TRACKING IN THE EDUCATIONAL SYSTEM AND INEQUALITIES

Vom Fachbereich Sozialwissenschaften
der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Philosophie (Dr. phil.)
genehmigte

Dissertation

vorgelegt von
Phil Kolbe

Tag der Disputation:	Kaiserslautern, 15.12.2022
Dekan:	Prof. Dr. Michael Fröhlich
Vorsitzende/r:	Juniorprof. Dr. Daniela Czernochowski
Gutachter/in:	1. Prof. Dr. Henning Best 2. Prof. Dr. Steffen Schindler

D 386

Februar 2023

Acknowledgements

Working and writing on a dissertation is a lengthy process and hard work, which can be a lot of fun, but also takes a lot of nerves. However, it can hardly be done without the help of others. Therefore, I would like to take this opportunity to thank these people, because without you this work would either not have been possible or the process would have taken much longer.

The first thank goes to my advisor Henning Best. Thank you for always showing me solutions when I had problems, critically questioning my ideas, and re-adjusting my work focus every once in a while. I really learned a lot from our conversations and I am very grateful for the experiences I was able to have while working here.

I would also like to thank colleagues and friends who have always encouraged me, discussed problems with me and helped me (even sometimes at very short notice) with various things in the work process. I would like to express a special thanks to Nico Seifert, Ingmar Rapp, as well as Tobias Rüttenauer. You supported me at very different stages of the work, were always available, helpful and patient. I enjoyed working with you very much. Many thanks also to Steffen Schindler for taking the time and effort to read and grade this work. I would also like to thank Marius Cziriak for the discussions, the support and the walks along the Neckar. I would also like to thank Volker Ludwig and Tanja Dannwolf, as well as Felix Bader, Ingmar Ehler, Charlotte Haußmann and Sam Handel. For proofreading and improving the language of this dissertation, I would like to thank John Cirilli, Khumo Kluge, and Wolfgang Amann.

I would like to express my gratitude to my partner Vanessa for her unwavering encouragement, for enduring some moody phases, but also for appreciating each other and for offering me peace and relaxation when I needed it. Finally yet importantly, I would like to thank my parents Irene and Raimut. You have supported me unconditionally from the very beginning, made the path to my studies possible, and stood by me through every obstacle.

Thank you all very much!

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aims and scope

The association between social origin and educational attainment has been repeatedly confirmed and studied in social science research (Breen & Jonsson 2005; Boudon 1974). This area of research examines various outcomes, such as educational performance (Grätz & Wiborg 2020), educational decision-making (Stocké 2007), and teacher assessment (Helbig & Morar 2017), each as a function of social origin. Results show that lower social origin correlates with lower performance (Grätz & Wiborg 2020), lower transition to more demanding educational paths (Stocké 2007) and lower grades for the same performance (Helbig & Morar 2017). Moreover, research examining the broader consequences of social inequality in education consistently demonstrates the negative effects of this inequality, such as poor individual labor market outcomes (Card 1999; Allmendinger 1989; Grätz & Pollak 2016) or poor health (Leuven et al. 2016; Negri et al. 2021). Much of the international comparative research to date has shown that countries differ in the extent of educational inequality (Hanushek & Wössmann 2011; Breen et al. 2010). This research suggests that the institutional design of the education system can affect multiple dimensions of educational inequality (Cordero et al. 2018; Le Donné 2014; Pfeffer 2008), such as school performance and educational decisions. In addition to international comparative research, other research also suggests that institutional characteristics moderate the link between social origin and educational inequality (Büchler 2016; Below 2002; Klein et al. 2019). Thus, the institutional features of the education system provide opportunities for policy interventions to influence the relationship between social origin and educational inequality and to reduce educational inequalities, which in turn would bring positive externalities (Psacharopoulos & Patrinos 2004).

The literature examines and discusses various institutional characteristics of the education system for their respective effects on or associations to educational inequalities (Van de Werfhorst, Herman G. & Mijs 2010; Le Donné 2014). Studies in this regard examine, among other things, the importance of private schools (Schütz et al. 2008) and class size (Angrist & Lavy 1999; Leuven et al. 2008). In this respect, tracking is also an institutional characteristic that has been studied repeatedly and could be an important link between social origin and education (Betts 2011; Terrin & Triventi 2022; Le Donné 2014). Tracking is the practice of separating students by performance. This separation can occur between schools or within schools. Thus, students are placed in a particular school type (between-school tracking) or class (within-school tracking) based on their performance. National and international research demonstrate the importance of tracking in relation to the emergence of educational inequalities (Hanushek &

Wössmann 2006; Piopiunik 2014; van Elk et al. 2011; Baier et al. 2022). In this context, previous research has often shown that early and strict tracking leads to greater educational inequality (Van de Werfhorst, Herman G. & Mijs 2010; Lavrijsen & Nicaise 2015; Hanushek & Wössmann 2011). However, there is also research that finds no effects from tracking (Van de Werfhorst 2019) or even inequality-reducing effects from early and strict tracking (Esser & Relikowski 2015; Figlio & Page 2002). Against this background, further research on the associations with - and effects of - tracking, including under different settings and contexts, is important for a better understanding of tracking and may be particularly interesting for the German education system. This is because, apart from some deviations, the German education system is characterized by an early and strict separation of students into different school types in secondary education (Le Donné 2014; Henninges et al. 2019).

Over the years, there have been many different educational reforms in Germany with different scopes, goals, and at different phases in the education system (Becker et al. 2017; Büchler 2016; Helbig & Nikolai 2015). The fact that the federal states in Germany can decide independently on education policy (*Kulturhoheit der Länder* - Cultural sovereignty of the states) means that they partially developed in different directions (Helbig & Nikolai 2015). The following contribution is therefore limited to three selected aspects of tracking in the education systems of the federal states in Germany and its influences on features of educational inequality: *integrated comprehensive schools*, *timing of tracking*, and *strictness of tracking*.

The first aspect within the education system that is the focus of this contribution is the relationship between integrated comprehensive schools and *Abitur* attainment. Comprehensive schools were introduced in Germany as early as 1969 to supplement existing traditional school types (Köller 2008). While the number of students in integrated comprehensive school was relatively low, the number of students in integrated comprehensive schools has been increasing since 2005, and since 2016, more students attend comprehensive schools than *Realschule*, the intermediate secondary school type (Autorengruppe Bildungsberichterstattung 2018). Unlike the traditional school types in Germany, integrated comprehensive schools use ability grouping instead of between-school tracking. In this approach, students are divided into ability groups within the school or class for some subjects. One goal of introducing integrated comprehensive schools was to reduce educational inequality by moving tracking to a later point in time (Köller 2008; Leschinsky & Mayer 1999). However, it is unclear to what extent this goal is being achieved today, especially since many research findings are quite old (Tillmann 1988). It is therefore also unclear whether the relationship between social origin and *Abitur* attainment is

lower in integrated comprehensive schools compared to the traditional tracked forms of secondary education. In addition, more and more federal states are expanding or changing the school types in the secondary education (Becker et al. 2017). In the process, school types are being introduced that, in some cases, divide students into ability groups similar to the integrated comprehensive schools. The relationship between integrated comprehensive schools and upper secondary school attainment is examined with data from the National Educational Panel Study (NEPS). NEPS data are particularly well suited for this analysis because integrated comprehensive schools are oversampled in the data set (IEA Data Processing and Research Center 2010).

The second focus of this contribution is the influence of timing of tracking on educational decisions. To determine the effect of timing of tracking on educational decisions, we examine an educational reform in Lower Saxony in 2004. As part of the education reform, Lower Saxony abolished a two-year orientation stage (*Orientierungsstufe*) starting in grade 5 that was independent of school type (Schuchart 2006). The orientation stage separated students into ability groups for certain subjects. Research has shown that the termination of the orientation stage has widened performance gaps (Roller & Steinberg 2020), but it is unclear whether the reform also had an impact on transition behavior to the *Gymnasium* (the academic school track), as theory would expect. To answer this question, data is needed that has a sufficient sample size of students in the federal states and is also collected repeatedly. For this reason, the federal state extensions of the PISA data, PISA-E and IQB-LV are used as the basis for the analysis.

The final focus is on the influence of the strictness of the transition to tracked secondary education on academic performance in secondary education. As for the study of the effect of timing of tracking, the analysis relies on an educational reform. The reform under study was implemented in Bremen in 2003. It made non-binding teacher recommendations for secondary education binding (Bremische Bürgerschaft 2003). Most previous research shows no effect of strictness of tracking on educational decisions (Jähnen & Helbig 2015; Neugebauer 2010; Roth & Siegert 2016). However, for the effect of strictness on academic performance, an effect on performance in elementary school has been demonstrated (Bach & Fischer 2020). For performance in secondary education, there are mixed results (Esser & Seuring 2020; Heisig & Matthewes 2021). Similar to the analysis of the education reform in Lower Saxony, the analysis of the reform in Bremen is based on the extension of PISA, namely PISA-E.

1.2 Educational system in Germany and educational reforms

Before discussing theoretical arguments and empirical results on the effects of tracking, the following section gives a general overview of the structure, characteristics, and some developments of the education system for primary and secondary education in Germany. Several aspects of the German school system, however, will be described in more detail in the study chapters.

The first grade of elementary school begins when children are around six years old and lasts until grade four or six, depending on the federal state. In elementary school, students are taught jointly without being separated into ability groups, regardless of the level of graduation they are aiming for (Henninges et al. 2019). This is followed by the transition to secondary education, where students are usually placed in different school types, each leading to different levels of educational degrees. For the transition to secondary education, the class teachers or the conference of class teachers give a recommendation for each student for the school type that corresponds to the student's performance potential. Depending on the state, the strictness of the transition varies, meaning that the teacher recommendation can be binding or non-binding. the freedom of educational decision-making for secondary education, i.e., how much families have to say in educational decision-making depends on binding or non-binding recommendations (Helbig & Nikolai 2015). Can families alone decide about a child's education (non-binding recommendations), or do teachers decide about the educational trajectory of a child (binding recommendations)? But even with binding recommendations, families can bypass teachers' recommendations about educational trajectories through a variety of options, allowing them to send their child on a more challenging school type despite the lack of a recommendation, such as requiring students to pass a trial phase (Helbig & Nikolai 2015).

After the transition to secondary education, the placement of students into separate school types takes place, as mentioned earlier. Compared to other countries, the division into school types with ten- to twelve-year-old students in Germany is early (Le Donné 2014). The traditional three-tier education system consists of two school types that lead to lower secondary qualifications, the lower secondary school, named *Hauptschule*, and the intermediate secondary school, named *Realschule*. These two schools lead to school-leaving qualifications that are mostly required for vocational training. Lower secondary education goes up to grade nine or ten, depending on the state (Eckhardt 2017). Unlike *Hauptschule* and *Realschule*, the academic track, called

Gymnasium, exists in every federal state. It leads to the university entrance qualification (*Abitur*). Depending on the federal state and the school, this qualification is obtained after the twelfth or thirteenth grade (Henninges et al. 2019; Eckhardt 2017). The *Abitur* gives unrestricted access to tertiary education. Educational reforms in the 1960s created other restricted possibilities to access tertiary education, in addition to the *Abitur* (Schindler 2017). However, the *Abitur* remains the most usual way for access to tertiary education (Müller et al. 2011). From the tenth or eleventh grade onwards, this stage of secondary education is known as upper secondary education (*gymnasiale Oberstufe*).

In the late 1960s, the German Education Council (*Deutscher Bildungsrat*) established comprehensive schools, initially on an experimental basis. In the early 1980s, comprehensive schools were established in most of the German states as an alternative school type (Köller 2008). All school-leaving qualifications can be obtained at comprehensive schools, although not all comprehensive schools offer this option (Köller 2008). There are two types of comprehensive schools, cooperative and integrative comprehensive schools. Cooperative comprehensive schools are very similar in structure to the traditionally tracked school system. The individual school types are combined under one management. However, students are taught separately in each school type. This means that the between-school tracking of the tracked school system has been replaced by within-school tracking (Henninges et al. 2019). In integrated comprehensive schools, students are also taught separately, but only in some subjects. For this purpose, students are divided into ability groups, but still remain in the same class, a practice also called setting (Henninges et al. 2019; Köller 2008).

Another school type that was introduced in some federal states but no longer exists today was the school type-independent orientation stage. The orientation stage was an administratively independent school in the first two years of secondary education (grade five to six), which led to a postponed placement into the different school types in the secondary education by two years. The teachers were composed of the different school types of secondary education. Students in the orientation stage were divided into ability groups in some subjects. The goal of the orientation stage was to achieve a better placement of students in secondary school types given their performance (Schuchart 2006, 2003).

The described traditional three-tier school system has been changed more and more over the years. The federal states differ to some extent in the changes they have made. Several states have moved away from the traditional three-tier school system more clearly than others (Becker

et al. 2017). Several states have merged the two school types that lead to lower secondary qualifications. At these so-called schools with two educational programs (*Schularten mit zwei Bildungsgängen*), students can attain both lower secondary degrees (Henninges et al. 2019). In other states, there are school types in which students can obtain all three school-leaving certificates. They are called schools with three educational programs (*Schularten mit drei Bildungsgängen*) (Henninges et al. 2019). Just as in integrated comprehensive schools, in these two school types students are also divided into ability groups for certain subjects. Thus, these school types are similar in some ways to the integrated comprehensive school (Becker et al. 2017).

1.3 Theoretical mechanisms

Tracking can affect educational inequality through different mechanisms. It can influence students' academic performance via the class's performance composition (Betts 2011), and the timing of the educational decision changes the framing of the decision, which can alter specific decision behavior by social background (Berger & Combet 2017). Moreover, the strictness of tracking could limit the influence of social background on educational transition (Dollmann 2016). However, before discussing these mechanisms, the first step will be to outline the origins of social educational inequality.

1.3.1 Primary, secondary, and tertiary effects

Students' access to different levels of education depends strongly on their socioeconomic status (SES) backgrounds (Breen & Jonsson 2005). Two origin effects usually explain this: First, a higher SES background inherently augments students' educational performance (primary effects). Second, higher-SES students are more likely to attend more demanding school tracks independent of their performance (secondary effects) (Boudon 1974). In addition, tertiary effects, which extend these two origin effects, are unequal assessments of performance by teachers, depending on a student's social background (Helbig & Morar 2017). Primary effects describe the influence of SES background on educational performance transmitted through cultural, social, and economic resources relevant to educational success. High-status families, for example, can invest money in tutoring with hopes of improving insufficient academic performance (Boudon 1974; Luplow & Schneider 2014). More resources usually mean higher academic performance. Additionally, the higher the social background, the more resources are available.

