attends the ninth grade, I define grades 7 and 8 as low grade levels for 15-year-old students. Column (4) shows that the share of students attending a low grade level increased significantly by 7 percentage points in the Bavarian non-Gymnasium tracks. By excluding the middle track in Bavaria, Column (5) shows that the share of students attending a low grade particularly increased in the basic school track. The magnitude of the effect, 9 percentage points, is sizeable compared to the baseline of 15.4 percent of Bavarian basic school students attending a low grade level in PISA-E 2003.[59] The impact on low grade-attendance of 3.3 percentage points is considerably smaller in the middle track (Column 6).[60]

Importantly, the average grade level attended and the share of students in low grades in the Gymnasium track developed very similarly in Bavaria and the control states between PISA-E 2003 and 2006 (results not shown). This strongly indicates that the negative effect on the grade level attendance in the basic and middle school tracks in Bavaria does indeed reflect a negative reform effect on grade repetition and not a change in late or early enrollment in Bavarian primary schools. Therefore, increased grade repetition partly explains the negative reform effect on the performance of basic and middle school students. The strong effect on the share of students in low grade levels in the basic track might especially explain the rather strong reform effect on the share of low-performing students. However, since grade repetition explains only a part of the negative effects on student performance, there must be other factors, such as peer effects, that cause the decrease in student performance.

6.5 Differential Effects on Basic and Middle School Students

Thus far we have investigated the impact of the tracking reform on student performance for basic and middle school students jointly. However, the reform effects might be different for the two tracks. Even though peer groups have changed in grades 5 and 6 for students in both tracks, the change could have different effects for the lower-performing students in the basic track and the higher-performing students in the middle track. To investigate the reform effect on basic school students, I exclude the Bavarian middle schools from the estimation

[59] Since only a few students attend a seventh or tenth grade in basic schools, students induced to attend a low grade level mostly attend the eighth grade instead of the ninth grade. Therefore, the effect size of 9 percentage points more students attending a low grade directly translates into a decrease in the average grade attended by 0.09 grade levels (see Column 2). The analogous translation of effect sizes is not straightforward for the middle school track because a considerable fraction of middle school students attend grade 10.

[60] Results are very similar if the effect on the basic track is estimated only on the basis of basic school tracks and Gymnasium tracks; similarly for effect on middle track.

sample such that the reform indicator reflects the effect on Bavarian basic schools. The effects on the performance of basic school students are very similar to those for basic and middle school students together (see Table 12). One exception is the negative effect on the share of high-performing students, which is much smaller (and statistically insignificant) for basic school students (see Column 4). The reduction of this effect is reasonable as the share of high-performing students in the basic track is rather low anyway; across all subjects, the respective share was only 11.1 percent in Bavarian basic schools in PISA-E 2003 and 8.7 percent in PISA-E 2006.[61]

Similarly, the basic school track in Bavaria is excluded from the sample to estimate the reform impact on middle school students. While the effect on the performance dispersion is slightly larger for the middle track, the effect on mean performance is very similar for basic and middle school students, with about 12 PISA points or, equivalently, the competency increase of almost half a school year (see Table 13). However, there are notable differences in the effects on the shares of very low-performing and high-performing students across the two school tracks. First, the effect on the share of very low-performing students is stronger for basic school students than for middle school students. This finding is plausible since very low-performing students are more likely to be found in basic schools: the share of students achieving at most competency level 1 in PISA-E across all subjects is about 30 percent in basic schools but only about 2 percent in middle schools in Bavaria. Second, while there is basically no effect on the share of high-performing students in basic schools (see above), the effect is quite strong in middle schools (-6.9 percentage points). Again, this finding is reasonable considering that high-performing students are very unlikely to attend basic schools, but are much more likely to attend a middle school.[62] Overall, the reform effects on basic and middle school students are similar. However, there are important differences with respect to the shares of low-performing and high-performing students, which are reasonable given the differences in competency levels across the two tracks.

6.6 Persistence of the Reform Effects

Because all results so far are based on student performance up to year 2006 only, the question arises whether the Bavarian tracking reform has a negative effect only on those students who went to school when the reform was implemented or whether it has permanent negative effects also on student cohorts who attended school when the new system was more established. In principle, the tracking reform might have had only short-run effects because

[61] While the reform effects on basic school students in Table 12 are estimated by comparing the basic school track in Bavaria with all non-Gymnasium school types in the control states, the effects are similar if only basic schools in the control states are used as the comparison group, with only the effect on mean performance decreasing somewhat.

[62] Similarly, using only the middle school tracks instead of all non-Gymnasium tracks in the control states as comparison group yields very similar results, with a somewhat smaller effect on the share of low-performing students and a stronger effect on the share of high-performing students.

its implementation might have been accompanied by difficulties and problems that affected the student cohorts at the time of the implementation. For example, many recently hired teachers in middle schools had no teaching experience and the more experienced middle school teachers had to teach a completely revised curriculum for the first time. This might have negatively affected student performance in the first school years after the reform, but have had no negative effect on later student cohorts.

Instead of extending the international PISA study 2009 to compare student performance across German states, the Secretariat of the Standing Conference of the Ministers of Education (*Kultusministerkonferenz*, or KMK) tested reading competencies of ninth grade students as part of a quality check of recently introduced educational standards (*Bildungsstandards*). These standards were implemented by the KMK in the school year 2003/2004 for students aiming at a middle school degree and are binding for all 16 German states. While the educational standards tested in 2009 actually apply to students striving for a middle school degree, the aim of the student assessment was more broad—to compare competencies of all ninth graders in all general education school tracks across German states.[63] Note that all Bavarian basic and middle school students who participated in the educational standards test in 2009 attended school under the new tracking regime since the state-wide implementation of the six-year middle schools was complete by school year 2003/2004. Testing the educational standards was linked to the international PISA 2009 survey in two ways. First, the student samples overlap. Two ninth grade classrooms in 201 general education schools that participated in PISA 2009 were tested in PISA on the first day. On the second day of testing, these ninth graders participated in the educational standards test. Additional schools were sampled to ensure comparability across German states. In the final sample, 36,605 ninth grade students from 1,655 classrooms and 1,466 schools participated in the reading assessment based on the educational standards. Second, the competency scale in the educational-standards-based test was linked to the international PISA scale to ensure maximal comparability with the competency levels in PISA. One aspect was that the reading competency scale was standardized such that the mean (496 points) and the standard deviation (92 points) are identical to the reading results of German ninth graders in PISA 2000. The fact that the three best-performing states on the educational-standards-based reading assessment were also the best three-performing states in the PISA-E 2006 reading ranking strongly suggests that the reading scales of the educational standards 2009 and PISA-E reading performance 2006 are comparable; furthermore, two of the three worst-performing states in 2009 were the two worst-performing states in PISA-E 2006.[64]

[63] The educational standards can be downloaded (only in German) from http://www.kmk.org/bildung-schule/qualitaetssicherung-in-schulen/bildungsstandards/dokumente.html. In addition to reading, listening and orthography competencies in the subject of German and reading and listening competencies in the subject of English were tested in 2009. For details and results of the educational-standards-based assessment, see Köller et al. (2010).