Secondary effects are SES-specific educational decisions independent of students' performance, which means students of high or low SES backgrounds and the same educational performance, for example, will make different track choices. High-SES students will more likely choose an academic track while, in contrast, low-SES students tend to choose a vocational track, all else being equal. The reason for this SES-specific decision-making process is a different cost-benefit calculation shaped by three subjectively assessed factors: cost of education, the success probability of completing a certain educational path, and that track's benefits (Breen & Goldthorpe 1997). First, education requires financial resources, and perceptions of those costs vary across social backgrounds based on different financial endowments. As a result, lower-SES students perceive the cost of additional education as higher (Stocké 2007; Breen & Goldthorpe 1997). Secondly, successful completion of a given school track depends on the student's abilities. Primary effects lead to unequal distribution of educational performance by SES background, leading in turn to a SES-specific difference in the average success probability of completing a particular educational track (Breen & Goldthorpe 1997; Stocké 2007). In addition, there is a SES-specific perception of the difficulty of an educational path. The lower the social origin, the more the difficulty of a higher educational path is overestimated (Erikson & Jonsson 1996). The last factor is the SES-specific benefit of a certain educational degree, which differs by social background due to the motive of status maintenance. Accordingly, minimizing the risk of intergenerational status decline motivates parents independent of SES. Because of the high correlation between education and social position, it is crucial for high-SES families that their offspring reaches high educational degrees to secure the social status for further generations. In turn, students from low-SES families do not need a high degree to maintain parental social status (Breen & Goldthorpe 1997; Stocké 2007).

Tertiary effects are teacher expectations and evaluations based on stereotyped expectations according to students' social background. For example, teachers evaluate the same performance differently by students depending on their social background (Esser 2016). This leads to different grades in school and can affect transition patterns in secondary education (Helbig & Morar 2017). The potential for parental support is also important. The higher the social background of students, the higher teachers assess the potential for parental support. This in turn may influence the recommendation for secondary education (Helbig & Morar 2017). This extension is controversial, however, as some researchers point out that teacher evaluations can already be included in the primary and secondary origin effects (Helbig & Morar 2017).

1.3.2 Tracking and students' performance development

Before describing how timing and strictness of tracking can affect student performance via the composition of performance within the classroom, the first part of this section initially clarifies how tracking might affect the composition of performance. The performance composition within a class differs, with the variation of timing of tracking. Later tracking into different school types should lead to classes being more heterogeneous in terms of performance. As students who are aspiring to different school-leaving qualifications and with different abilities are in the same class and are only taught separately for certain subjects, for example, as in the case of integrated comprehensive schools or the orientation stage. Early tracking, on the other hand, leads to more performance-homogeneous classes, as students are placed into different school types according to their aspired school-leaving qualifications and abilities. This is the case in traditional tracked school types of secondary education in Germany. In addition to timing of tracking, the strictness of tracking can also influence the performance composition of classrooms. Depending on the level of strictness, the performance composition of classes should also vary, since the basis for teachers' recommendations, as opposed to parents' decisions, should be stronger correlated with students' performance (Ditton et al. 2005). The more strict the tracking, the lower the variance of the performance composition. Therefore, with binding teacher recommendations, the composition of performance in classrooms in secondary education should be more homogeneous compared to non-binding recommendations (Esser & Hoenig 2018).

As described above, with timing and strictness of tracking the performance composition of classes varies. This in turn can affect students' performance development, through homogeneity and heterogeneity of performance composition within a classroom (Betts 2011). Two different views emerged in the literature on the effects of variance of performance composition on students' performance development. One perspective advocates early homogenization because homogeneity of performance is seen as a more efficient learning environment. It is argued that homogenization of students' performance within the classroom can positively influence students' performance development through more accurate matching of the curriculum and instructional tempo, as well as ability-appropriate teaching methods, making learning more efficient, benefiting all students, and increasing overall performance (Matthewes 2021; Esser & Hoenig 2018; Figlio & Page 2002; Betts 2011). Therefore, according to this perspective, it is important to place students in different school types early and strictly based on performance.

The other view questions the efficiency argument of performance homogenization, noting that promoting homogenization could reinforce educational inequalities related to social origin. They argue that this is particularly the case when tracking to different school types begins early in children's educational careers and characteristics other than prior performance determine placement in a school type in secondary education (Betts 2011). In homogeneous classes, students in less challenging school types lose their high-performing classmates. However, their presence would have benefited them in their learning, for example, by discussing questions or problems directly with high-performing peers. High-performing students, on the other hand, may also benefit from the presence of lower-performing students because it allows them to consolidate their knowledge (Sacerdote 2011). Thus, separating students by ability can widen the performance gap between students in lower and higher tracks and lead to a decline in average performance (Betts 2011).

In summary, it is theoretically unclear how timing of tracking and strictness of tracking affect student performance. Both effects of homogeneity and heterogeneity of classroom performance composition are theoretically plausible. On the one hand, performance homogenization could increase teaching efficiency, benefiting all students. On the other hand, the homogenization of performance could hinder interactions between students with different performance levels, which is important for learning, and thus may increase social inequalities.

1.3.3 Tracking and educational decisions

In addition to performance, tracking can also influence educational decision-making. Timing of tracking changes the conditions of the educational decision through the timing of the decision (Berger & Combet 2017). The strictness of tracking can limit the influence of parents and thus the influence of social background in the transition process (Dollmann 2016). However, before presenting the influence of tracking on educational decisions, the relationship between social background and educational decisions should be looked at again in more detail.

As already established, the motive of status maintenance states that parents want to secure their own social status over the generations (Breen & Goldthorpe 1997). For this, children with a higher social background must also achieve a higher level of education in order to maintain the status over the generations, compared to children with a lower social background. This motive can also be derived from prospect theory (Berger & Combet 2017). Prospect theory focuses on decision-making and highlights that possible outcomes are evaluated based on a reference point. If an outcome deviates negatively from this reference point, actors react more strongly than if

it deviates positively. Actors particularly want to avoid losses, i.e. negative deviations from the reference point, and therefore actors faced with a possible loss behave in a more risk-affirming manner. In contrast, in the case of a possible gain, actors want to avoid risks (Kahneman & Tversky 1979). In the context of social status and the educational decision, the reference point differs by social origin, as it is the current family status. Thus, families with a high status face a possible loss if the status cannot be maintained. Therefore, they would still choose a more demanding school type even under poor conditions, for example, if the child's academic performance was low. Low-status families, on the other hand, face a possible gain, so they want to avoid risk in the child's educational decision. For this reason, they are more likely to choose a less demanding school type, even if their performance is equivalent to high-status students. Because investing in higher education carries uncertainty about the probability of success and with it the risk of failure (Berger & Combet 2017).

To influence decision behavior, one can vary the uncertainty of the probability of success and thus the investment risk. The more uncertain the probability of success, the greater the investment risk. Depending on the social background, actors react differently to these variations. In order to maintain status, families with a high status must also accept a low probability of success, since they act in a risk-affirming manner due to the potential loss of status. While families with a low status, will only at a low investment risk, that is a high probability of success, opt for higher education, since they have a potential gain in front of them and act risk-averse (Berger & Combet 2017).

Much of the uncertainty about the probability of success depends on the uncertainty about a student's performance. The better a student's performance, the more likely it is that a high level of education will be successfully achieved (Berger & Combet 2017). Nevertheless, actors cannot be certain that a given level of performance will be maintained over time. The more information actors have about prior performance, however, the more certain they can be about the future trajectory of performance development. In this way, timing of tracking can influence the educational decision. As the timing of tracking varies, so does the amount of information obtained about performance development. With more information about performance development, uncertainty about performance development decreases and so does uncertainty about the probability of success. This should allow for a more rational decision and reduce the influence of subjective assessments that depend on social background (Berger & Combet 2017; Bauer & Riphahn 2006). A later timing of tracking should therefore increase the probability of students with a lower social background opting for higher education. This is especially true for higher

performing students with lower social backgrounds. Students with a higher social background are more risk affirming anyway and are more likely to opt for higher education even with greater uncertainty.

However, the timing of tracking should not only influence secondary effects but also tertiary effects, as teachers should have more time to evaluate and adjust stereotyped expectations about students' performance development, based on students' social background, with actual student performance development. This should lead to a decrease in tertiary effects, as teacher evaluations, especially pre-tracking, should become more student performance oriented and the influence of student social background decreases. Yet, the effects on tertiary effects are not the focus of this contribution.

As noted earlier, as the strictness of tracking varies so does the freedom for families to make their own decisions about the transition to secondary education. The stricter the educational transition is thereby designed, the less freedom of choice families have and the more teacher recommendation determine post-transition education. Thus, parents' socially biased decision-making behavior should have less influence on children's education (Dollmann 2016; Ditton et al. 2005; Roth & Siegert 2016). The recommendation of teachers therefore increases in importance in the transition process. However, the recommendation is not influenced by the degree of strictness.

Taken together, the described mechanisms allow tracking through the timing and the strictness of tracking to affect student performance as well as educational choices and, consequently, educational inequities. In this context, within-class variance in performance affects student performance and the within-class variance of performance varies with the timing of tracking and the strictness of tracking. The timing of tracking also changes the context of educational decision-making and thus can influence educational inequality. In addition, the strictness of tracking alters the influence of secondary effects in the formation of educational inequality. A detailed analysis of tracking within the educational system can thus provide valuable policy implications for future reforms.

1.4 Previous research on the effects of tracking on students' performance, educational decisions, and research gaps

The following section provides an overview of the current state of research. For this purpose, this section reports the results from different research areas. First, we look shortly at the basic relationship between social origin and educational inequality, specifically at the origin effects. Next, we briefly present results on the relationship between educational systems and educational inequality. Third, we focus on the relationship between performance variance in classes and students' performance development. In this context, we focus on results of the effects of timing of tracking and strictness of tracking on performance. Finally, results of the effects of timing and strictness of tracking on educational decisions and attainment are presented.

1.4.1 Social origin and education

Research consistently has shown that higher social background is positively associated to obtaining higher educational credentials (Breen & Jonsson 2005). Students with a higher social background have shown higher academic performance on average, and these students are also more likely to go on to more demanding education, even at the same performance levels as students from lower social backgrounds (Jackson et al. 2007). In addition, research has demonstrated that students are graded differently. For the same level of performance, students from higher social backgrounds receive higher grades compared to students from lower social backgrounds (Helbig & Morar 2017). The different transitions in the educational system also indicate that secondary effects increase in importance over the course of the educational career. In the transition from primary to secondary education, primary effects are most significant, but in the transition from secondary to tertiary education, secondary effects are more important in the formation of educational inequality between students with different social backgrounds (Neugebauer et al. 2013; Scharf et al. 2020).

1.4.2 Educational systems and educational inequality

International comparative research has often showed that certain characteristics of education systems are associated with inequalities in educational attainment and school performance. In particular, the mode of tracking seems to explain much variance across countries. The earlier students are placed in different school types, the stronger the link between social origin and school performance as well as educational attainment (Hanushek & Wössmann 2006; Van de Werfhorst, Herman G. & Mijs 2010; Pfeffer 2008). However, these studies are often based on

data that do not show an attained degree but placement in a school type in secondary education (Brunello & Checchi 2007). Studies based on full school careers find different or less strong associations for tracking and educational attainment as well as school performance (Van de Werfhorst, Herman G. 2019; Heisig et al. 2020). In addition, the differences in tracking between countries are oversimplified into early and late tracking in most studies. Yet, types of performance differentiation also exist in the so-called late tracking systems. As in early tracking systems, this too can lead to certain path dependencies for students (Schindler et al. 2021). Thus, it turns out that a more precise classification of tracking in international comparative research hardly explains the connection between social origin and education (Schindler et al. 2021). An analysis of differences in tracking within a country can therefore be useful since differences between countries do not play a role there. Although education systems differ between federal states in Germany, this difference should be less pronounced compared to differences between countries.

1.4.3 Variance in performance composition and students' performance

Regarding the influence of performance variance in classrooms on the performance development of students, there are contrasting results. There are indications that heterogeneous or homogeneous learning groups do not show different performance development (Gröhlich et al. 2009). However, there is also evidence showing that both heterogeneity and homogeneity can be beneficial for performance (Esser & Seuring 2020; Scharenberg 2012).

Previous research has repeatedly shown for different countries and contexts that later placement of students in school types of secondary education has a positive effect on general and subject-specific performance development (Bygren 2016; Hanushek & Wössmann 2006; Horn 2013; Jakubowski et al. 2016; Korthals & Dronkers 2016; Piopiunik 2014). These results suggest that greater performance variance is useful for students' performance development. However, a meta-study shows that tracking has no effect on average performance (Terrin & Triventi 2022). Further analyses for Germany that distinguish students by performance groups show that higher-performing students are more likely to benefit from early tracking in performance development (Roller & Steinberg 2020). A variance in the effect of tracking on performance was also found by other research (Lavrijsen & Nicaise 2016). While additional research could also show, that the positive effects of late tracking on performance are attributable to the improved performance of lower-performing students (Matthewes 2021). Social inequalities in academic performance can be reduced by later tracking (Terrin & Triventi 2022). These results illustrate

that timing of tracking does not have the same effect on performance development for all students, but that the effect differs by students' performance level.

Regarding the relationship between strictness of tracking and student performance, there are mixed results. For primary education, binding recommendations show a positive effect on performance. For students, binding recommendations function as an incentive for better performance. Yet there is a trade-off because, binding recommendations also decrease intrinsic motivation to learn and students feel more pressure (Bach & Fischer 2020). In particular, the decrease in intrinsic motivation to learn could lead to slower or stagnant performance development after the transition to secondary education. For performance in secondary education, however, some results suggest that performance increases when transition to secondary education is strict in Germany (Esser & Relikowski 2015) and that performance equity increases in the transition process to secondary education (Esser & Hoenig 2018). This positive effect on educational performance seems particularly pronounced in a strict transition setting for students in less ambitious educational tracks (Esser & Seuring 2020). However, a reanalysis and extension of Esser and Seuring's analysis shows no effect of classroom homogeneity on students' performance (Heisig & Matthewes 2021). Heisig and Matthewes (2021) also show that the performance composition in the classroom does not mediate the relationship between strictness of tracking and academic performance. Therefore, this finding disputes the mechanism underlying the positive effect of strict tracking on performance. While the results for secondary education rely mostly on multilevel analysis, the results for primary education are based, among others, on difference-in-differences analysis. Despite the difference in the methods used, the empirical findings on the effect of strictness of tracking on academic performance do not allow for a clear conclusion and more research is needed. Results suggest a positive effect of strict tracking on performance in primary education, but the effect in secondary education is inconclusive.

The state of research on the effect of the timing and strictness of tracking on performance varies. While it cannot be clearly stated for the strictness of tracking how it affects performance development, for the timing of tracking, however, it can be noted that later tracking can reduce inequalities in academic performance.