[64] The state ranking on the basis of mean reading performance in 2006 and 2009 is similar looking at all German states. Also note that the difference in reading performance between the best-performing (512

Table 14 presents evidence that the negative effects of the tracking reform on students' reading performance are indeed permanent. The dependent variable is the mean reading performance of non-Gymnasium and Gymnasium students within each state, which was previously used to investigate the reform effect across the three PISA-E waves (see Column (4) in Tables 3 and 4).[65] The specifications in Columns (1) and (2) pool all three PISA-E waves and the educational-standards-based test of 2009. The coefficient on the reform indicator, which equals 1 for non-Gymnasium students in Bavaria for the surveys in 2006 and 2009 and 0 otherwise, indicates that early tracking in Bavaria lowers students' reading performance even for later student cohorts. The reform coefficient of -13.6 points in Column (1) reflects the joint effect of the surveys 2006 and 2009; the reform dummy in Column (2) reflects the effect on reading performance in 2006 only.[66] The dummy *reform 2009*, which equals 1 for non-Gymnasium students in Bavaria in 2009, is quite small and statistically insignificant, indicating that there is hardly any difference in the reform effect on students' reading performance between 2006 and 2009. Columns (3) and (4) exclude the PISA-E 2000 wave from the sample, finding very similar results if the reform effects are identified on the basis of the same pre-reform period (PISA-E 2003) used in the triple-differences models. In sum, the effect of early tracking is not only due to immediate reform changes or difficulties, but appears to have negative effects on later student cohorts too. Because results based on the PISA-E waves show that effects are similar across subjects, this suggests that the reform also likely has permanent negative effects on students' math and science performance.

7 Conclusion

Using students' performance in primary school to sort students into different secondary school types is common practice in European countries. This paper studies the impact of early tracking on the level and distribution of student performance, exploiting an education reform in the German state of Bavaria that advanced by two years the tracking of basic and middle school students. Performance data come from the German extensions (PISA-E) of several PISA waves that tested students under both the old and new tracking regimes. The reform effects are identified in a triple-differences model, taking advantage of the fact that the reform affected students in the basic and middle school tracks in Bavaria, but did

points) and worst-performing state (474) in 2006 is very similar to the respective difference in 2009 (509 vs. 469). It is especially remarkable that the state ranking is very similar across both assessments because 15-year-old students were tested in PISA-E 2006, whereas the educational standards assessment tested ninth grade students.

[65] Note that reading performance refers to ninth grade students in 2009 and to 15-year-old students in 2003 and 2006. Similar to the PISA-E waves, the mean reading performance for non-Gymnasium students in 2009 (which is not published) is computed on the basis of the mean reading performance of the state, the mean reading performance of Gymnasium students, and the share of Gymnasium students within the state. For differences of the dependent variable across the three PISA-E waves, see Section 6.2.

[66] Note that this coefficient is identical to the coefficient in Column (4) of Table 3, which identifies the effect for 2006 only.

not affect Gymnasium students in Bavaria or students in any school track in other German states.

The results consistently indicate that the reform reduced the performance level and increased inequality in performance. The performance level falls because both the share of low-performing students increases and the share of high-performing students decreases. Additional results indicate that early tracking lowers the performance level both in the more vocational-oriented basic school track and in the middle school track with a more academic curriculum. Part, but not all, of the negative effect is mediated through increased grade repetition, which implies that tested 15-year-old students in basic and middle schools on average attend lower grade levels. Finally, the negative effect on student performance seems to be persistent and not solely due to difficulties that might have occurred with implementation of the reform.

It should be kept in mind that the effects of a reform necessarily depend both on the specifics of the reform and on the environment in which it takes place. Therefore, one should be careful when transferring the findings of the Bavarian reform to other countries or education systems. For example, tracking decisions elsewhere might be completely different from those made in Bavaria, where school grades at the end of primary school are the crucial determinant, and parents' wishes basically do not count. Furthermore, higher mobility across school tracks might mitigate the negative effects of early tracking. Also, teaching styles (e.g., whole classroom vs. individual instruction) and support for low-performing students might be different in other countries; or effects might differ in a system where grade repetition is not an option. Another important issue, which requires the use of micro data, concerns whether the effects of early tracking on student performance depend on family background. If early tracking primarily hurts students from low socioeconomic backgrounds, such a system would not only decrease efficiency but also increase inequality of opportunities. Because tracking systems and educational environments differ across countries, more studies are needed to better understand the consequences of tracking students into different school types.

References

Ammermüller, A. (2005): "Educational Opportunities and the Role of Institutions," ZEW Discussion Paper No. 44.

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

Bauer, P. and R. Riphahn (2006): "Timing of School Tracking as a Determinant of Intergenerational Transmission of Education," *Economics Letters*, 91, 90–97.

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

Baumert, J., M. Becker, M. Neumann, and R. Nikolova (2009): "Frühübergang in ein grundständiges Gymnasium - Übergang in ein privilegiertes Entwicklungsmilieu? Ein Vergleich von Regressionsanalyse und Propensity Score Matching," *Zeitschrift für Erziehungswissenschaft*, 12, 189–215.

Bavarian State Ministry of Education (2008): *Schule und Bildung in Bayern 2008: Zahlen und Fakten*, Munich: Bavarian State Ministry of Education.

Bavarian Statistical Office (1996): *Die allgemeinbildenden Schulen in Bayern Schuljahr 1995/96: Realschulen, Realschulen für Behinderte, Abendrealschulen*, Munich: Bavarian Statistical Office.

Betts, J. (2011): "The Economics of Tracking in Education," in *Handbook of the Economics of Education*, ed. by E. Hanushek, S. Machin, and L. Woessmann, Elsevier, vol. 3, 341–381.

Brunello, G. and D. Checchi (2007): "Does School Tracking Affect Equality of Opportunity? New International Evidence," *Economic Policy*, 22, 781–861.

Dustmann, C. (2004): "Parental Background, Secondary School Track Choice, and Wages," *Oxford Economic Papers*, 56, 209–230.

Fahrmeir, L., R. Kuenstler, I. Pigeot, and G. Tutz (2003): *Statistik: Der Weg zur Datenanalyse*, Berlin: Springer.

Galindo-Rueda, F. and A. Vignoles (2007): "The Heterogeneous Effect of Selection in UK Secondary Schools," in *Schools and the Equal Opportunity Problem*, ed. by L. Woessmann and P. Peterson, Cambridge, Mass.: The MIT Press, 103–128.

Hamermesh, D. and S. Trejo (2000): "The Demand for Hours of Labor: Direct Evidence from California," *Review of Economics and Statistics*, 82, 38–47.

Hanushek, E. (2002): "Publicly Provided Education," in *Handbook of Public Economics*, ed. by A. Auerbach and M. Feldstein, Elsevier, vol. 4, 2046–2141.

Hanushek, E. and L. Woessmann (2006): "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries," *Economic Journal*, 116, C63–C76.

——— (2008): "The Role of Cognitive Skills in Economic Development," *Journal of Economic Literature*, 46, 607–668.

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

Köller, O., M. Knigge, and B. Tesch, eds. (2010): *Sprachliche Kompetenzen im Ländervergleich*, Münster: Waxmann.

Kropf, M., C. Gresch, and K. Maaz (2010): "Überblick über die rechtlichen Regelungen des Übergangs in den beteiligten Ländern," in *Der Übergang von der Grundschule in die weiterführende Schule - Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten*, ed. by Bundesministerium für Bildung und Forschung (BMBF), Bonn, 399–429.

Kultusministerkonferenz (2007): *Schüler, Klassen, Lehrer und Absolventen der Schulen 1997 bis 2006. Dokumentation Nr. 184*, Bonn: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs.

Lohmar, B. and T. Eckhardt (2010): "The Education System in the Federal Republic of Germany 2008: A Description of the Responsibilities, Structures and Developments in Education Policy for the Exchange of Information in Europe," Bonn: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs.

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

Meier, V. and G. Schütz (2008): "The Economics of Tracking and Non-Tracking," *Zeitschrift für Betriebswirtschaft*, 78, 23–43.

Mühlenweg, A. (2008): "Educational Effects of Alternative Secondary School Tracking Regimes in Germany," *Schmollers Jahrbuch (Journal of Applied Social Science Studies)*, 128, 351–379.

OECD (2001): *Knowledge and Skills for Life: First Results from PISA 2000*, Paris: OECD.

——— (2002): *PISA 2000 Technical Report*, Paris: OECD.

——— (2004): *Learning for Tomorrow's World: First Results from PISA 2003*, Paris: OECD.

Pekkarinen, T., R. Uusitalo, and S. Pekkala (2009): "School Tracking and Intergenerational Income Mobility: Evidence from the Finnish Comprehensive School Reform," *Journal of Public Economics*, 93, 965–973.

Pietsch, M. and T. Stubbe (2007): "Inequality in the Transition from Primary to Secondary School: School Choices and Educational Disparities in Germany," *European Educational Research Journal*, 6, 424–445.

Pischke, J.-S. and A. Manning (2006): "Comprehensive versus Selective Schooling in England in Wales: What Do We Know?" NBER Working Paper No. 12176.

Prenzel, M., C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, and R. Pekrun, eds. (2008): *PISA 2006 in Deutschland: Die Kompetenzen der Jugendlichen im dritten Ländervergleich*, Münster: Waxmann.

Prenzel, M., J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost, and U. Schiefele, eds. (2005): *PISA 2003: Der zweite Vergleich der Länder in Deutschland: Was wissen und können Jugendliche?*, Münster: Waxmann.

Sacerdote, B. (2011): "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" in *Handbook of the Economics of Education*, ed. by E. Hanushek, S. Machin, and L. Woessmann, Elsevier, vol. 3, 249–277.

Schütz, G., H. Ursprung, and L. Woessmann (2008): "Education Policy and Equality of Opportunity," *Kyklos*, 61, 279–308.

Vigdor, J. and T. Nechyba (2007): "Peer Effects in North Carolina Public Schools," in *Schools and the Equal Opportunity Problem*, ed. by L. Woessmann and P. Peterson, Cambridge, Mass.: MIT Press.

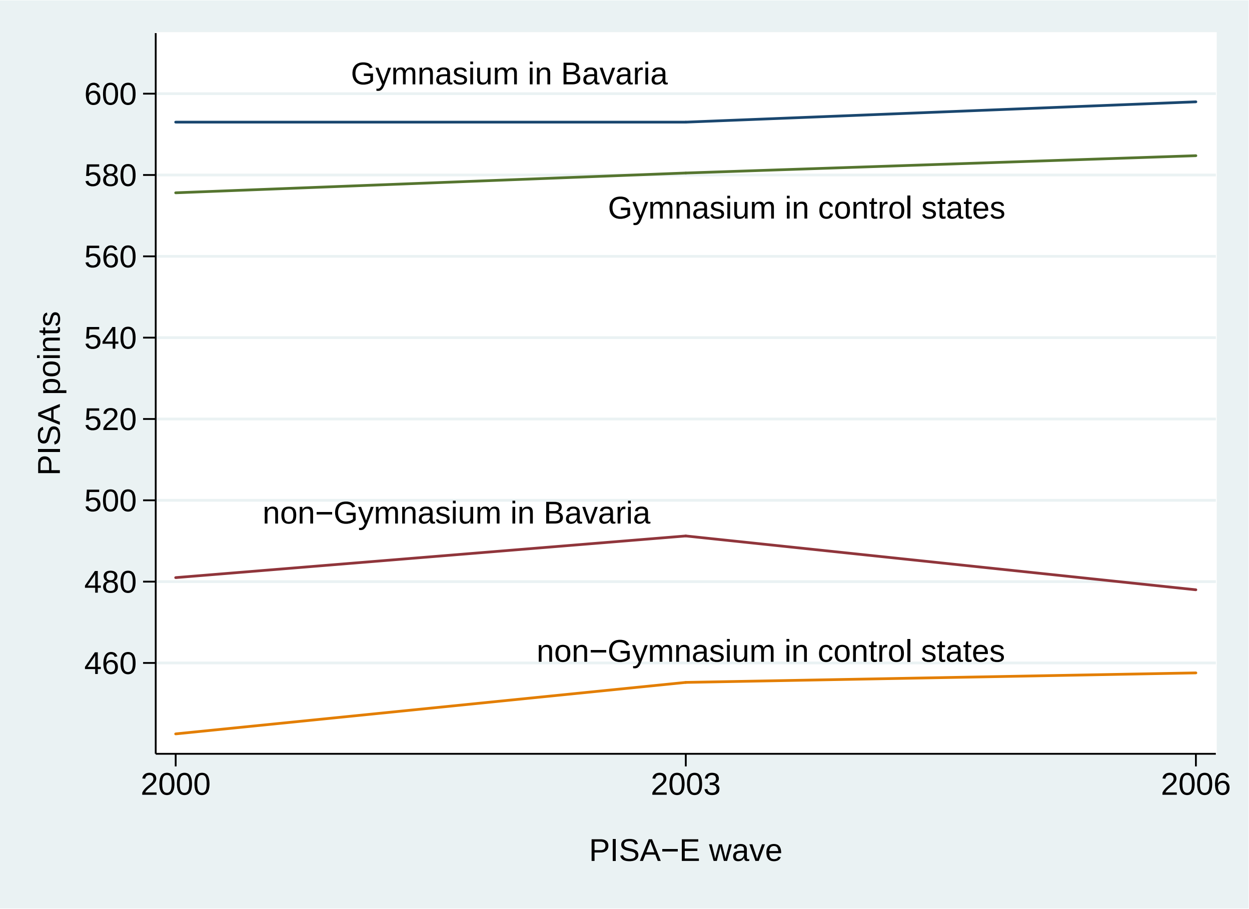
Woessmann, L. (2009): "International Evidence on School Tracking: A Review," *CESifo DICE Report*, 7, 26–34.

——— (2010): "Institutional Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries," *Jahrbücher für Nationalökonomie und Statistik*, 230, 234–270.

Woessmann, L., E. Lüdemann, G. Schütz, and M. West (2009): *School Accountability, Autonomy, and Choice around the World*, Cheltenham: Edward Elgar.