1.4.4 Timing of tracking and educational decisions

The empirical findings on the effect of timing of tracking on educational decision-making are based on different research approaches. One experimental study uses a choice experiment to examine the effect of timing of decision on individuals who either stand to gain (lower social

background) or stand to lose (higher social background). In the experiment, anagrams must be solved before and after the decision. Thus, it is also possible to examine effects for different performance groups. It is found that high performers who can gain are more likely to continue the experiment at the time of decision if the time of the decision is later, and that the decision of subjects who can lose is not influenced by the time of the decision (Berger & Combet 2017).

Other research uses educational reforms to examine a change in the timing of tracking on educational attainment or track choice. For Sweden, research has shown that a change in the education system from an early to a late tracking school system substantially raised educational attainment, especially for students with a low educational background (Meghir & Palme 2005). For a similar reform in Finland, however, the research shows that there is no significant effect of later timing of tracking on the choice of an academically oriented track for all students as well as for students with high or low educational backgrounds (Pekkarinen 2008). For an educational reform in the German federal state of Lower Saxony, that introduced the orientation stage, which led to a delayed timing of tracking, research has found that while there is no effect on average educational attainment, the educational attainment of students with low educational backgrounds increases, while the educational attainment of students with high educational backgrounds decreases (Lange & Werder 2017).

Results for Switzerland have demonstrated that the relationship between educational background and educational attainment is stronger in cantons with early tracking than in cantons with later tracking (Bauer & Riphahn 2006). However, these results cannot be replicated by other research (Combet 2019). Looking at the overall research results to date, the findings on the effect of the timing of tracking on educational decision-making or educational attainment do not show clear results and are thus inconclusive.

1.4.5 Strictness of tracking and educational decisions

Regarding the strictness of tracking, it appears that students with a higher social background are more likely to attend a *Gymnasium* and that this correlation is stronger in federal states with a binding teacher recommendation (Gresch et al. 2010). This is inconsistent with the theoretical arguments that binding recommendations should limit the influence of parental background. Other findings that examine the effects of teacher recommendations on educational attainment, mostly through a legislative change, also neither have indicated an effect of strictness of tracking on educational attainment nor found effects for specific background groups (Neugebauer

2010; Jähnen & Helbig 2015; Roth & Siegert 2016). Only individual findings, which are spatially limited, have shown a decrease in social educational inequality (Dollmann 2016).

1.4.6 Research gaps

Research findings to date draw a mixed picture of the impact of tracking. Therefore, it is important to investigate further the impact of different aspects of tracking, especially in different contexts. This contribution addresses three gaps in this regard. First, the extent to which comprehensive schools are related to educational inequalities has been studied since the introduction of comprehensive schools in Germany. However, this research is relatively old (Tillmann 1988). Today, the number of students in comprehensive schools is increasing (Autorengruppe Bildungsberichterstattung 2018). In addition, some German states have introduced school types that are in part structured quite similar to comprehensive schools (Becker et al. 2017). In light of the mixed empirical findings on the impact of timing of tracking, an updated analysis of the relationship between comprehensive schools and educational inequality is therefore a gap that this contribution aims to fill. Second, another context in which this contribution examines tracking is the termination of the orientation stage in Lower Saxony in 2004. It has already been shown that there was an effect on the performance development of students, in particular giving higher-performing students an advantage due to the shortening of the comprehensive school period (Roller & Steinberg 2020). However, it is not yet clear whether the reform, which altered the timing of tracking, also led to a change in transition behavior to secondary education, as theory would suggest. Third, for Germany, it has already been shown quite consistently that the strictness of tracking has no influence on educational decisions (Neugebauer 2010; Jähnen & Helbig 2015; Roth & Siegert 2016). However, this consistency in research findings is not the case with regard to the influence on performance development (Esser & Seuring 2020; Heisig & Matthewes 2021; Bach & Fischer 2020). Therefore, this contribution examines the impact of a legislative change in the strictness of tracking through an educational reform on the performance development of students in secondary education for the first time.

1.5 Overview

This contribution includes three empirical studies and a concluding discussion of the results. Each study examines specific aspects of tracking in different contexts. The studies can be read independently from one another, as they each contain the necessary theoretical background and research review. The following section provides a brief overview of each study.

Study 1 (Chapter 2) examines the relationship between integrated comprehensive schools and *Abitur* attainment. This involves a comparison between integrated comprehensive schools and traditionally tracked schools of secondary education using data from the National Educational Panel Study (NEPS). Since comprehensive schools represent an alternative school type and coexist with the traditional tracked school types, there is a self-selection of students attending a comprehensive school. Therefore, comprehensive schools and traditional tracked schools should not be compared directly. However, in order to compare students from these school types and to investigate whether comprehensive schools change the likelihood of obtaining an *Abitur* compared to traditional tracked schools, propensity score matching procedures are used. With this method, similar students from comprehensive schools and the traditional tracked schools are selected based on observed variables. As a result, the two student populations should be comparable. This is the first study to use this method to examine the relationship between comprehensive schools and *Abitur* attainment.

Study 2 (Chapter 3) examines the effect of an educational reform in Lower Saxony on attendance at a *Gymnasium* in ninth-grade. This reform abolished a two-year orientation stage independent of school type. As a result, the timing of tracking was preponed by two years. For this purpose, the study uses PISA-E and IQB-LV data to examine the educational reform with difference-in-difference and difference-in-difference-in-difference estimators. A similar study already examined this reform (Roller & Steinberg 2020), however, it examined the effect of the reform on student performance. The results show a positive effect on the performance of high-performing students and a negative effect for low-performing students. In addition, students with a high educational background benefit more. The results of the previous study, as well as the theoretical arguments regarding the effect of timing tracking on educational decision-making, suggest an effect of the reform on transition behavior. However, the extent to which the reform also influenced transition behavior is still unknown. Therefore, *Study 2* contributes to the literature by analyzing the effect of timing of tracking on the decision to attend *Gymnasium* in another context compared to the previous research.

Study 3 (Chapter 4) also examines the effect of an educational reform. In this reform, the strictness of tracking in Bremen changed from non-binding recommendations to binding recommendations. For this reform, the effect on students' reading and math performance in the ninth grade is examined. For this purpose, it uses PISA-E data and analyzes the data with difference-in-difference estimators. Previous research focusing on performance effects in secondary educa-

tion mostly used multilevel models, with mixed results (Esser & Seuring 2020; Heisig & Matthewes 2021). Therefore, this analysis uses an educational reform in Bremen that made non-binding recommendations binding. *Study 3* investigates the effect of strictness of tracking on performance in secondary education using a different research design and methods as the previous research.

This contribution ends with a final discussion of the results and a conclusion (Chapter 5). To do so, the chapter summarizes the findings of the three studies and places them in the larger research context to develop a conclusion on tracking and its relationship to educational inequality. This section also discusses the limitations of the analyses of this contribution, based on which it also provides recommendations for further research.

2 Association between late tracking and *Abitur* attainment: A comparison between comprehensive schools and tracked schools in Germany

Abstract

Germany has an early-stratified education system. However, comprehensive schools, an alternative school type in Germany, divide students differently and later in their educational careers. Comprehensive schools' implementation in 1969 reflected reform goals to reduce educational inequalities. International research shows that timing of tracking influences educational inequalities through effects on performance development and on educational decision-making. Typically, late tracking relates to less socioeconomic status (SES) inequality while early tracking produces greater SES inequalities in the context of educational disparities. It is unclear how comprehensive schools alter educational inequalities in Germany. Using data from the National Educational Panel Study (NEPS), we analyze the association of comprehensive schools with upper secondary completion using propensity score matching. The results show no change in the probability of completing upper secondary education for students in comprehensive schools. However, consistent with prior research on the timing of tracking, the results show a statically significant increase in the likelihood of completing upper secondary education for low-SES students and for students with low initial performance in comprehensive schools. Reform goals connected to comprehensive schools seem realized. Comprehensive schools could help reduce social inequalities in the education system.

2.1 Introduction

Educational inequalities are responsible for many dimensions of inequality (e.g. labor market success or personal health) (Card 1999; Leuven et al. 2016). This highlights the social relevance of minimizing educational inequalities. Between countries, inequalities within educational systems vary. One explanation for the variation is timing of tracking (Le Donné 2014). Countries with a late-tracking education system, differentiating later between two or more educational tracks (e.g. an academic-orientated track and a vocational-orientated track), tend to have less inequality within their education systems in terms of performance development and educational decision-making. In contrast, countries with early tracking systems tend to have greater educational inequalities (Schütz et al. 2008).

Timing of tracking can influence educational inequalities in two ways. First, it can influence students' performance development in school through different mechanisms in both positive and negative ways. In homogenous classes, teachers can more accurately match pedagogical methods to student performance, which could have a positive effect (Figlio & Page 2002). The argument in favor of heterogeneous classes, on the other hand, is that interchange between classmates with different performance levels could be performance-enhancing for all (Maaz et al. 2008). Empirical studies mostly indicate a positive effect of late tracking on performance (Cordero et al. 2018). However, timing of tracking seems to affect high and low performing students differently (Roller & Steinberg 2020). Secondly, it can affect the decision-making process for secondary education because late tracking should reduce uncertainty about future performance and enable more rational decisions (Berger & Combet 2017). A positive effect on low-SES students seems especially plausible. Studies focus less frequently on the effect of timing of tracking on the decision-making process than they do on that on performance. These studies have mixed results, with some confirming theoretical expectations while others do not (Meghir & Palme 2005; Bauer & Riphahn 2006; Combet 2019).

The educational system in Germany is considered an early and highly stratifying system. In most federal states, students separate after the fourth grade into different secondary school types, which lead to different school degrees with various properties for continuing education (Eckhardt 2017). Within Germany, a long political debate about the early and highly stratifying educational system dates back to the 1960's. One outcome of this debate was the introduction of a late tracking school type, the comprehensive school (Köller 2008). Comprehensive schools generally combine all secondary school types of the German school system in one type of

school. By focusing on Germany, it is possible to test certain claims about the inequality-reducing effects of comprehensive schools, particularly claims concerning the educational-inequality-reducing effect of comprehensive schools (Köller 2008), a school type with increasing numbers of students (Autorengruppe Bildungsberichterstattung 2018). The aim of this study is to estimate the association between comprehensive schools and students' completion of upper secondary education (*Abitur*) using propensity score matching.

2.2 Education system in Germany and comprehensive schools

In Germany, children begin attending primary school (*Grundschule*) around the age of six. Primary education covers the first four grades.[1] Around the age of ten, the first transition takes place: Students split into two or three different secondary educational school tracks. The number of secondary school tracks depends on the federal state. The lowest secondary track is the lower secondary school (*Hauptschule*), offering a certificate after grade nine or ten. After grade ten, students in Germany receive a certificate from the intermediate secondary school (*Realschule*). Some federal states are currently replacing or have already replaced the lower and intermediate schools with a combination of lower and intermediate secondary schools offering both certificates, called schools with two educational programs (*Schule mit zwei Bildungsgängen*). While *Hauptschule*, *Realschule*, and schools with two educational programs are vocational tracks leading into apprenticeships, the academic track is *Gymnasium*. After grade twelve or thirteen (depending on the federal state), it offers the higher education entrance qualification (*Abitur*), the most common way to enter university in Germany (Eckhardt 2017; Maaz et al. 2008).[2] In addition to the *Gymnasium*, some states have introduced or are introducing schools with three educational programs (*Schularten mit drei Bildungsgängen*), including another path to the *Abitur*. As a result, these states are moving further from the traditional two- or three-tier education system (Autorengruppe Bildungsberichterstattung 2020; Becker et al. 2017). From grade ten or eleven onwards (depending on the federal state), the grades at a *Gymnasium* and other school types which offer the *Abitur* are also part of the general upper secondary track (*gymnasiale Oberstufe*) (Henninges et al. 2019). In the following analysis, however, we compare comprehensive schools to tracked schools without considering further schools with

[1] The *Grundschule* covers six years only in Berlin and Brandenburg (Eckhardt 2017).

[2] The *Abitur* was for some time the only access to university. In the 1960s educational reforms introduced other possibilities, e.g. through professional qualification. However, these new paths were limited in their access to university, unlike the *Abitur*, which allows full access (Schindler 2017). While the status of the *Abitur* as the only possibility has changed, it is still the most usual path for enrollment into university (Müller et al. 2011).

two or three educational programs, especially given comprehensive schools' introduction as a kind of response to inequalities in the structured school system (Leschinsky & Mayer 1999) without these newer forms of schooling in mind and because schools with three educational programs' resemble comprehensive schools in some respects.

Most states offer comprehensive schools in addition to the other school tracks and in most states students in comprehensive schools can acquire any of the different school certificates (Helbig & Nikolai 2015). There are two types of comprehensive schools: cooperative and integrated. The cooperative comprehensive school is very similar to the tracked educational system as described above. Instead of between-school tracking, students in cooperative comprehensive schools subdivide into different educational tracks within the same school. However, the data employed do not enable ready identification of cooperative comprehensive schools, which (fortunately) are not the focus of the analysis. In contrast to cooperative comprehensive schools, integrated comprehensive schools (referred to as comprehensive schools) instruct students in some subjects together as a class and sort students by performance into ability groups for specific subjects, a practice of within-school tracking called setting (Henninges et al. 2019; Köller 2008). Hence, classes contain students seeking different educational aspirations (Eckhardt 2017). Comprehensive schools do not postpone the first transition for all students equally, because ability grouping takes place already in grades 5–7 (depending on the school and the subject) (Köller 2008). Like between-school tracking, ability grouping also increases the performance differences for both low- and high-performing students (Gamoran et al. 1995). Still, mobility between less ambitious and more ambitious courses is not uncommon (Köller 2008) so that high-performing students may participate in high-performance courses, thus postponing their educational decisions relative to those of low-performing students (postponed only slightly if at all).

In 1969, the German education council (*Deutscher Bildungsrat*) (Deutscher Bildungsrat 1969) implemented comprehensive schools as an educational experiment. The experiment ended in 1982, when most states established comprehensive schools as an alternative school type (Köller 2008) in pursuit of certain specific reform goals: postponing selection into different educational paths in the transition from comprehensive primary education to tracked secondary education in order to reduce the risk of educational investments for disadvantaged groups to reduce educational inequalities and promote equality of opportunity (Köller 2008; Leschinsky & Mayer 1999). However, not everyone agreed comprehensive schools would help reduce social inequalities. Opponents saw the traditional education system as sufficient for future developments,

rendering massive structural reform of the education system unnecessary and urged spending more resources on quality of teachers and instruction rather than on educational reform. Besides, conservative politicians did not want to support an educational system with clear similarities to those in socialist states (Leschinsky & Mayer 1999; Wenzler 2003).