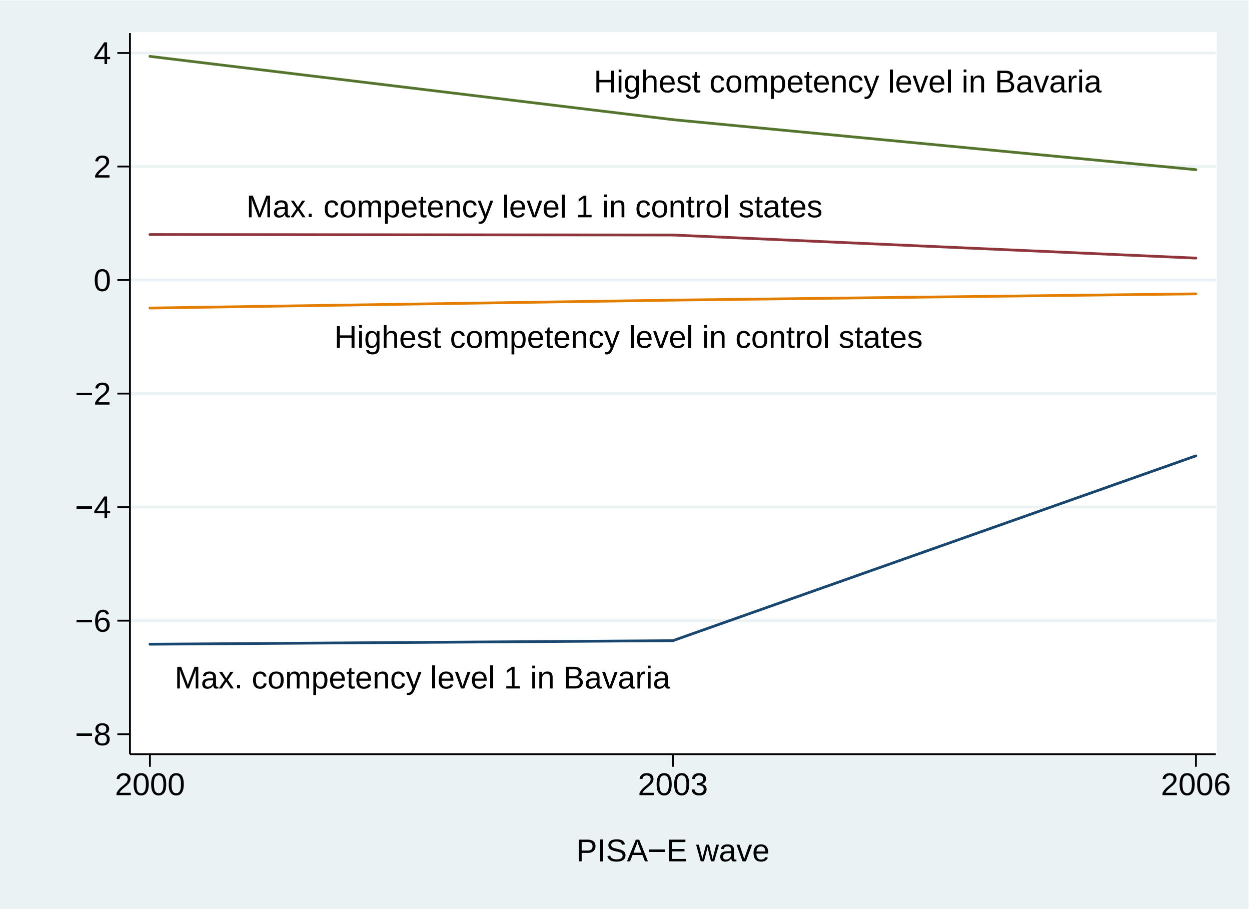
Figures and Tables

Figure 1
Reading Performance in Non-Gymnasium and Gymnasium Schools in Bavaria and Control States



Notes: mean reading performance of PISA-E waves 2000, 2003, and 2006 plotted. Non-Gymnasium mean reading performance is computed on the basis of the mean reading performance of all 15-year-old students in the state (including students in special education and vocational schools), the mean reading performance in Gymnasium, and the share of Gymnasium students in the state; the latter two measures refer to ninth graders in PISA-E 2000 and to 15-year-old students in PISA-E 2003 and 2006. See main text for definition of control states.
Data: PISA-E 2000, 2003, and 2006.

Figure 2
Share of Low Performers and Top Performers in Bavaria and Control States



Notes: residuals from regressions of performance measures on survey year dummies plotted. Performance measures are the share of 15-year-old students achieving at most competency level 1 and the share of students achieving the highest competency level within a state. PISA-E performance is pooled across the three subjects math, reading, and science. See main text for definition of control states.
Data: PISA-E 2000, 2003, and 2006.

Table 1
Descriptive Statistics of Outcomes at State Level

	Bavaria			Control states		
	2000	2003	2006	2000	2003	2006
Variable	(1)	(2)	(3)	(4)	(5)	(6)
PISA-E performance (across all subjects)						
Mean	511.3	527.0	522.0	487.2	497.2	504.8
Share achieving competency level 1 or below	17.5	13.9	14.3	24.7	21.0	17.8
Share achieving highest competency level	13.0	9.0	11.0	8.5	5.9	8.8
Percentiles						
p5	330.7	341.7	341.7	302.9	316.3	334.5
p10	373.3	392.0	386.0	347.9	360.1	374.6
p90	639.0	646.7	644.7	617.8	626.1	632.7
p95	667.3	674.0	672.7	649.0	655.0	662.2
PISA-E performance (reading)						
Mean	510.0	518.0	511.0	482.1	489.5	495.2
Share achieving competency level 1 or below	14.5	14.1	15.6	23.2	21.0	19.2
Share achieving highest competency level	12.2	12.5	12.0	7.8	8.3	9.7
Percentiles						
p5	322.0	322.0	314.0	283.5	299.9	304.4
p10	373.0	382.0	369.0	336.8	346.8	356.6
p90	635.0	635.0	634.0	613.5	616.4	624.5
p95	661.0	662.0	661.0	642.5	644.5	652.3
Share of 15-year-olds affected by new tracking regime in Bavaria (*)						
Basic school students	9.2	26.2	77.4			
Middle school students	8.5	25.4	74.7			
Share of 15-year-olds attending ...						
Gymnasium	26.6	26.3	27.5	27.7	28.3	30.5
Special education schools	3.6	2.6	2.5	4.4	3.6	3.6
Vocational schools	14.1	11.0	9.4	4.3	5.3	5.3

Notes: all variables are measured at the state level. Variables for the control states are unweighted averages across all control states; for definition of control states, see main text. All performance measures refer to 15-year-old students. *PISA-E performance (across all subjects)* is the simple average of the indicated performance measure across the three subjects math, reading, and science.

Data: PISA-E 2000, 2003, and 2006; (*) Bavarian Ministry of Education (2008); own calculations.

Table 2
Descriptive Statistics of Outcomes at School-Type Level

	Bavaria				Control states			
	2003		2006		2003		2006	
	non-Gym	Gym	non-Gym	Gym	non-Gym	Gym	non-Gym	Gym
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PISA-E performance (across all subjects)								
Mean	502.7	605.7	496.3	608.3	470.8	588.9	477.4	594.1
Standard deviation	90.5	62.9	90.3	62.0	86.5	65.0	83.1	65.3
Share achieving competency level 1 or below	17.2	0.5	18.0	0.0	26.3	0.8	21.9	0.4
Share achieving competency level 4 or higher	31.9	80.8	28.4	81.7	18.2	72.4	18.7	74.2
Share of 15-year-olds attending school track	59.4	26.3	60.0	27.5	56.7	28.3	54.3	30.5
Share of migrants	24.6	13.2	23.8	12.2	19.2	12.5	17.9	9.9
PISA-E participation rate	90.7	96.0	89.5	95.0	92.7	95.6	94.0	95.9
Average class size in grades 5-6 (*)	25.9	28.4	26.0	28.5	25.0	27.5	25.1	27.4
Average class size in grades 7-9/10 (*)	25.1	27.0	25.0	27.3	23.5	25.5	23.2	25.6
Share of 15-year-olds in 7th grade	1.2	0.7	1.8	0.2	2.8	0.1	2.2	0.1
Share of 15-year-olds in 8th grade	16.4	9.0	18.5	6.7	19.5	5.7	16.1	3.1
Share of 15-year-olds in 9th grade	68.3	59.6	61.7	54.2	57.4	61.3	57.6	54.8
Share of 15-year-olds in 10th grade	14.1	30.6	18.1	38.6	20.1	32.6	24.1	41.4
Average grade attended	8.95	9.19	8.97	9.32	8.94	9.27	9.04	9.39
Share of students below 9th grade	17.6	9.7	20.3	6.9	22.4	5.8	18.3	3.1

Notes: all variables are measured at the school-track x state level. Variables for the control states are unweighted averages across all control states; see main text for definition of control states. *Non-Gym* variables are aggregated across all non-Gymnasium school types within each state (excluding Gymnasium and comprehensive schools), with each non-Gymnasium school track being weighted by the share of 15-year-old students attending the respective school track; *Gym* refers to outcomes for the Gymnasium track in each state. All performance measures refer to 15-year-old students, averaged across the PISA-E performance in the subjects math, reading, and science. *Average class size in grades 5-6* refers to two years prior to the PISA-E school year, while *average class size in grades 7-9/10* refers to the PISA-E school year. A negligible share of PISA-E students attended grades 6 and 11, respectively.
Data: PISA-E publications 2003 and 2006; (*) Kultusministerkonferenz (2010).

Table 3
Reform Effects with Performance at State Level

			Share max. level 1	Mean reading performance
Dependent variable:	perc 5 – perc 95	perc 10 – perc 90	– share highest level	non-Gym – Gym
	(1)	(2)	(3)	(4)
Reform	–11.19**	–8.40**	5.32***	–15.09***
	(3.61)	(2.96)	(0.88)	(2.57)
Bavaria	7.94**	7.81*	–10.99***	22.28***
	(3.08)	(3.86)	(1.52)	(3.09)
Year 2003	6.96	4.74	–0.84	8.06**
	(4.77)	(4.41)	(1.26)	(2.71)
Year 2006	18.17***	12.27**	–7.14***	6.01
	(4.69)	(4.32)	(1.42)	(3.31)
Constant	–345.92***	–270.35***	16.08***	–133.20***
	(4.50)	(5.15)	(1.73)	(4.13)
Adj. R-squared	0.104	0.062	0.279	0.311
Observations	81	81	81	27

Notes: Dependent variables: 5th performance percentile minus 95th percentile (Column 1); 10th percentile minus 90th percentile (Column 2); share of students achieving at most competency level 1 minus share of students achieving the highest competency level (Column 3); all performance measures in Columns (1)-(3) are measured at the state level and pooled across the subjects math, reading, and science. The dependent variable in Column (4) is mean reading performance in non-Gymnasium tracks minus mean reading performance in the Gymnasium track. In all PISA-E waves, non-Gymnasium mean reading performance is computed on the basis of the mean reading performance of all 15-year-old students in the state (including students in special education and vocational schools), the mean reading performance in Gymnasium, and the share of Gymnasium students in the state; the latter two measures refer to ninth graders in PISA-E 2000 and to 15-year-old students in PISA-E 2003 and 2006. Coefficients from ordinary least squares regressions with standard errors (in parentheses) clustered at the state level. Samples include Bavaria and all control states (see main text) with data from PISA-E waves 2000, 2003, and 2006. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 4
Robustness Check: Reform Effects with Performance at State Level

			Share max. level 1	Mean reading performance
Dependent variable:	perc 5 – perc 95	perc 10 – perc 90	– share highest level	non-Gym – Gym
	(1)	(2)	(3)	(4)
Reform	–11.11***	–7.53**	4.11***	–13.64***
	(3.06)	(2.79)	(1.06)	(3.65)
Bavaria	9.85*	11.89**	–11.09***	20.05***
	(4.52)	(4.85)	(1.52)	(3.43)
Year 2003	9.21*	8.31*	1.15	2.73
	(4.17)	(3.64)	(1.64)	(3.44)
Year 2006	21.45***	16.45**	–2.41	–4.11
	(4.27)	(5.34)	(2.68)	(6.84)
Share of 15-year-olds in Gymnasium	–0.53	–0.37	–1.28**	2.30*
	(1.46)	(1.68)	(0.50)	(1.10)
Share of 15-year-olds in special schools	2.81	4.78	1.96	–5.99*
	(2.68)	(2.67)	(1.73)	(3.06)
Constant	–343.76***	–281.22***	43.01***	–170.45***
	(37.63)	(41.71)	(7.46)	(20.87)
Adj. R-squared	0.090	0.092	0.340	0.437
Observations	81	81	81	27

Notes: Dependent variables: 5th performance percentile minus 95th percentile (Column 1); 10th percentile minus 90th percentile (Column 2); share of students achieving at most competency level 1 minus share of students achieving the highest competency level (Column 3); all performance measures in Columns (1)-(3) are measured at the state level and pooled across the subjects math, reading, and science. The dependent variable in Column (4) is mean reading performance in non-Gymnasium tracks minus mean reading performance in the Gymnasium track. In all PISA-E waves, non-Gymnasium mean reading performance is computed on the basis of the mean reading performance of all 15-year-old students in the state (including students in special education and vocational schools), the mean reading performance in Gymnasium, and the share of Gymnasium students in the state; the latter two measures refer to ninth graders in PISA-E 2000 and to 15-year-old students in PISA-E 2003 and 2006. Coefficients from ordinary least squares regressions with standard errors (in parentheses) clustered at the state level. Samples include Bavaria and all control states (see main text) with data from PISA-E waves 2000, 2003, and 2006. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 5
Reform Effects in Triple-Differences Model with Performance at School-Type Level

Dependent variable:	Mean	SD	Share max. level 1	Share min. level 4
	(1)	(2)	(3)	(4)
Reform	–11.45***	5.35***	5.57***	–3.43***
	(2.52)	(1.17)	(1.07)	(0.99)
Bavaria	16.79***	–2.06*	–0.31**	8.38***
	(2.47)	(0.94)	(0.10)	(1.29)
Year 2006	5.21***	0.33	–0.38**	1.79***
	(1.29)	(0.49)	(0.12)	(0.49)
Bavaria 2006	–2.54*	–1.23**	–0.09	–0.92*
	(1.29)	(0.49)	(0.12)	(0.49)
Bavaria non-Gymnasium	35.23***	–1.91	–17.72***	9.93***
	(7.66)	(3.06)	(3.04)	(2.65)
Non-Gymnasium 2006	–0.86	–4.01**	–3.02**	–1.85*
	(2.42)	(1.21)	(1.03)	(0.93)
School type dummies	Yes	Yes	Yes	Yes
Adj. R-squared	0.941	0.535	0.898	0.953
Observations	144	144	144	144

Notes: Dependent variables: mean performance (Column 1), standard deviation (Column 2), share of students achieving at most competency level 1 (Column 3), and share of students achieving competency level 4 or higher (Column 4) within each school track and state. All dependent variables are pooled across the subjects math, reading, and science. *Reform* equals 1 for basic and middle school track in Bavaria for PISA-E wave 2006 and 0 otherwise. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Non-Gymnasium school tracks are weighted by the share of 15-year-old students attending the respective school track relative to all non-Gymnasium school tracks; the sum of all non-Gymnasium weights and the Gymnasium weight equals 1 in each state. Samples contain all school tracks, excluding comprehensive schools, in Bavaria and the control states (see main text). PISA-E waves 2003 and 2006 included. Significance levels: * p<0.10, ** p<0.05,*** p<0.01.