This division between supporters and opponents of comprehensive schools led to comprehensive schools developing variously across Germany's federal states. Conservative-governed states chose to keep numbers of comprehensive schools low, while progressive-governed states established more comprehensive schools (Wenzler 2003). For example, Baden-Wuerttemberg, Bavaria, and—later—Saxony never established comprehensive schools as a regular school type. Concurrently, the percentage of students entering upper secondary education from comprehensive schools varies considerably between federal states. In Saxony-Anhalt, Thuringia, and Mecklenburg-Western Pomerania, only 1–4 percent of students in the general upper secondary track attend a comprehensive school, whilst in North Rhine-Westphalia, Hamburg, Berlin, and Brandenburg the proportion is from 15 to over 20 percent (Helbig & Nikolai 2015). Since 2005, the number of students in comprehensive schools has been increasing and, as of 2016, the number of students in comprehensive schools is as high as the number of students attending *Realschule* (Autorengruppe Bildungsberichterstattung 2018). In the 2012/13 school year, however, students from comprehensive schools, who transition to the general upper secondary track, form only a small minority (about 1 percent) of all students in Germany who make this transition. Around 93 percent of all students in the general upper secondary track attend a *Gymnasium* (Malecki et al. 2014). In general, however, the number of students obtaining the *Abitur* at comprehensive schools is increasing in federal states with comprehensive schools. In states with a lower density of comprehensive schools, the transition to the general upper secondary track is higher for students in comprehensive schools compared to comprehensive school students in states with a higher density of comprehensive schools. This suggests that, where comprehensive schools occupy a higher proportion of schools, they serve as a substitute for the *Gymnasium* for parents wishing their children to enter the *Abitur* (Helbig & Nikolai 2015). Students in comprehensive schools have a performance distribution more similar to that of a *Realschule* than those in a *Gymnasium*, with underrepresentation of students with a low or high learning ability (Leschinsky & Mayer 1999). Attending a comprehensive school appears to reflect a selection process based simultaneously on students' performance and SES (Köller 2008). Comparing comprehensive with tracked schools must take into account the selection of students based on performance, parental aspiration, and SES, as well as the state-specific opportunity structures for comprehensive school attendance.

2.3 Theoretical considerations and previous research on the effects of timing of tracking

Comprehensive schools have the goal of postponing selection into different educational paths. While within-school tracking practices make delays for all students arguable. Especially high-performing students should enroll in more demanding courses and should experience actual postponement of educational decisions. Ability grouping should also impact class composition as regards educational performance: composition by performance should be broader than in stratified schools. A potential explanation for effects of comprehensive schools on educational outcomes could be their postponement of tracking. The following section will present discussion of different theoretical arguments and research results with respect to timing of tracking.

2.3.1 Effect of tracking on performance

For low-SES students the primary and secondary effects translate into lower educational degrees and lower labor market outcomes. Comparative research repeatedly finds varying degrees of the correlation between social background and educational outcome depending on several arrangements of the educational system (Le Donné 2014; Pfeffer 2008). Prior research has identified timing of school tracking as one important difference between educational systems and a possible explanation for the variance of educational inequalities between countries (Pfeffer 2008). Still, the effects of timing of tracking on students' academic performance are theoretically ambiguous. As earlier tracking leads to more academically homogeneous classes, later tracking promotes more heterogeneous classes. On one hand, researchers argue that homogeneous classes allow teachers to adopt specific pedagogical methods for each performance group, thus increasing the performance of students of all educational levels. On the other, researchers have suggested that predominately high-SES students benefit in homogeneous classes while simultaneously leaving low-SES students behind (Figlio & Page 2002; Maaz et al. 2008; Piopiunik 2014). Late tracking and heterogeneous classes, in contrast, increase interaction between high-performing and low-performing students within classes. Students with high performance can help lower-performing classmates, thereby increasing their classmates' performance while consolidating their own knowledge. Thus, late tracking could have a positive effect on students' performance for all performance groups (Maaz et al. 2008; Piopiunik 2014).

Previous findings in this research area show predominantly positive causal effects from heterogeneous classes on performance development in general and also for specific fields (e.g. read-

ing, mathematics, and science), supporting a positive impact of late tracking on academic performance (Bygren 2016; Hanushek & Wössmann 2006; Horn 2013; Jakubowski et al. 2016; Korthals & Dronkers 2016; Lavrijsen & Nicaise 2015; Piopiunik 2014).[3] However, research for Germany shows that high-performing students benefit in their development from early tracking (Roller & Steinberg 2020), while positive effects of postponed tracking center upon low-performing students (Matthewes 2021), suggesting a variance in effects due to late tracking for different performance groups rather than a uniformly positive effect.

2.3.1 Effect of tracking on educational decisions

In addition to performance, timing of tracking can also influence the educational decision-making process. The earlier an educational system tracks its students, the greater the uncertainty about students' future abilities. For the process of educational decision-making, early tracking means a higher correlation between SES background and the assessed probability of success completing a certain educational path. As late tracking reduces the uncertainty regarding school performance and its future development, it enables more rational decisions about investments in education. Postponing the educational decision makes the information about students' performance more reliable, increasing in turn the assessed probabilities of successfully obtaining more demanding levels of education (Bauer & Riphahn 2006; Berger & Combet 2017). Thus, late tracking within the educational system should reduce the association between SES background and secondary educational decisions (Dustmann 2004), which should lead to a higher number of low-SES students trying to complete a more demanding upper secondary education, especially for low-SES students with higher performances, since they will have a higher probability of completing upper secondary education compared to low-SES students with lower educational performance. Late tracking should not affect high-SES students' decision-making: Because of the motive of status maintenance, high-SES students, regardless of performance, need to reach more demanding and prestigious degrees.

In line with this assumption, previous research finds an increase in upper secondary and tertiary education for students with low-SES backgrounds following postponed tracking. Research shows greater inequality in early tracking systems (Bauer & Riphahn 2006). However, other findings cannot confirm this result (Combet 2019). Contrary to theoretical considerations, low-SES students with low educational performances, not low SES students with high performance, drive this increase in upper secondary students (Meghir & Palme 2005). However, for tertiary

[3] For more international research, see for instance Cordero et al. (2018) or Webbink (2005).

education, early tracking appears to have a negative effect for high-SES students, students with higher performances, and low-SES students with high performance (Meghir & Palme 2005; van Elk et al. 2011). In addition, late tracking seems to promote the expectancy of students completing tertiary education (Lee 2014).

2.3.3 Hypotheses

The theoretical considerations and institutional setting of the educational system in Germany lead to different hypotheses about the associations between comprehensive schools and *Abitur* attainment. Ambiguous arguments about the effect of late tracking on performance development, as illustrated above, lead to conflicting hypotheses as to the effect of late tracking on educational inequalities. However, the effect of late tracking on the decision-making process is theoretically clear: late tracking is especially beneficial for lower-SES students but gives no disadvantage to upper-SES students. In addition, empirical research shows that late tracking has differing effects on performance development depending on students' performance, reducing inequality with regard to the attainment of educational qualifications. Therefore, the overall hypotheses assume no association between comprehensive schools and *Abitur* attainment.

H1: Comprehensive schools do not increase the probability of obtaining a higher secondary degree in general.

Theoretical considerations and empirical research suggest SES-specific associations of late tracking with obtaining the *Abitur*. Because of the reduced uncertainty as regards performance development and the following smaller association between students' SES and educational decisions, low-SES students should benefit from comprehensive schools.

H2: Comprehensive schools increase the probability of obtaining a higher secondary degree, especially for low-SES students.

In addition to an SES-specific association, some theoretical arguments and empirical findings also suggest a performance-specific association of comprehensive schools with obtaining the *Abitur*. Low-performing students in particular should benefit in their performance development from comprehensive schools, enabling them to obtain the *Abitur* despite poorer initial performance.

H3: Comprehensive schools increase the probability of obtaining a higher secondary degree, especially for students with no recommendation for upper secondary education.

2.4 Data, analytical strategy and variables

In our analyses, we use data from Starting Cohort 4 (SC4) of the National Educational Panel Survey (NEPS) (Blossfeld et al. 2011).[4] SC4 has 15,239 students in 629 schools. All students were in the ninth grade when data collection started in school year 2010/11. In wave 9 of 2015/16, with 9,044 students remaining, all students had attained their secondary school diplomas with observable educational attainment. NEPS collects information from both students and also their parents, teachers, and school administrators (Leibniz Institute for Educational Trajectories 2017). All students who graduated from a *Hauptschule*, *Realschule*, *Gymnasium*, or comprehensive school and have no missing values are part of the sample. Not included are students from other school types (e.g. schools with two or three educational programs). Due to panel mortality and nonresponse, 5,095 students have no missing values for the variables used in the analyses.

Because students' assignments to a comprehensive school or a tracked school are not random, we use propensity score matching to estimate the change in the probability of obtaining the *Abitur* for comprehensive school students in Germany. While causal analyses typically employ propensity score matching, we will use it to account for the differences between students in comprehensive and tracked schools described earlier for the analysis. The basic idea behind propensity score matching is comparison of outcomes for observed individuals i with identical or nearly identical covariate values (Rosenbaum & Rubin 1985). The individuals are either in the treatment group $D_i = 1$ or in the control group $D_i = 0$, i.e. a student attended a comprehensive school or a *Hauptschule*, *Realschule* or *Gymnasium*. An individual can have the potential outcomes ${Y_i}^1$ if $D_i = 1$ or ${Y_i}^0$ if $D_i = 0$ (Rubin 1974), i.e. a student receiving the *Abitur* or not. To calculate the individual effect of the treatment on the outcome, one must subtract the potential outcomes of the individual ($\delta_i = {Y_i}^1 - {Y_i}^0$). However, an individual can only experience one of the two potential outcomes and it is not possible to observe an individual in both states simultaneously (Holland 1986). To overcome this problem, it is possible to identify the average treatment effect ($ATE = E({Y_i}^1 - {Y_i}^0)$). The ATE is the average change in outcome for all subjects, in this case: the average change in the probability of receiving the *Abitur*. Yet, in this

[4] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:9.1.1. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

specific case, it is more meaningful to calculate the average treatment effect on the treated ($ATT = E(Y_i^1 - Y_i^0|D = 1)$) because the focus lies on students in comprehensive schools and not students in general. The ATT is the average change of receiving the treatment on the outcome for those who received the treatment. The ATT can be interpreted as the average change in the probability of receiving the *Abitur* for a student from a comprehensive school.

Every individual has some true probability of receiving the treatment of going to a comprehensive school. The true propensity score $\pi(X)$ represents this treatment probability. X signifies pretreatment observed characteristics. We estimate this true propensity score using a logistic regression model of the treatment indicator on the observed pre-treatment covariates, called the assignment model, to generate predicted probabilities $\hat{\pi}(X)$. The predicted probability of receiving treatment, conditional on X, serves as an approximation of the true propensity score ($\pi(X) = \Pr(D_i = 1|X)$). Propensity score matching compares the outcomes of individuals with identical or almost identical propensity scores (Rosenbaum & Rubin 1985). With sufficient observed characteristics for the assignment model of the treatment to estimate the propensity score, matching methods can identify causal effects from treatment to outcome (Gangl 2015).

For identification of the causal effect of the treatment—here comprehensive school—on the outcome, here *Abitur* attainment, the observable covariate must emulate the social processes leading to allocation into the treatment and control groups, thereby removing the connection between treatment and outcome via what is called the conditional independence assumption (Gangl 2015). As discussed above, the selection of students into comprehensive schools in part reflects performance, parental aspiration, SES, and the state-specific opportunity structure The assignment model for propensity score matching must consider these characteristics.

Some challenges remain. First, students' assignment into treatment is not random and propensity score matching does not solve this problem. The results of the analyses enable conclusions on the change in *Abitur* completion for the self-selected student population in comprehensive schools. Nevertheless, it is not possible to infer from these analyses the associations of comprehensive schools with students in general. Secondly, comprehensive school vary in the intensity of treatment. For example, some comprehensive schools divide their students earlier than others. Additionally, low-performing students do not get to decide about their further education later in time, as students in low-performance courses do not have the option to opt for the *Abitur*. That implies an ambiguous treatment for all students in comprehensive schools. Late tracking only influences high-performance students in their educational decision-making, because those

students are sorted into high-performance groups. One consequence is that late tracking affects only part of the student population.

The outcome variable is dichotomous, indicating whether a student obtained the *Abitur* or not. The treatment variable is also dichotomous, showing whether a student attended a comprehensive school (treatment group) or not (control group). Covariates are different individual background variables, as well as specific structural characteristics. First, the individual background variables: We measure a students' SES background using highest parental ISEI (International Socioeconomic Index of Occupational Status) and highest parental secondary education. Parental education is dichotomized, indicating whether a parent has the *Abitur* or not. Educational performance is another covariate. SC4 measures a student's academic performance at multiple points in time. However, these measurements follow and thus already reflect tracking into a given school type and are therefore endogenous. At the end of primary education, students get school type recommendation for secondary education from teachers. While this is not a perfect representation of performance, it remains a relatively good indicator for performance (Birkelbach 2010). Moreover, teacher recommendation is not endogenous and thus serves as a proxy for educational performance. The teacher recommendation variable signifies a recommendation (or not) for upper secondary education. A dichotomous variable measures parental educational aspiration, indicating whether parents intend their child for *Abitur*.

Second, we also need measurements for specific structural characteristics: Indicators for the different federal states for state-specific characteristics, which could influence the probability of attainment of degrees. For instance, German states vary in how much discretion as to educational transitions at the end of primary education they allow parents independent of teacher recommendation. This varying strictness in transitions perhaps influences social inequalities in the educational system, but this effect and its extent remain under debate (Dollmann 2016; Roth & Siegert 2016). Yet, it may still have an effect on school choice decisions for parents and attainment of degrees. This could lead to a greater possibility for students to visit comprehensive schools in more strict educational systems. Besides the state indicator, some schools have specific admission conditions as well, e.g. an entry exam or trial lesson.[5] The admission conditions appear summarized in an index. The higher the value, the more important the criteria for the admission of students. Missing values (NAs) are coded to 0 and the models include an indicator for NAs in admission conditions. Comprehensive schools are more frequent in some

[5] Other admission conditions include a suitable teacher recommendation for the particular school track or a specific performance a student must demonstrate.

states than in others (Helbig & Nikolai 2015). This affects the opportunity to send a student to a comprehensive school. In addition, a region's urbanity can also influence the opportunity structure. The percentage of students for whom there is an alternative school with the same educational path nearer the current school—provided by the schools themselves—represents the opportunity structure. Similarly, regarding admission conditions, NAs are coded to 0 and the models include an indicator for NAs to reduce the loss of cases for analysis.