Table 6
Pre-Reform Trend

Dependent variable:	perc 5 – perc 95	perc 10 – perc 90	Share max. level 1 – share highest level	Mean reading performance non-Gym – Gym
	(1)	(2)	(3)	(4)
Bavaria 2003	–2.96	7.04	1.33	2.46
	(5.38)	(4.89)	(1.41)	(3.07)
Bavaria	9.42*	4.29	–11.65***	21.05***
	(4.74)	(5.35)	(1.78)	(4.33)
Year 2003	7.29	3.96	–0.99	7.78**
	(5.38)	(4.89)	(1.41)	(3.07)
Constant	–346.08***	–269.96***	16.15***	–133.06***
	(4.74)	(5.35)	(1.78)	(4.33)
Adj. R-squared	0.016	0.003	0.234	0.403
Observations	54	54	54	18

Notes: Dependent variables: 5th performance percentile minus 95th percentile (Column 1); 10th percentile minus 90th percentile (Column 2); share of students achieving at most competency level 1 minus share of students achieving the highest competency level (Column 3); all performance measures in Columns (1)-(3) are measured at the state level and pooled across the subjects math, reading, and science. The dependent variable in Column (4) is mean reading performance in non-Gymnasium tracks minus mean reading performance in the Gymnasium track. In all PISA-E waves, non-Gymnasium mean reading performance is computed on the basis of the mean reading performance of all 15-year-old students in the state (including students in special education and vocational schools), the mean reading performance in Gymnasium, and the share of Gymnasium students in the state; the latter two measures refer to ninth graders in PISA-E 2000 and to 15-year-old students in PISA-E 2003 and 2006. Coefficients from ordinary least squares regressions with standard errors (in parentheses) clustered at the state level. Samples include Bavaria and all control states (see main text) with data from PISA-E waves 2000 and 2003. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 7
Robustness Check: Pre-Reform Trend

			Share max. level 1	Mean reading performance
Dependent variable:	perc 5 – perc 95	perc 10 – perc 90	– share highest level	non-Gym – Gym
	(1)	(2)	(3)	(4)
Bavaria 2003	–3.08	7.81	0.98	2.63
	(4.61)	(4.42)	(1.43)	(2.99)
Bavaria	11.12**	7.85	–11.50***	19.36***
	(4.58)	(4.65)	(1.60)	(4.40)
Year 2003	11.68**	9.35**	0.64	3.27
	(4.19)	(3.56)	(1.64)	(4.16)
Share of 15-year-olds in Gymnasium	–1.97	–1.62	–1.00	2.06
	(1.92)	(2.14)	(0.62)	(1.43)
Share of 15-year-olds in special schools	4.86	6.65**	1.59	–4.97
	(2.67)	(2.66)	(1.80)	(3.31)
Constant	–312.94***	–254.41***	36.92**	–168.26***
	(48.55)	(54.15)	(11.07)	(26.67)
Adj. R-squared	0.037	0.078	0.271	0.476
Observations	54	54	54	18

Notes: Dependent variables: 5th performance percentile minus 95th percentile (Column 1); 10th percentile minus 90th percentile (Column 2); share of students achieving at most competency level 1 minus share of students achieving the highest competency level (Column 3); all performance measures in Columns (1)-(3) are measured at the state level and pooled across the subjects math, reading, and science. The dependent variable in Column (4) is mean reading performance in non-Gymnasium tracks minus mean reading performance in the Gymnasium track. In all PISA-E waves, non-Gymnasium mean reading performance is computed on the basis of the mean reading performance of all 15-year-old students in the state (including students in special education and vocational schools), the mean reading performance in Gymnasium, and the share of Gymnasium students in the state; the latter two measures refer to ninth graders in PISA-E 2000 and to 15-year-old students in PISA-E 2003 and 2006. Coefficients from ordinary least squares regressions with standard errors (in parentheses) clustered at the state level. Samples include Bavaria and all control states (see main text) with data from PISA-E waves 2000 and 2003. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 8
Triple-Differences Results Robust to Including Controls at School-Type Level

Dependent variable:	Mean	SD	Share max. level 1	Share min. level 4
	(1)	(2)	(3)	(4)
Reform	–12.65***	4.99**	5.70***	–4.19**
	(3.19)	(1.63)	(1.18)	(1.38)
Bavaria	17.18***	–1.01	–1.23	7.54***
	(3.86)	(1.13)	(1.44)	(1.49)
Year 2006	4.62*	0.13	–0.50	1.29
	(2.45)	(1.01)	(0.75)	(1.01)
Bavaria 2006	–2.72	–0.73	–0.21	–1.28
	(1.82)	(0.95)	(0.53)	(0.85)
Bavaria non-Gymnasium	38.52***	–1.60	–19.36***	10.60***
	(7.56)	(2.04)	(2.94)	(2.51)
Non-Gymnasium 2006	0.87	–3.42*	–3.10*	–0.65
	(4.43)	(1.77)	(1.64)	(1.74)
Share of 15-year-olds in track	–0.04	0.38*	–0.07	–0.10
	(0.66)	(0.18)	(0.25)	(0.23)
Share of migrants in school track	–0.11	0.24	–0.15	–0.23
	(0.62)	(0.18)	(0.20)	(0.24)
PISA-E participation rate	–0.44	0.37**	–0.08	–0.45*
	(0.71)	(0.12)	(0.29)	(0.24)
Average class size in grades 5-6	–3.95***	0.35	1.18**	–1.48***
	(1.12)	(0.21)	(0.42)	(0.38)
Average class size in grades 7-9/10	2.17	–0.61	–0.09	1.50**
	(1.72)	(0.42)	(0.78)	(0.52)
School type dummies	Yes	Yes	Yes	Yes
Adj. R-squared	0.947	0.622	0.907	0.960
Observations	144	144	144	144

Notes: Dependent variables: mean performance (Column 1), standard deviation (Column 2), share of students achieving at most competency level 1 (Column 3), and share of students achieving competency level 4 or higher (Column 4) within each school track and state. All dependent variables are pooled across the subjects math, reading, and science. *Reform* equals 1 for basic and middle school track in Bavaria for PISA-E wave 2006 and 0 otherwise. *Share of 15-year olds in track* is the share of all 15-year-old students in a state attending the respective school track. The other control variables are measured at the school-type level within each state. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Non-Gymnasium school tracks are weighted by the share of 15-year-old students attending the respective school track relative to all non-Gymnasium school tracks; the sum of all non-Gymnasium weights and the Gymnasium weight equals 1 in each state. Samples contain all school tracks, excluding comprehensive schools, in Bavaria and the control states (see main text). PISA-E waves 2003 and 2006 included. Significance levels: * p<0.10, ** p<0.05,*** p<0.01.

Table 9
Triple-Differences Results Robust to Aggregating Performance across Non-Gymnasium School Types

Dependent variable:	Mean	SD	Share max. level 1	Share min. level 4
	(1)	(2)	(3)	(4)
Reform	–14.01**	4.20	6.19**	–4.65**
	(4.75)	(2.30)	(2.01)	(1.56)
Bavaria	15.17***	–1.36	0.32	7.88***
	(2.78)	(1.56)	(0.87)	(1.31)
Non-Gymnasium schools	–129.72***	15.51**	28.70***	–59.80***
	(14.76)	(6.27)	(6.52)	(4.64)
Year 2006	1.60	1.01	1.04	0.74
	(2.45)	(1.30)	(0.94)	(0.92)
Bavaria 2006	–0.97	–1.86*	–1.01*	–0.81
	(1.46)	(0.84)	(0.49)	(0.59)
Bavaria non-Gymnasium	15.48**	3.78*	–9.51***	4.55**
	(5.92)	(1.70)	(2.71)	(1.67)
Non-Gymnasium 2006	6.80	–3.58	–5.58*	0.92
	(5.95)	(2.68)	(2.54)	(1.85)
Share of 15-year-olds in track	0.67	0.06	–0.26	0.23
	(0.61)	(0.22)	(0.27)	(0.18)
Share of migrants in school track	–0.67	0.29**	0.26	–0.17
	(0.49)	(0.12)	(0.18)	(0.16)
PISA-E participation rate	0.21	–0.27	–0.34	–0.20
	(0.59)	(0.20)	(0.24)	(0.19)
Average class size in grades 5-6	–1.50	–0.48	0.18	–0.75
	(1.31)	(0.28)	(0.49)	(0.44)
Average class size in grades 7-9/10	3.09	–0.17	–0.90	1.20
	(2.62)	(0.79)	(1.09)	(0.80)
Adj. R-squared	0.959	0.807	0.907	0.973
Observations	108	108	108	108

Notes: Dependent variables: aggregated performance across all non-Gymnasium school tracks within each state and performance for Gymnasium track pooled; mean (Column 1); standard deviation (Column 2); share of students achieving at most competency level 1 (Column 3); share of students achieving competency level 4 or higher (Column 4). The aggregated outcome across the non-Gymnasium school tracks is computed by weighting each non-Gymnasium school track by the share of 15-year-old students attending the respective school track relative to all non-Gymnasium school tracks. All dependent variables are pooled across the subjects math, reading, and science. *Reform* equals 1 for basic and middle school track in Bavaria for PISA-E wave 2006 and 0 otherwise. *Share of 15 years old in track* is the share of all 15-year-old students in a state attending the respective school track. The other control variables are measured at the school type level within each state. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Non-Gymnasium school tracks are weighted by the share of 15-year-old students attending the respective school track relative to all non-Gymnasium school tracks; the sum of all non-Gymnasium weights and the Gymnasium weight equals 1 in each state. Samples contain all school tracks, excluding comprehensive schools, in Bavaria and control states (see main text). PISA-E waves 2003 and 2006 included. Significance levels: * p<0.10, ** p<0.05,*** p<0.01.

Table 10
Grade Repetition as Possible Channel for Reform Effects

Dependent variable:	Mean	SD	Share max. level 1	Share min. level 4
	(1)	(2)	(3)	(4)
Reform	–6.34	4.79**	1.93	–3.97
	(4.81)	(1.69)	(1.70)	(2.38)
Bavaria	23.27***	–2.00	–3.73***	9.34***
	(3.87)	(1.16)	(0.95)	(1.72)
Year 2006	–10.83**	3.18*	5.04**	–4.36***
	(4.46)	(1.65)	(2.14)	(1.10)
Bavaria 2006	–3.70*	–0.48	0.54	–1.37
	(1.97)	(0.94)	(0.71)	(0.82)
Bavaria non-Gymnasium	20.99**	–0.23	–9.95***	8.46*
	(7.29)	(2.79)	(2.34)	(4.23)
Non-Gymnasium 2006	7.34	–5.17**	–4.94**	2.50
	(4.68)	(1.95)	(2.07)	(1.61)
Share of 15-year-olds in 7th grade	–3.52	0.89	2.05	–0.82
	(3.68)	(0.90)	(1.23)	(1.49)
Share of 15-year-olds in 8th grade	–0.36	–0.16	0.19	0.11
	(0.78)	(0.24)	(0.26)	(0.39)
Share of 15-year-olds in 10h grade	1.66**	–0.40	–0.56*	0.69**
	(0.69)	(0.23)	(0.29)	(0.25)
Share of 15-year-olds in track	–0.39	0.37**	0.09	–0.13
	(0.35)	(0.12)	(0.13)	(0.10)
Share of migrants in school track	–0.32	0.21	0.01	–0.17
	(0.49)	(0.19)	(0.10)	(0.24)
PISA-E participation rate	–1.16**	0.46***	0.26	–0.60**
	(0.44)	(0.11)	(0.19)	(0.24)
Average class size in grades 5-6	–4.09***	0.43	1.06**	–1.71***
	(1.04)	(0.26)	(0.45)	(0.39)
Average class size in grades 7-9/10	2.33	–0.53	–0.21	1.41**
	(1.82)	(0.51)	(0.71)	(0.57)
School type dummies	Yes	Yes	Yes	Yes
Adj. R-squared	0.959	0.638	0.941	0.966
Observations	144	144	144	144

Notes: Dependent variables: mean performance (Column 1), standard deviation (Column 2), share of students achieving at most competency level 1 (Column 3), and share of students achieving competency level 4 or higher (Column 4) within each school track and state. All dependent variables are pooled across the subjects math, reading, and science. *Reform* equals 1 for basic and middle school track in Bavaria for PISA-E wave 2006 and 0 otherwise. *Share of 15-year olds in track* is the share of all 15-year-old students in a state attending the respective school track. The other control variables are measured at the school-type level within each state. The ninth grade level, the omitted category, is the most frequently attended grade in each school track. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Non-Gymnasium school tracks are weighted by the share of 15-year-old students attending the respective school track relative to all non-Gymnasium school tracks; the sum of all non-Gymnasium weights and the Gymnasium weight equals 1 in each state. Samples contain all school tracks, excluding comprehensive schools, in Bavaria and the control states (see main text). PISA-E waves 2003 and 2006 included. Significance levels: * p<0.10, ** p<0.05,*** p<0.01.

Table 11
Reform Effects on Share of Students Repeating a Grade

Dependent variable:	Average grade attended			Share of students below 9th grade		
		Excluding Bavarian...			Excluding Bavarian...	
Sample:	All	middle school	basic school	All	middle school	basic school
	(1)	(2)	(3)	(4)	(5)	(6)
Reform	–0.09***	–0.09***	–0.06**	7.04***	9.03***	3.33**
	(0.03)	(0.02)	(0.02)	(1.14)	(1.15)	(1.15)
Bavaria	–0.08***	–0.08***	–0.08***	3.85**	3.85**	3.85**
	(0.02)	(0.02)	(0.02)	(1.38)	(1.39)	(1.39)
Year 2006	0.12***	0.12***	0.12***	–2.72**	–2.72**	–2.72**
	(0.02)	(0.02)	(0.02)	(1.09)	(1.10)	(1.10)
Bavaria 2006	0.00	0.00	0.00	–0.07	–0.07	–0.07
	(0.02)	(0.02)	(0.02)	(1.09)	(1.10)	(1.10)
Bavaria non-Gymnasium	0.20***	0.34***	0.06*	–13.59***	–25.94***	–0.38
	(0.03)	(0.05)	(0.03)	(2.21)	(3.71)	(0.67)
Non-Gymnasium 2006	–0.04	–0.04	–0.04	–1.02	–0.93	–0.93
	(0.02)	(0.02)	(0.02)	(1.16)	(1.15)	(1.15)
School type dummies	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R-squared	0.865	0.887	0.878	0.750	0.815	0.817
Observations	48	46	46	48	46	46

Notes: Dependent variables: average grade attended by 15-year-old students within each school track in each state (Columns 1 to 3); share of students below ninth grade within each school track in each state (Columns 4 to 7). *Reform* equals 1 for basic and middle school tracks in Bavaria in 2006; 0 otherwise. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Non-Gymnasium school tracks are weighted by the share of students in the respective school track relative to all non-Gymnasium school tracks (sum of all non-Gymnasium weights equals 1 in each state); in each state, the Gymnasium track receives weight 1. Samples contain all school tracks, excluding comprehensive schools, in Bavaria and the control states (see main text). PISA-E waves 2003 and 2006 included. Significance levels: * p<0.10, ** p<0.05,*** p<0.01.