Beyond the variables mentioned, the assignment models also include interaction effects as revealed by previous research. The teacher recommendation and the aspiration to obtain the *Abitur* depend on parental background (Boudon 1974). Interaction variables for recommendation and aspiration with parental ISEI and education, respectively, are thus included. Another interaction between recommendation and aspiration is added also to the models. Because parental aspiration influences performance in school (Fan & Chen 2001) and thereby teacher recommendations.

Federal states vary in the amount of comprehensive schools (Helbig & Nikolai 2015) as well as in the number of cases in the sample. This leads to some federal states with no or a low number of observations of students in comprehensive schools in the sample. Additionally, as discussed above, some states have departed from stratified secondary education in Germany. As a result, some states have no or hardly any students in the *Hauptschule* and the *Realschule* (the *Gymnasium* exists in every state). Therefore, 2,037 observations from Baden-Wuerttemberg, Bremen, Brandenburg, Bavaria, Hamburg, Mecklenburg Western Pomerania, Saarland, Saxony, Saxony-Anhalt, and Thuringia are excluded from the analysis (see table 3 in the appendix). Students from Berlin, Hesse, Lower Saxony, North Rhine Westphalia, Rhineland Palatinate, and Schleswig-Holstein are part of the sample, yielding a final sample size of 3,058 students, including 386 students in comprehensive schools.

2.5 Analysis of the association of comprehensive schools with obtaining *Abitur*

The following analysis investigates the association of comprehensive school with *Abitur* completion in Germany. First, we present descriptive statistics, then the bias reduction from the matching algorithms, and finally presentation and discussion of the results of the analyses. Table 1 provides a summary of descriptive statistics for the sample, as well as for the two subgroups of students who graduated from comprehensive or tracked schools. In the overall sam-

ple, 67.3 percent of students achieved the *Abitur*, with 60.9 percent of students in comprehensive schools and 68.2 percent of students in tracked schools achieving the *Abitur*. Students are distributed evenly according to the highest ISEI in the family. However, there are significant differences in terms of educational background. 46.4 percent of parents of students who graduated from comprehensive schools have *Abitur*. This is around ten percentage points lower than in the tracked school system. 28.2 percent of students in comprehensive school have teacher recommendations for *Gymnasium*, while 57.6 percent of students in the tracked school types have recommendations to *Gymnasium*. This illustrates that the average initial performance of students in comprehensive schools is lower compared to the average initial performance of those in tracked schools. However, parents' aspiration that their own child reach the *Abitur* is very similar in both groups. In addition, comprehensive schools seem to be located in areas with higher density of alternative schools nearer students and lower admission conditions for comprehensive school. When looking at the distribution of cases in the federal states, it is noticeable that most of the students in the sample come from North Rhine-Westphalia, followed by Lower Saxony and Hesse. Only in Hesse and Berlin are there more comprehensive school students in the sample than students from tracked schools.

We have already mentioned the relatively high number of missing values. Table 7 in the appendix provides an overview of the descriptive statistics of the analysis sample and of the cases not included in the analysis sample originating from states in the analysis sample. It is notable that the cases not included in the analysis sample have about eight percentage points more students from comprehensive schools, are about thirty percentage points less likely to graduate from high school, have an ISEI on average about ten points lower, and have fewer parents who graduated from high school compared to the cases in the analysis sample. In addition, students are less likely to be recommended for upper secondary education and parents less likely to aspire to the *Abitur*. The dropouts are relatively similar between students in comprehensive and in tracked schools. The differences between the analysis sample and the dropouts suggest attrition bias. Therefore, the results below were also checked with weights for wave 9. However, the results were mostly the same (results not shown). Therefore, we assume that attrition bias does not significantly bias the results.

Table 1: Descriptive statistics of the full sample, the comprehensive school sample, and the sample of tracked schools

	Full sample				CS		Tracked Schools	
	Mean	sd	Min	Max	Mean	sd	Mean	sd
CS	0.126	0.332	0	1				
Abitur	0.673	0.469	0	1	0.609	0.489	0.682	0.466
ISEI (highest in family)	56.734	19.406	11.560	88.960	55.076	19.867	56.973	19.331
Educational background (*Abitur*)	0.554	0.497	0	1	0.464	0.499	0.567	0.496
Recommendation to *Gymnasium*	0.539	0.499	0	1	0.282	0.451	0.576	0.494
Aspiration (*Abitur*)	0.732	0.443	0	1	0.707	0.456	0.736	0.441
Alternative school nearer	18.243	25.467	0	100	30.689	32.903	16.445	23.677
Alternative school nearer (NA)	0.250	0.433	0	1	0.207	0.406	0.256	0.436
Admission condition	2.434	2.198	0	10	1.873	2.114	2.515	2.198
Admission condition (NA)	0.239	0.427	0	1	0.337	0.473	0.225	0.418
Schleswig Holstein	0.083	0.277	0.000	1.000	0.078	0.268	0.084	0.278
Lower Saxony	0.189	0.391	0.000	1.000	0.150	0.358	0.194	0.396
North Rhine Westphalia	0.500	0.500	0.000	1.000	0.433	0.496	0.510	0.500
Hesse	0.107	0.309	0.000	1.000	0.249	0.433	0.087	0.282
Rhineland Palatinate	0.093	0.291	0.000	1.000	0.049	0.217	0.100	0.299
Berlin	0.027	0.163	0.000	1.000	0.041	0.200	0.025	0.158
N	3058				386		2672	

Notes: CS = comprehensive school. Own calculations. Source: NEPS SC4:9.1.1 (2018).

We base our multivariate analysis on kernel matching with an *Epanechnikov* kernel and radius matching. Kernel matching and radius matching show the best results in reducing bias (see table 2). However, to check for robustness (Caliendo & Kopeinig 2008), we also conducted the analysis with nearest neighbor matching (in the following NN) with no replacement using one neighbor (k = 1) and nearest neighbor matching with caliper (in the following NNC) with replacement using three neighbors (k = 3), yielding mostly similar results (results not shown). To improve precision of the estimates from kernel and radius matching, we deployed logistic regression models on the matched sample. All propensity score matching procedures are estimated with Stata 16.1 and the ado `psmatch2` (Leuven & Sianesi 2003).

The assignment models for estimating the propensity scores appear in the appendix (see table 6), as well as in a table with a detailed bias overview by variable for kernel and radius matching (table 8 in the appendix). All assignment models are logistic regressions on the probability of a student attending a comprehensive school. Table 2 reports the mean standardized bias for unmatched data and each propensity score matching method by assignment model. Standardized bias can help with assessing the balancing. It is desirable to reduce bias through matching. Usually, bias around 5 percent is still considered acceptable (Gangl 2015). Each propensity score matching method is able to reduce the bias substantially compared to the unmatched data. However, neither NN matching method consistently reduces the bias to at most 5 percent. For the two NN methods, the mean standardized bias ranges from 5.5 percent to 15.1 percent for NNC, respectively from 5.0 percent to 16.8 percent for NN. Kernel and radius matching perform better, ranging 3.1–5.7 percent for kernel matching and 4.0–4.6 percent for radius matching. Yet, the mean bias of the last model (M3b: recommendation to *Gymnasium*) breaks away for kernel (9.7 percent) and radius (8.5 percent) matching. The number of treated cases for this model is also lower, with eighty-nine and seventy-seven matched observations for kernel and radius matching, respectively. Thus, we must approach the last model with caution.

Common support test (results not shown) expresses no major differences between students in comprehensive and in tracked schools in the distribution of propensity scores. It is apparent that the overlap of students in comprehensive and in tracked schools is mainly present in the area of lower propensity scores. For higher propensity scores, the number of cases for students in tracked schools is relatively low.

Table 2: Treated cases on support and mean standardized bias (MB) by propensity score matching algorithm and specification model

	Unmatched data		Kernel			Radius			NN (k=1)		NNC (k=3)		
	$N_{Di=1}$	MB	$N_{Di=1}$	MB	bw	$N_{Di=1}$	MB	c	$N_{Di=1}$	MB	$N_{Di=1}$	MB	c
General Model	386	27.9	386	3.1	.1	386	4.0	.1	386	6.2	386	9.2	.1
ISEI (Upper Half)	235	28.8	235	5.7	.1	235	4.5	.1	235	11.5	159	15.1	.001
ISEI (Lower Half)	151	29.6	144	4.4	.01	144	4.4	.01	151	6.1	144	8.2	.01
No Recommendation	277	24.8	277	4.3	.1	277	4.6	.1	277	5.0	208	5.5	.001
Recommendation	109	21.8	89	9.7	.01	77	8.5	.001	98	16.8	77	9.2	.001

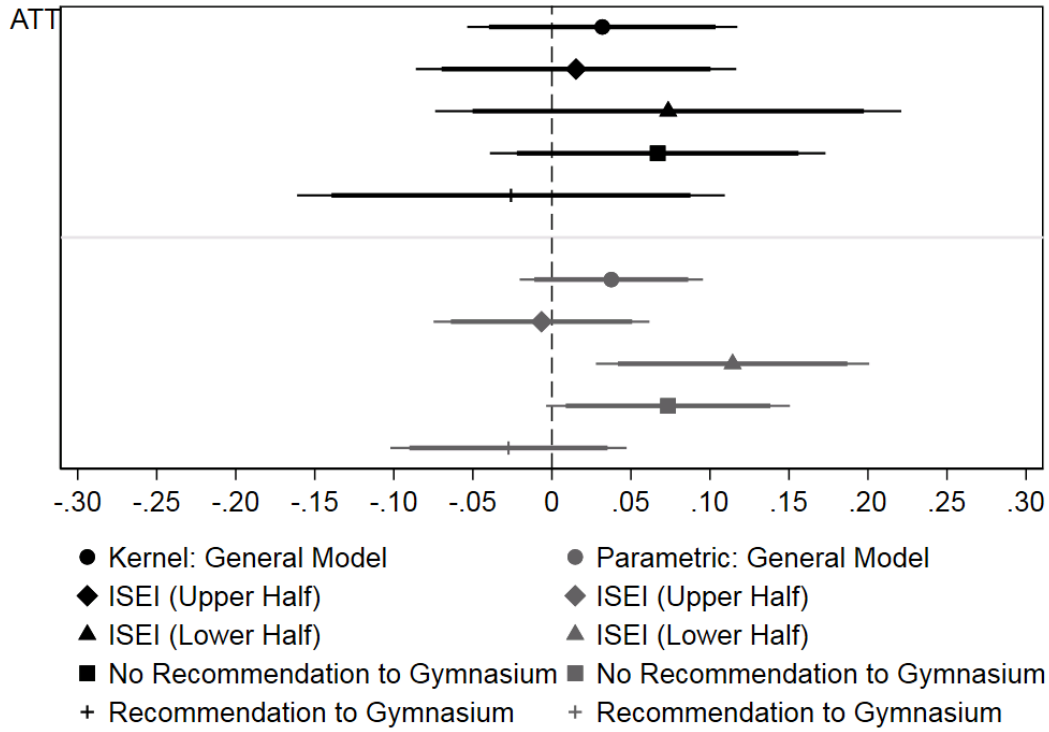
Notes: bw = bandwidth, c = caliper, k = number of neighbors. Own calculations. Source: NEPS SC4:9.1.1 (2018).

The following paragraphs present and discuss the ATT estimates from kernel and radius matching, as well as parametric models based on the obtained propensity scores. The discussion of the results will refer to parametric models. Estimation of ATT employed 250 bootstrap repetitions. Additionally, the calculation of clustered standard errors, using schools as clusters, reflects the multilevel structure of the data. All ATT estimators are plotted with a 90 and 95 percent confidence level. We formulated various models to test the hypotheses. The general model tests hypothesis 1 while separate models for the upper and lower halves of ISEI test hypothesis 2 and models for students with and without upper secondary recommendations test hypothesis 3. Descriptive statistics of the models for the hypothesis 2 and 3 samples are in the appendix (table 4 and 5). Figure 1 shows all results for kernel matching and its corresponding parametric models while figure 2 displays all results for radius matching and its parametric models. The general model analyzes how comprehensive schools change the probability to obtain the *Abitur* for all students at a comprehensive school. While the parametric models of both matching algorithms show a small positive increase of about four percentage points in the probability of obtaining the *Abitur* in a comprehensive school, the change is not statistically significant. Overall, comprehensive schools do not change students' probability of obtaining the *Abitur* significantly. The general model confirms hypothesis 1.

However, it is not possible with the general model to test the SES-specific hypotheses about the positive association of comprehensive schools with low SES students obtaining the *Abitur*. To examine SES-specific associations of comprehensive schools, we estimate separate models for higher and lower SES students. The model for higher SES students indicates no significant association of comprehensive schools with *Abitur* attainment (the point estimate is close to zero). However, for students of lower SES, the parametric model shows a significant positive association of comprehensive schools on the probability of obtaining the *Abitur*, which increases by about twelve percentage points. In particular, students of low SES benefit from comprehensive schools in achieving the *Abitur*, confirming hypothesis 2. It also shows that students with higher SES are not at a disadvantage in achieving the *Abitur*.

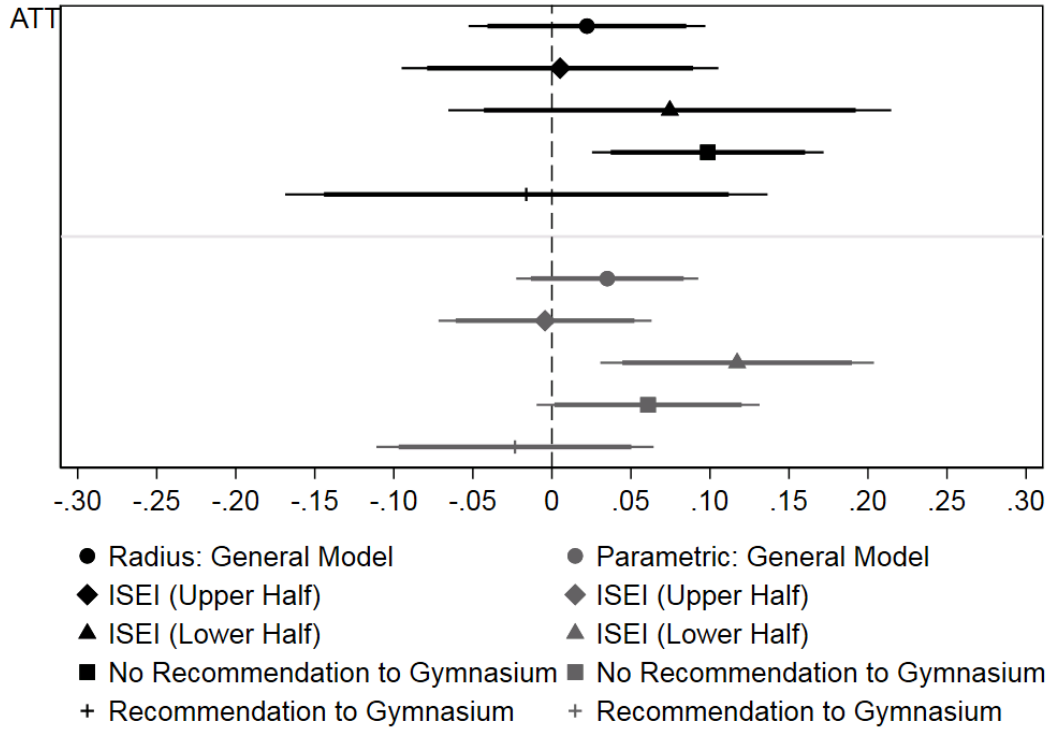
To test the ability-specific hypothesis that students with weaker initial performance benefit particularly from attending comprehensive schools in achieving the *Abitur*, we estimated separate models for students with and without teacher recommendations for the *Gymnasium*. The model for students with a low initial performance, i.e. students without a recommendation for *Gymnasium*, displays a significant positive change in the probability of reaching the *Abitur*. The point estimate indicates a change of about six percentage points but is significant only at the 10-percent level. Still, the results support hypothesis 3 and show that, through comprehensive schools, students without a recommendation for the *Gymnasium* have a higher probability of achieving the *Abitur* nonetheless. As referred to above, the models for students with higher initial performance, i.e. those with a recommendation for *Gymnasium*, have relatively few cases; in addition, the matching algorithms have a high standard bias of more than five percentage points. Thus, we do not discuss the results further and show them only for completeness.

Figure 1: ATT estimates from kernel matching of the change in *Abitur* completion for comprehensive school students; estimates by model



Notes: ATT estimator with 90 and 95 percent confidence level. Clustered standard errors with schools as clusters. Own calculations. Source: NEPS SC4:9.1.1 (2018).

Figure 2: ATT estimates from radius matching of the change in *Abitur* completion for comprehensive school students; estimates by model



Notes: ATT estimator with 90 and 95 percent confidence level. Clustered standard errors with schools as clusters. Own calculations. Source: NEPS SC4:9.1.1 (2018).

2.6 Conclusion and limitations

Many consider the timing of tracking important in the formation and persistence of educational inequalities. Countries with early tracking educational systems usually display higher inequalities in respect to performance development and educational decision-making. In countries with late tracking systems, these inequalities are significantly reduced. The aim of this study is to examine the association of comprehensive schools, a late tracking alternative to the early tracking school system in Germany, with upper secondary attainment. Therefore, we analyze NEPS data with propensity score matching.

Regarding the general association of comprehensive schools on the probability of obtaining the *Abitur*, hypothesis 1 assumes no association, particularly because different theoretical arguments go in different directions and empirical findings vary as to the effect of late tracking depending on students' performance. The results indicate no change in the probability of obtaining the *Abitur* for students in comprehensive schools generally. However, there are group-specific hypotheses relating comprehensive schools to obtaining the *Abitur*. For example, it is possible to hypothesize less uncertainty about performance development and consequently lower association between a student's SES and educational attainment. The separate models for higher and lower SES students find no influence of comprehensive schools on high SES students' likelihood of obtaining an *Abitur* but a significant increase in the probability of graduating with an *Abitur* for students of low SES, confirming hypothesis 2. Because later tracking affects performance trajectories differently for high- and low-performing students, we also assume comprehensive schools to have a positive impact on the likelihood of achieving the *Abitur* for students without recommendations for *Gymnasium*. Results show an increase in the probability of obtaining the *Abitur* for students without a recommendation for upper secondary education, confirming hypothesis 3. The creation of the comprehensive schools in Germany reflected reform goals for a more inclusive and less socially stratified educational system, which the results of this analysis apparently support. Taken together, the results show that, while comprehensive schools do not increase the probabilities of achieving the *Abitur* for all students in attendance, they do increase those probabilities especially for students with low SES and for students without a recommendation for upper secondary education. Thus, comprehensive schools can contribute to a reduction of social inequalities in the education system.

Several limitations to this study require mention. The first concern is about self-selection into comprehensive schools. Parents can decide freely whether they want to send their child to a comprehensive school. Matching cannot resolve this issue completely. However, because of

balancing and common support, it is possible to draw conclusions about students attending comprehensive schools. It is not possible to draw conclusions for all students because the self-selected students attending comprehensive schools may be systematically different from the rest of the student population. Therefore, it is unclear whether the student population as a whole also benefit from comprehensive schools through increased probability of obtaining the *Abitur*.

Second, they are some limitations that arise from data quality. Some pretreatment measurements are not at all or not sufficiently available, most problematically in the lack of primary school performance. Primary school performance, an important predictor for school and track choice (Maaz et al. 2008), is not measured. The teacher recommendation in SC4 serves as a good alternative measurement for primary school performance. Nevertheless, teacher recommendation data are only measured through the parental survey. Because of high parental non-response, the use of teacher recommendation in the assignment models discards many cases due to missing values. One could argue for using SC3 (which started with students in grade 5) instead of SC4 and using school performance at the beginning of secondary education. However, the sample of students is smaller and displays similar problems of non-response. In terms of non-response, however, attrition bias does not appear to be a major problem, since analyses with weights come to very similar results. Another issue with the data is that it is only possible to represent the opportunity structure for school choice in part. The proximity of a school is important to predicting whether a student will visit the school or not. One could also partially address this problem by including an indicator for the percentage of students with an alternative school offering the same educational pathway and nearer than their current school. While this indicator is a subjective approximation by the principal, an individual indicator of proximity and the objective number of alternative schools for each student within a given radius would be much more accurate. The third limitation concerns statistical power. While comprehensive schools are oversampled in the NEPS SC4 (IEA Data Processing and Research Center 2010), in combination with non-responses, it is hardly sufficient for in-depth subgroup analysis and analysis of different subgroup associations.

Further research should investigate comprehensive schools more closely and monitor and analyze the implementation of schools with three educational programs because this relatively new school type is similar to comprehensive schools. Further, research should focus on the mechanisms behind the effects of timing of tracking. The influence on the decision-making process for SES groups and performance groups should be of major concern. The quantity of studies, especially on the influence of tracking on educational decision-making, is limited and more

empirical testing of theoretical predictions under different settings is crucial for understanding how timing of tracking affects educational inequalities. Research should also examine the variation in effects of timing of tracking on performance development for various performance groups and of that variation's conditions and consequences (Roller & Steinberg 2020).

2.A Appendix Chapter 2

Table 3: Distribution of students by school type and federal state

Federal State	Tracked schools	*Haupt-schule*	*Real-schule*	*Gymnasium*	CS
Schleswig Holstein	**225**	**46**	**70**	**109**	**30**
Hamburg	30	2	5	23	26
Lower Saxony	**519**	**98**	**146**	**275**	**58**
Bremen	35	0	7	28	0
North Rhine Westphalia	**1363**	**215**	**349**	**799**	**167**
Hesse	**232**	**17**	**58**	**157**	**96**
Rhineland Palatinate	**266**	**32**	**54**	**180**	**19**
Baden-Wuerttemberg	1018	172	302	544	0
Bavaria	551	97	155	299	0
Saarland	14	3	5	6	0
Berlin	**68**	**5**	**19**	**44**	**16**
Brandenburg	60	2	0	58	19
Mecklenburg Western Pomerania	58	3	2	53	21
Saxony	113	0	0	113	0
Saxony-Anhalt	38	3	0	35	0
Thuringia	49	0	0	49	5
Total	4639	695	1172	2772	457

Notes: CS = comprehensive school. Tracked schools consist of *Hauptschule*, *Realschule*, and *Gymnasium*. Bold print indicates the federal states in the analysis sample. Source: NEPS SC4:9.1.1 (2018).

Table 4: Descriptive statistics of comprehensive school sample and the sample of other school types by recommendation to *Gymnasium*

	CS				Tracked schools			
	Recommendation to *Gymnasium*		No Recommen-dation to *Gymna-sium*		Recommendation to *Gymnasium*		No Recommen-dation to *Gym-nasium*	
	Mean	Sd	Mean	sd	Mean	sd	Mean	Sd
Abitur	0.881	0.326	0.502	0.501	0.912	0.284	0.370	0.483
ISEI (highest in family)	65.102	17.598	51.130	19.341	62.883	17.472	48.958	18.843
Educational background (*Abitur*)	0.688	0.465	0.375	0.485	0.717	0.451	0.365	0.482
Aspiration (*Abitur*)	0.954	0.210	0.610	0.489	0.956	0.206	0.437	0.496
Alternative school nearer	44.972	36.015	25.069	29.838	16.559	24.435	16.292	22.619
Alternative school nearer (NA)	0.110	0.314	0.245	0.431	0.311	0.463	0.182	0.386
Admission condition	2.239	2.063	1.729	2.120	2.635	2.217	2.352	2.163
Admission condition (NA)	0.229	0.422	0.379	0.486	0.270	0.444	0.163	0.370
N	109		277		1538		1134	

Notes: CS = comprehensive school. Own calculations. Source: NEPS SC4:9.1.1 (2018).

Table 5: Descriptive statistics of comprehensive school sample and the sample of other school types by half of ISEI

	CS				Tracked schools			
	Upper Half of ISEI		Lower Half of ISEI		Upper Half of ISEI		Lower Half of ISEI	
	Mean	Sd	Mean	sd	Mean	sd	Mean	sd
Abitur	0.681	0.467	0.497	0.502	0.794	0.405	0.490	0.500
ISEI (highest in family)	68.643	10.828	33.961	9.644	69.294	10.930	35.939	10.293
Educational background (*Abitur*)	0.638	0.482	0.192	0.395	0.744	0.437	0.266	0.442
Recommendation to *Gymnasium*	0.374	0.485	0.139	0.347	0.684	0.465	0.390	0.488
Aspiration (*Abitur*)	0.770	0.422	0.609	0.490	0.837	0.370	0.563	0.496
Alternative school nearer	33.191	33.629	26.795	31.455	17.039	24.727	15.433	21.743
Alternative school nearer (NA)	0.196	0.398	0.225	0.419	0.275	0.447	0.224	0.417
Admission condition	1.970	2.147	1.722	2.060	2.571	2.213	2.418	2.170
Admission condition (NA)	0.319	0.467	0.364	0.483	0.247	0.432	0.186	0.390
N	235		151		1685		987	

Notes: CS = comprehensive school. Own calculations. Source: NEPS SC4:9.1.1 (2018).

Table 6: Assignment models 1, 2, and 3 (logistic regression on the probability of a student attending a comprehensive school, model 2 by SES, and model 3 by recommendation status, logits with standard errors in parentheses)

	M1: General Model	M2a: ISEI (Upper Half)	M2b: ISEI (Lower Half)	M3a: No Recommendation to *Gymnasium*	M3b: Recommendation to *Gymnasium*
ISEI (highest in family)	0.015*	0.026	-0.003	0.014	0.059
	(0.01)	(0.02)	(0.02)	(0.01)	(0.04)
Educational background (*Abitur*)	-0.145	-0.139	-0.220	-0.181	-0.528
	(0.27)	(0.35)	(0.47)	(0.29)	(1.30)
Recommendation to *Gymnasium*	-1.520*	-1.372	-1.575		
	(0.67)	(1.26)	(1.33)		
Aspiration (*Abitur*)	1.714***	2.785*	1.682*	1.653***	3.268
	(0.42)	(1.33)	(0.74)	(0.45)	(2.21)
Alternative school nearer	0.029***	0.031***	0.028***	0.027***	0.032***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Alternative school nearer (NA)	-0.030	-0.119	0.262	0.257	-0.512
	(0.20)	(0.25)	(0.31)	(0.24)	(0.40)
Admission condition	-0.128**	-0.164**	-0.086	-0.112*	-0.153*
	(0.04)	(0.05)	(0.06)	(0.05)	(0.07)
Admission condition (NA)	0.850***	0.739**	1.044***	1.298***	0.057
	(0.19)	(0.25)	(0.31)	(0.24)	(0.36)
Aspiration X ISEI	-0.017	-0.032	-0.020	-0.015	-0.054
	(0.01)	(0.02)	(0.02)	(0.01)	(0.04)
ISEI X Recommendation	0.008	0.005	0.007		
	(0.01)	(0.02)	(0.03)		
Aspiration X Recommendation	-0.720	-0.547	-0.753		
	(0.53)	(0.70)	(0.84)		
Educational background X Recommendation	-0.134	-0.210	0.042		
	(0.35)	(0.42)	(0.64)		
Aspiration X Educational background	-0.178	-0.231	-0.071	-0.184	0.062
	(0.34)	(0.43)	(0.57)	(0.36)	(1.33)
Constant	-3.216***	-3.902***	-2.653***	-3.285***	-6.849**
	(0.36)	(1.10)	(0.58)	(0.40)	(2.21)
Pseudo-R^2	0.193	0.189	0.224	0.192	0.182
N	3058	1920	1138	1411	1647

Notes: Federal state indicators included (not shown). * p<0.05, ** p<0.01, *** p<0.001. Own calculations. Source: NEPS SC4:9.1.1 (2018).

Table 7: Descriptive statistics of the full sample, the comprehensive school sample, the sample of tracked schools, and the dropouts

	Full sample		Dropouts			CS		Dropouts (CS)			Tracked schools		Dropouts (Tracked schools)		
	Mean	sd	Mean	sd	Obs.	Mean	sd	Mean	sd	Obs.	Mean	sd	Mean	sd	Obs.
CS	0.126	0.332	0.207	0.405	5260										
Abitur	0.673	0.469	0.376	0.484	5603	0.609	0.489	0.397	0.490	872	0.682	0.466	0.405	0.491	4172
ISEI (highest in family)	56.734	19.406	45.855	20.357	4757	55.076	19.867	47.420	19.760	837	56.973	19.331	46.653	20.370	3330
Educational background (*Abitur*)	0.554	0.497	0.187	0.390	4772	0.464	0.499	0.211	0.408	874	0.567	0.496	0.183	0.387	3353
Recommendation to *Gymnasium*	0.539	0.499	0.349	0.477	981	0.282	0.451	0.359	0.481	273	0.576	0.494	0.373	0.484	627
Aspiration (*Abitur*)	0.732	0.443	0.539	0.499	4594	0.707	0.456	0.601	0.490	892	0.736	0.441	0.527	0.499	3533
Alternative school nearer	18.243	25.467	16.792	24.928	4454	30.689	32.903	27.981	29.921	622	16.445	23.677	15.588	23.361	3139
Alternative school nearer (NA)	0.250	0.433	0.215	0.411	4454	0.207	0.406	0.193	0.395	622	0.256	0.436	0.241	0.427	3139
Admission condition	2.434	2.198	2.264	2.236	4454	1.873	2.114	1.772	2.196	622	2.515	2.198	2.257	2.120	3139
Admission condition (NA)	0.239	0.427	0.241	0.428	4454	0.337	0.473	0.391	0.488	622	0.225	0.418	0.212	0.409	3139
N	3058		6042			386		1088			2672		4172		

Notes: CS = comprehensive schools. Own calculations, Source: NEPS SC4:9.1.1 (2018).