Table 12
Reform Effects for Basic School Track

Dependent variable:	Mean	SD	Share max. level 1	Share min. level 4
	(1)	(2)	(3)	(4)
Reform	–11.19***	4.93***	6.35***	–1.46
	(2.42)	(1.22)	(1.02)	(0.94)
Bavaria	16.79***	–2.06*	–0.31**	8.38***
	(2.47)	(0.94)	(0.10)	(1.30)
Year 2006	5.21***	0.33	–0.38**	1.79***
	(1.30)	(0.49)	(0.12)	(0.49)
Bavaria 2006	–2.54*	–1.23**	–0.09	–0.92*
	(1.30)	(0.49)	(0.12)	(0.49)
Bavaria non-Gymnasium	29.25***	4.72*	–22.96***	–0.50
	(5.40)	(2.20)	(3.19)	(1.37)
Non-Gymnasium 2006	–0.81	–4.07**	–2.98**	–1.77*
	(2.42)	(1.22)	(1.02)	(0.94)
School type dummies	Yes	Yes	Yes	Yes
Adj. R-squared	0.943	0.571	0.904	0.960
Observations	138	138	138	138

Notes: Dependent variables: mean performance (Column 1), standard deviation (Column 2), share of students achieving at most competency level 1 (Column 3), and share of students achieving competency level 4 or higher (Column 4) within each school track and state. All dependent variables are pooled across the subjects math, reading, and science. *Reform* equals 1 for basic and middle school track in Bavaria for PISA-E wave 2006 and 0 otherwise. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Non-Gymnasium school tracks are weighted by the share of 15-year-old students attending the respective school track relative to all non-Gymnasium school tracks; the sum of all non-Gymnasium weights and the Gymnasium weight equals 1 in each state. Samples contain all school tracks in Bavaria and the control states (see main text), excluding comprehensive schools and the middle school track in Bavaria. PISA-E waves 2003 and 2006 included. Significance levels: * p<0.10, ** p<0.05,*** p<0.01.

Table 13
Reform Effects for Middle School Track

Dependent variable:	Mean	SD	Share max. level 1	Share min. level 4
	(1)	(2)	(3)	(4)
Reform	–12.52***	6.67***	4.05***	–6.86***
	(2.42)	(1.22)	(1.02)	(0.94)
Bavaria	16.79***	–2.06*	–0.31**	8.38***
	(2.47)	(0.94)	(0.10)	(1.30)
Year 2006	5.21***	0.33	–0.38**	1.79***
	(1.30)	(0.49)	(0.12)	(0.49)
Bavaria 2006	–2.54*	–1.23**	–0.09	–0.92*
	(1.30)	(0.49)	(0.12)	(0.49)
Bavaria non-Gymnasium	41.58**	–8.95*	–12.12**	21.10***
	(13.60)	(4.78)	(5.01)	(4.47)
Non-Gymnasium 2006	–0.81	–4.07**	–2.98**	–1.77*
	(2.42)	(1.22)	(1.02)	(0.94)
School type dummies	Yes	Yes	Yes	Yes
Adj. R-squared	0.938	0.554	0.900	0.956
Observations	138	138	138	138

Notes: Dependent variables: mean performance (Column 1), standard deviation (Column 2), share of students achieving at most competency level 1 (Column 3), and share of students achieving competency level 4 or higher (Column 4) within each school track and state. All dependent variables are pooled across the subjects math, reading, and science. *Reform* equals 1 for basic and middle school track in Bavaria for PISA-E wave 2006 and 0 otherwise. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Non-Gymnasium school tracks are weighted by the share of 15-year-old students attending the respective school track relative to all non-Gymnasium school tracks; the sum of all non-Gymnasium weights and the Gymnasium weight equals 1 in each state. Samples contain all school tracks in Bavaria and the control states (see main text), excluding comprehensive schools and the basic school track in Bavaria. PISA-E waves 2003 and 2006 included. Significance levels: * p<0.10, ** p<0.05,*** p<0.01.

Table 14
Persistence of Reform Effects on Reading Performance

Included survey years:	2000 + 2003 + 2006 + 2009		2003 + 2006 + 2009	
	(1)	(2)	(3)	(4)
Reform	–13.56***	–15.09***	–14.80***	–16.32***
	(2.86)	(2.64)	(2.45)	(2.55)
Reform 2009		3.06		3.06
		(5.40)		(5.44)
Bavaria	16.76***	16.76***	12.50***	12.50***
	(4.39)	(4.43)	(2.46)	(2.49)
Non-Gymnasium	–133.20***	–133.20***	–125.28***	–125.28***
	(4.21)	(4.24)	(2.38)	(2.41)
Non-Gymnasium x Bavaria	22.28***	22.28***	23.52***	23.52***
	(3.14)	(3.17)	(2.38)	(2.41)
Year 2003	4.74	4.74		
	(2.80)	(2.82)		
Bavaria 2003	–3.64	–3.64		
	(3.27)	(3.30)		
Non-Gymnasium 2003	8.06**	8.06**		
	(2.76)	(2.78)		
Year 2006	9.14**	9.06**	4.33*	4.25*
	(3.27)	(3.28)	(1.97)	(1.97)
Bavaria 2006	–4.27	–3.51	–0.01	0.75
	(3.79)	(3.48)	(2.43)	(1.97)
Non-Gymnasium 2006	5.84	6.01	–2.08	–1.91
	(3.31)	(3.40)	(2.37)	(2.55)
Year 2009	–5.65	–5.57	–10.46**	–10.37**
	(5.01)	(5.13)	(3.25)	(3.40)
Bavaria 2009	–6.12	–6.88	–1.86	–2.63
	(4.67)	(5.27)	(2.50)	(3.40)
Non-Gymnasium 2009	21.26***	21.09***	13.34**	13.17**
	(5.37)	(5.64)	(4.23)	(4.54)
Adj. R-squared	0.975	0.974	0.982	0.981
Observations	72	72	54	54

Notes: Dependent variable: mean reading performance in non-Gymnasium tracks and Gymnasium track within each state pooled. In all PISA-E waves, non-Gymnasium mean reading performance is computed on the basis of the mean reading performance of all 15-year-old students in the state (including students in special education and vocational schools), the mean reading performance in Gymnasium, and the share of Gymnasium students in the state; the latter two measures refer to ninth graders in PISA-E 2000 and to 15-year-old students in PISA-E 2003 and 2006. Non-Gymnasium mean reading performance for the 2009 survey is based on the mean reading performance of ninth graders in the state (without students in special education and vocational schools), the mean reading performance of ninth graders in Gymnasium, and the share of ninth grade Gymnasium students in the state. *Reform* equals 1 for non-Gymnasium outcome in Bavaria in 2006 and 2009; 0 otherwise. *Reform 2009* equals 1 for non-Gymnasium outcome in Bavaria in 2009; 0 otherwise. Ordinary least squares regressions with robust standard errors clustered at the state level in parentheses. Samples contain Bavaria and the control states (see main text) for the PISA-E waves 2000, 2003, and 2006. 2009 performance is reading performance from the "Educational Standards" (*Bildungsstandards*) survey conducted on behalf of the Secretariat of the Standing Conference of the Ministers of Education (*Kultusministerkonferenz*). Significance levels: * p<0.10, ** p<0.05,*** p<0.01.