Table 8: Standardized bias by propensity score matching algorithm and assignment model

	M1: General Model		M2a: ISEI (Upper Half)		M2b: ISEI (Lower Half)		M3a: No Recommendation to *Gymnasium*		M3b: Recommendation to *Gymnasium*	
	Radius	Kernel	Radius	Kernel	Radius	Kernel	Radius	Kernel	Radius	Kernel
ISEI (highest in family)	1.3	3.0	5.8	8.8	3.4	3.1	-0.3	0.1	3.6	14.9
Educational background (*Abitur*)	-2.0	-0.1	0.8	4.4	-6.6	-7.6	-2.6	-1.0	0.2	17.3
Recommendation for *Gymnasium*	-7.9	-4.8	-2.9	0.5	-0.6	-1.1				
Aspiration (*Abitur*)	1.3	2.3	5.4	8.4	-14.0	-13.2	1.1	1.7	-6.6	-6.8
Alternative school nearer	5.8	3.8	12.4	11.5	-4.6	-3.9	-6.1	-7.9	-0.4	3.8
Alternative school nearer (NA)	1.0	1.6	-1.0	-0.9	-2.3	-2.3	5.0	5.6	-2.2	-6.7
Admission condition	-1.4	0.9	0.5	3.7	0.6	1.2	-2.0	-1.0	0.2	2.3
Admission condition (NA)	1.7	0.6	-1.9	-3.8	-1.8	-2.2	2.5	0.6	-10.8	2.4
Schleswig Holstein	-4.5	-4.3	-5.8	-8.0	-6.5	-5.7	-7.6	-7.4	2.0	-1.6
Lower Saxony	0.3	0.6	3.9	4.2	-0.5	1.1	-0.3	1.2	-32.9	-27.9
Hesse	9.9	8.5	12.2	11.6	15.8	15.0	19.3	17.1	27.1	12.9
Rhineland Palatinate	-5.8	-5.5	-6.3	-6.5	-1.8	-1.0	-8.6	-8.5	-16.9	-13.1
Berlin	7.6	7.2	7.7	8.3	1.1	0.8	4.8	5.1	-13.8	-1.4
Aspiration X ISEI	1.0	2.7	6.4	9.6	-10.2	-9.4	-0.1	1.1	0.8	7.4
ISEI X Recommendation	-6.1	-3.5	-2.0	1.3	-0.3	-0.6				
Aspiration X Recommendation	-7.6	-4.5	-2.2	1.1	-0.5	-0.7				
Educational background X Recommendation	-4.7	-2.6	3.0	2.7	1.4	1.4				
Aspiration X Educational background	-1.9	0.3	-0.7	7.1	-8.2	-9.6	-3.7	-1.5	1.7	16.7
Mean Bias	4.0	3.1	4.5	5.7	4.4	4.4	4.6	4.3	8.5	9.7

Notes: Own calculations, Source: NEPS SC4:9.1.1 (2018).

3 Changes in the timing of tracking and its effects on educational inequalities: A natural experiment in Germany

Abstract

Socioeconomic status (SES) is a predictor for access to educational tracks. Nevertheless, the relationship between a family's SES and access to education varies between countries. A debated explanation for this variation is the timing of tracking. Tracking is the placement of students into two or more educational paths. Largely, existing findings show a decline in educational inequality when tracking is postponed. While research established that postponing tracking has effects on several inequality dimensions, such as income and performance, the overall effect of timing of tracking on the relationship between SES and access to educational tracks is often neglected. This study aims to examine the causal effects of tracking on academic track attendance of students in general, as well as low-SES students and high-performing low-SES students. This work utilizes the quasi-experimental structure of an educational reform in Lower Saxony. The reform preponed tracking from grade seven to grade five by two years. By using difference-in-difference and difference-in-difference-in-difference estimators with PISA-E and IQB-LV data, we aim to estimate the causal effect of the reform on academic track attendance. The results show no significant effect of preponing tracking on academic track attendance for students in general, low SES students, and high performing low SES students.

3.1 Introduction

Inequality within the educational system varies between countries. An explanation for the variation is the timing of tracking (Le Donné 2014). Countries with a late tracking education system, which is a late differentiation between two or more educational tracks (e.g. an academic orientated track and a vocational orientated track), tend to have fewer inequalities within their education system concerning performance development and educational decision-making (Hanushek & Wössmann 2011; Le Donné 2014). Theoretically, the timing of tracking can influence educational inequalities in two ways. First, it can influence the performance development in school through different mechanisms in both positive and negative ways (Figlio & Page 2002). Empirical studies mostly indicate a positive effect of late tracking on performance (Cordero et al. 2018). One explanation is that late tracking promotes interaction between high- and low-performing students within classes (Piopiunik 2014). However, high- and low-performing students seem to be differently affected by the timing of tracking. Particular high-performing students benefit from early tracking (Roller & Steinberg 2020). Secondly, the timing of tracking can affect the decision-making process for secondary education (Berger & Combet 2017). Especially, a positive effect on low socioeconomic status (SES) students seems plausible. Because the earlier tracking takes place, the more SES background determines educational decisions, as uncertainty about a student's performance development increases. Studies that focus on the effect of timing of tracking on the decision-making process are less frequent than on performance. They yielded mixed results on the effects of tracking (Combet 2019).

The present study examines the effect of an educational reform in Lower Saxony on average attendance to the academic track (*Gymnasium*) for students in general and for low SES students and low SES students with above average performance. In 2004, a reform preponed tracking by two years, from grade seven to grade five. The reform is a natural experiment, with a distinct pre and post-reform period and an "as good as random" assignment into treatment and control groups. The analysis uses PISA-E and IQB-LV data. The causal effect of the reform on *Gymnasium* attendance in Lower Saxony is estimated with difference-in-difference and difference-in-difference-in-difference estimators.

3.2 Timing of tracking and its effects on educational inequalities

The access to different levels of education strongly depends on the SES background of a student (Breen & Jonsson 2005). Comparative research repeatedly finds varying levels of the relationship between social background and educational outcome, depending on several arrangements of the educational system (Le Donné 2014; Pfeffer 2008). This link between the educational system and educational inequality is an important aspect of social sciences research, thereby offering major policy implications.

Prior research has identified the timing of tracking as one important difference between educational systems and a possible explanation for the variance between countries (Pfeffer 2008). Still, the effects of timing of tracking on performance are theoretically ambiguous. While early tracking leads to more academically homogeneous classes, late tracking promotes more heterogeneous classes. On the one hand, it is argued that homogeneous classes increase the efficiency of learning by allowing teachers to apply specific pedagogical methods to each performance group, thus improving student performance at all educational levels. On the other hand, research has proposed that predominately the high SES students benefit in homogeneous classes, while simultaneously leaving the low SES students behind (Figlio & Page 2002; Piopiunik 2014). Late tracking and heterogeneous classes, in contrast, increases the contact between high- and low-performing students within classes. Students with high performance could help lower-performing classmates, thereby increasing their classmates' performance and consolidating their knowledge. Thus, late tracking could have a positive effect on students' performance for all achievement groups (Figlio & Page 2002; Piopiunik 2014).

Previous findings in this research area show mostly positive effects from heterogeneous classes on performance development in general and also for specific fields (e.g. reading, mathematics, and science), thereby supporting the positive effect of late tracking on academic performance (Bygren 2016; Hanushek & Wössmann 2006; Horn 2013; Jakubowski et al. 2016; Korthals & Dronkers 2016; Piopiunik 2014). However, results for Germany show that especially high-performing students benefit from early tracking (Roller & Steinberg 2020), whereas positive effects of late tracking are mainly caused by lower-performing students (Matthewes 2021). These findings suggests a variance in effects for different performance groups and not a uniformly positive effect of late tracking.

In addition to performance, the timing of tracking can also influence educational decision-making. The earlier an educational system tracks its students, the higher is the uncertainty regarding student's future abilities. For the process of educational decision-making, this suggests a higher correlation between SES background and the assessed success probability of completing a certain educational path in early-tracking educational systems. As late tracking reduces the uncertainty of school performance and its future development, it allows for a better-informed decision about the investment in education. By postponing the educational decision, information about one's performances become more reliable, which in turn increases the assessed probabilities to successfully obtain a high level of education within the educational decision making process (Bauer & Riphahn 2006; Berger & Combet 2017). Thus, late tracking within the educational system should reduce the association between one's SES background and educational decision-making (Dustmann 2004). Early tracking, in turn, reduces the available information on performance development and increases uncertainty in decision-making. This increases the correlation between SES background and the assessed success probability of completing a certain educational path and leads to greater inequality in educational decisions. Because of the status maintenance motive, students with higher SES will nevertheless continue to choose a more demanding educational career, and low-SES students in particular will be affected by the increased uncertainty. However, the reduction of uncertainty about performance and its future development suggests that it is not the shift in timing of tracking as such that affects educational inequalities. Rather, tracking must be shifted for a certain amount of time to reduce uncertainty enough for the educational decision-making process to change significantly.

Previous research shows that high performing students with low SES in particular are influenced by later tracking in their educational decisions (Berger & Combet 2017; Meghir & Palme 2005). Research also shows greater inequality in early tracking systems (Bauer & Riphahn 2006). However, other findings cannot confirm this result (Combet 2019). While (Combet 2019) finds no significant differences in educational attainment between early and late tracking, when the time difference between early and late tracking is small, other research finds a significant effect of late tracking on educational attainment, when the time difference between early and late tracking is bigger (Berger & Combet 2017; Meghir & Palme 2005). This suggests that the uncertainty in performance development must be reduced by a certain amount. However, it is unclear how long tracking must be postponed to sufficiently reduce uncertainty to decrease inequalities in educational decisions.

Based on the theoretical considerations and the empirical evidence presented above, the three main hypotheses are as follows:

H1: If tracking occurs earlier in the educational career of students, then the likelihood for students for a more demanding track decreases.

The effect of *H1* may result from a different decision-making process of lower SES students in early tracking settings compared to late tracking settings, which causes a compositional change in students attending the *Gymnasium*. For students with higher SES, there is no change in educational decision-making. This difference in behavior between low and high SES students leads to the formulation of *H2*. Although there is evidence that later tracking has an effect on inequality only after a certain number of years, we still adopt this standard hypothesis in the sociology of education because it is still unclear how many years that is.

H2: If tracking takes place earlier in the educational career of students, then the likelihood for students with low SES for a more demanding track decreases.

It is also hypothesized that earlier tracking will cause high-performing students with low SES, in particular, to adjust their educational decisions.

H3: If tracking takes place earlier in the educational career of students, then the likelihood for high-performing students with low SES for a more demanding track decreases.

3.3 The German education system and educational reform in Lower Saxony

In Germany, children begin to attend primary school (*Grundschule*) around the age of six. Primary education covers the first four grades.[6] Around the age of ten, the first transition usually takes place, where students split into two or three different secondary educational school tracks. The number of secondary school tracks differs between the federal states of Germany. The lowest secondary track is the basic school (*Hauptschule*) offering a certificate after grade nine or ten. After grade ten, students get the certificate from the middle school (*Realschule*). Some federal states are replacing or have replaced the basic and middle school with a multitrack combination, called schools with two educational programs (*Schularten mit zwei Bildungsgängen*), which offers both certificates. Additionally, some states introduced school types in which students can obtain all three school-leaving certificates. They are called schools with

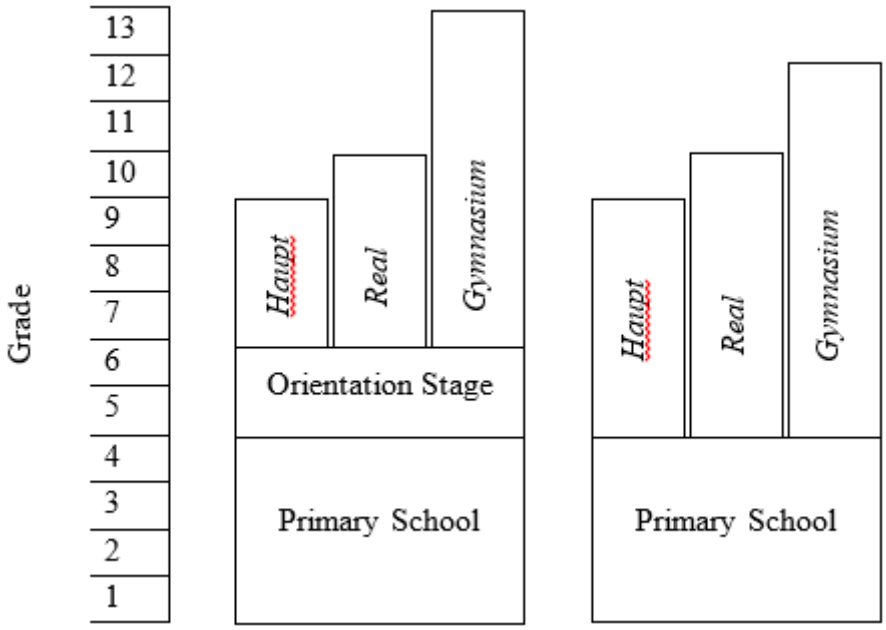
[6] Only in Berlin and Brandenburg, the *Grundschule* covers six years (Eckhardt 2017).

three educational programs (*Schularten mit drei Bildungsgängen*) (Henninges et al. 2019). While *Hauptschule*, *Realschule*, and schools with two educational programs are vocational tracks, leading into apprenticeship, the academic track is called *Gymnasium*. After grade twelve or thirteen, it offers the higher education entrance qualification (*Abitur*), which is the most common way to enter university in Germany.[7]

The German education system is highly stratifying compared to other countries, with a strong connection from SES background to educational qualification and little educational intergenerational mobility. Early tracking is considered as one important factor in the formation of these strong educational inequalities in Germany (Hanushek & Wössmann 2006; Maaz et al. 2008; Pfeffer 2008). However, each federal state in Germany has sovereignty over its educational system. Thus, precise arrangements of the educational system, such as timing of tracking, differ between most German states. One such peculiarity of the secondary education system was the orientation stage (*Orientierungsstufe*, OS) in Lower Saxony. Implemented in 1973 and abolished in 2004, the OS was structured as a two-year comprehensive school in secondary education between the comprehensive primary education and the tracked secondary education (Roller & Steinberg 2020). Figure 3 illustrates the position of the OS within the education system. The goal of this school was to support the student's choice of the right track for given academic performance at the end of the grade six (Lange & Werder 2017; Schuchart 2003). The class composition between the elementary school and the OS stayed the same (Roller & Steinberg 2020). Unlike elementary schools, OS uses ability grouping, where students are placed in ability-specific courses for some subjects. However, switching between courses was possible and not unusual (Schuchart 2003). Thus, the OS postponed tracking for two years from grade five to grade seven, especially for high performing students. The OS was administratively independent, and the teachers at the OS came from all subsequent tracked school types (Lange & Werder 2017).

[7] The *Abitur* was for some time the only possibility to enter university. In the 1960s educational reforms introduced other possibilities to enter university, e.g. through professional qualification. However, these new ways to enter university were limited in their access to university, unlike the *Abitur*, which allows a full access (Schindler 2017). While the status of the *Abitur* as the only possibility to enter university changed, it is still the most usual way for enrollment into university (Müller et al. 2011).

Figure 3: Lower Saxony's school system before (left) and after the reform (right)



Notes: *Haupt* means *Hauptschule* (basic school), and *Real* stands for *Realschule* (middle school).

For the implementation of the OS in the seventies, previous research shows a positive effect on educational performance for students with a low education background, but an opposite effect for students with a high education background (Lange & Werder 2017). This suggests a reduction in the relation between SES and the level of education. Nevertheless, just before the abolishment of the OS, Lower Saxony once more displayed high levels of inequality in the chance of visiting more demanding school tracks for students in tracked secondary education (Schuchart 2003). The educational reform in 2004, which preponed tracking again, led to a positive effect on test scores for high-performance students and an inverted effect for low-performance students (Roller & Steinberg 2020). Similar results are also reported for the orientation stage in Hesse (Mühlenweg 2008). Furthermore, the positive effects on performance are greater for students with a high education background (Roller & Steinberg 2020). Though this indicates effects on educational performance, there is no insight on the effect of preponing tracking on the educational decision for the *Gymnasium* for all students and by different SES backgrounds in Lower Saxony.

3.4 Data, Analytical strategy, Method, and Variables

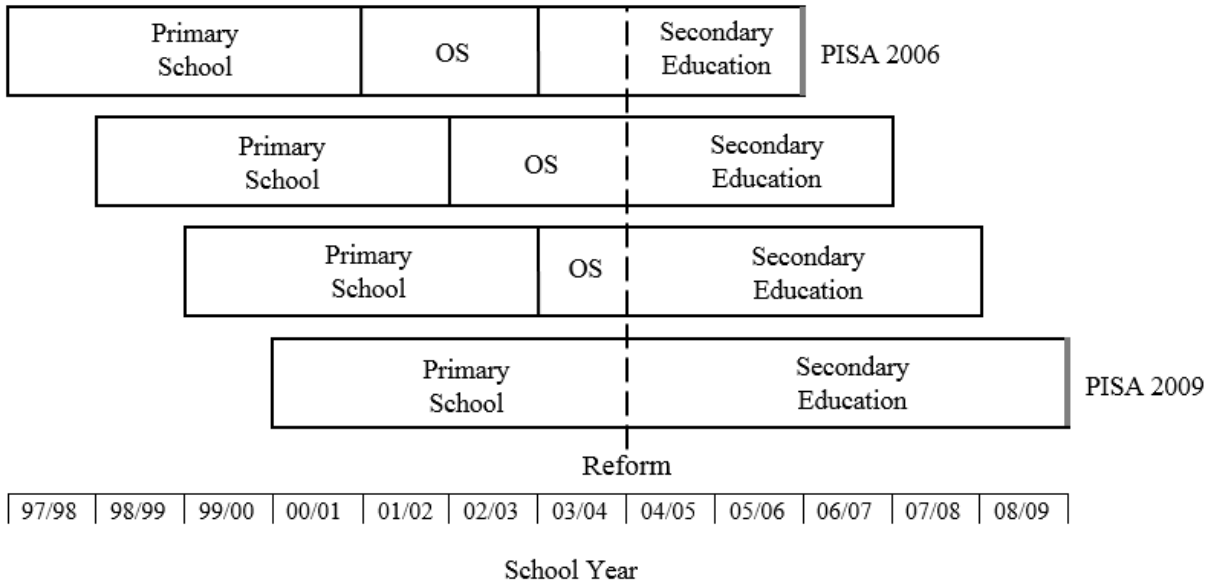
In order to investigate the hypotheses, this study will use data from the Program for International Student Assessment[8] (PISA) by the Organization for Economic Co-operation and Development (OECD). The PISA survey has been conducted every three years in several countries since its inception in 2000. The survey gathers information about the competencies of students in reading, mathematics, and science, as well as data on attitudes, demographics, social origin, and other information. In Germany, the samples consist of 15-year-old students, who are usually in the 9th grade. In addition, PISA conducts interviews with parents, teachers, and school management for other relevant information (Jude & Klieme 2010).

Figure 4 provides an overview of treated (post-reform) and untreated (pre-reform) student cohorts in Lower Saxony. For the analysis, the PISA waves of 2006 and 2009 are of special interest. Students participating in PISA 2006 were the last student cohort surveyed with PISA unaffected by the reform in 2004. These students were in the 7th grade when the reform was implemented. In contrast, students surveyed in PISA 2009 were the first cohort not affected by the OS. These students were in the 4th grade while the OS was abolished.

The focus of this research lies on the federal state level. Therefore, it is necessary to use the extended data PISA-E 2000 (Baumert et al. 2009), 2003 (Prenzel et al. 2007) and 2006 (Prenzel et al. 2010) and a replacement survey for the terminated PISA-E by the *Institut für Qualitätssicherung im Bildungswesen* (IQB) for 2009 and 2012, called IQB National Assessment Study (IQB-LV) 2008/2009 (Köller et al. 2011; Sachse et al. 2012) and 2012 (Lenski et al. 2016; Pant et al. 2015). Otherwise, the number of observations would be too small for analysis on federal state level. PISA-E and IQB-LV are specifically designed to examine federal states more closely.

[8] PISA 2000 was designed in Germany as a national research program by the German PISA Consortium (Juergen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, Manfred Weiß). It was lead-managed by Professor Dr. Juergen Baumert, Max-Planck Institute for Human Development, Berlin. Primary research results have been published, e.g., in Baumert et al. (2001); Baumert et al. (2002, 2003). Survey questionnaires have been documented in Kunter et al. (2002). We thank the German PISA Consortium and the Research Data Center (*Forschungsdatenzentrum*, FDZ) in Berlin for granting permission to conduct this secondary analysis and for their support.

Figure 4: Treatment over student cohorts in Lower Saxony



Notes: Figure adapted from (Roller & Steinberg 2020).

To investigate the relationship between educational inequality and distinct educational system arrangements like timing of the tracking, research often relies on reforms within an educational system. This allows drawing causal inference on how educational systems affect educational inequalities (for international overviews see: (Cordero et al. 2018; Hanushek & Wössmann 2011; Raudenbush & Eschmann 2015; Webbink 2005)). The abolishment of the orientation stage in 2004 can be understood as a natural experiment. The treatment is the preponing of tracking in Lower Saxony. Other German states are not part of this reform and stand as the control group. The selection criteria for the federal states in the control group is that they do not have a similar educational reform as Lower Saxony in the period under investigation. In 2005, Bremen also had a major educational reform, which abolished its OS. Furthermore, Bremen also changed the binding teacher recommendation for secondary school, to a non-binding recommendation. This means that parents are free to decide on the secondary school career of their children. Additionally, Bremen introduced a new school type, the *Sekundarschule* (later named *Oberschule*), which is a school with three educational programs (Büchler 2016). Therefore, Bremen needs to be excluded from the analysis. Assignment to the treatment or control group depends on the place of residence and year of birth, which is assumed to be "as good as random".

For the identification of the causal effect of reforms, difference-in-difference (DiD) regression estimators are used. DiD is often used as a tool for policy evaluation (Athey & Imbens 2017),

e.g. for reforms in the education system to study the effect of a specific reform on educational achievement (Lange & Werder 2017). DiD requires a natural experiment with at least two time periods, a treatment and a control group. The assignment to either the treatment or the control group must essentially be random. It is possible to use DiD not only with panel data, but also with aggregate level data or repeated cross-sectional data, like PISA-E and IQB-LV. So it is not necessary to have repeated measures of the same individuals (Angrist & Pischke 2009; Gangl 2010; Wooldridge 2010). The model is represented by the following equation:

$$y_i = \beta_0 + \beta_1 S_i + \beta_2 T_i + \beta_3 (S_i * T_i) + \beta_C X_i + u_i \quad (1)$$

Equation 1 gives the DiD estimator for the effect of the reform for individual *i* on enrollment into *Gymnasium* y, which is the outcome of interest. The dichotomous variable S_i is the state indicator and equals one for the treatment group, in this case Lower Saxony, and is zero for the control group. This variable accounts for possible differences between both groups in the model. T_i is a dichotomous variable for the time, which equals one for the post-reform period and zero otherwise. The time dummy captures possible changes over time, which could lead to a change in y, even without the reform. The interaction term between S_i and T_i equals one for Lower Saxony in the post-reform period. β_3 is the coefficient of this interaction and it is the effect to answer *H1*. It indicates the effect of the reform on the enrollment probabilities for the *Gymnasium*. X_i is a vector with control variables and u_i is the error term.

With this model, it is not possible to answer *H2* with DiD because *H2* assumes a different effect of the reform due to students' SES. The effect on enrollment probability in *Gymnasium* is expected to be negative for low SES students and negligible for high SES students. DiD can only give a treatment effect on the probability to attend a *Gymnasium*. However, SES-specific effects are needed to be included into the model to answer *H2*, thus it is necessary to extend the DiD estimator. With a difference-in-difference-in-difference (DDD) estimator it is possible to include another subgroup, next to federal state and period, as in the case of DiD (Wooldridge 2010).

$$y_i = \beta_0 + \beta_1 S_i + \beta_2 T_i + \beta_3 (S_i * T_i) + \beta_4 SES_i + \beta_5 (S_i * SES_i) + \beta_6 (T_i * SES_i) + \beta_7 (S_i * T_i * SES_i) + \beta_C X_i + u_i \quad (2)$$

Equation 2 specifies the DDD estimator for the effect of the reform by SES background of a student, which is an extension of equation 1. SES_i is a dichotomous variable for the family's SES of a student. Indicating if a student's ISEI is part of the lower half of the ISEI distribution.

It equals one for low SES background and zero otherwise. β_7 is the coefficient of interest. It multiplies the result of the interaction between the federal state, time and SES background and indicates the effect of the reform for low SES students in Lower Saxony. To answer *H3*, however, we estimate equation 1 only for students with low SES and above-average performance. Compared to equation 1, the state-time interaction then indicates the reform effect for students with above-average performance and a low SES.

The dichotomous dependent variable would suggest using logistic regression models for estimating the DiD and DDD models. Yet, the transfer of logistic regressions into the logic of DiD is problematic and can lead to violations of the parallel trend assumption. One alternative to avoid such problems is to use linear probability models (LPM) with robust standard errors instead of logistic regression (Lechner 2011). Consequently, all DiD and DDD models use LPM for estimation.

The following section describes the variables for the analysis. The dependent variable is a dichotomous variable for the school track enrollment of a student into *Gymnasium* or not. Vocational schools, special schools, and Waldorf schools are excluded from the analysis. The most prestigious educational track in Germany is the *Gymnasium*. At the end of *Gymnasium*, it is possible to obtain the *Abitur*, which is the only way for complete and direct access to a university. The school track enrollment is used as a proxy variable for the actual degree, which is not observed because students have not finished school at the point of data collection. This is in some way problematic, because some students could still change tracks or could obtain a higher degree through other means, such as evening classes. This could lead to biases in the analysis, in which the effect of the reform would be overestimated. However, it is assumed that this behavior occurs only for a small number of students (Maaz 2006). Moreover, students who had to repeat a class were excluded from the analyses.

Independent variables in the models are a dichotomous indicator showing affiliation to treatment or control group, a dichotomous variable indicating the year of survey (PISA-E 2006 or IQB-LV 2009), and the SES background of a student. The highest International Socio-Economic Index of Occupational Status (ISEI) in the family measures the SES background. Control variables in the models include a dichotomous variable indicating if parental education is below *Abitur* or not, German used as the first language at home (Ruhose & Schwerdt 2016), and gender (Pekkarinen 2008).

The DDD models for *H2* also include a dichotomous variable indicating whether a student belongs to the lower half of the ISEI distribution. In each case, the models for *H3* include only students who demonstrate above-average reading performance and belong to the lower half of the ISEI distribution. The performance is represented with plausible values. Five plausible values are given for each student. Plausible values are random draws from the posterior distributions and therefore they are not suitable as an estimate for the individual performance of a student. For dealing with plausible values, the OECD proposes a procedure in which, among other things, a single model is to be estimated with each plausible value and the results are to be averaged (OECD 2012). The properties of plausible values also mean that students who perform above average on one plausible value may not perform above average on another plausible value. Therefore, the number of cases varies between models for *H3* and the OECD procedure cannot be applied here. Consequently, five independent models are estimated to test *H3*, each with a different plausible value as a performance indicator.

The causal interpretation of the DiD estimator rests on the parallel (or common) trend assumption. This assumption states that had no reform (i.e. no treatment) occurred, *Gymnasium* attendance in both groups would have developed parallel to each other (Angrist & Pischke 2009; Gangl 2010). Pre-treatment data is required to test the parallel trend assumption. In this case from PISA-E 2000 to 2006. If the control and treatment groups show parallel trends in the outcome variable before the treatment, this is seen as an indication the assumption holds. To test the assumption, we perform a placebo DiD analysis (Gertler et al. 2016). For this purpose, the period before the reform is used. If the parallel trend assumption holds, then the interaction term between year and treatment group should be not significant and close to zero. Table 18 in the appendix shows the placebo DiD analysis, which is a DiD analysis of the pre-reform period from 2000 to 2006. Based solely on significance, the placebo DiD suggests that the parallel trend assumption holds because both models, with and without covariates, do not display significant different pre-reform trends for 2003 and 2006 compared to 2000. However, the interaction between the year 2006 and the treatment group indicates a positive coefficient that is not close to zero. Model 2 shows a positive time trend. This indicates that the parallel trend assumption does not hold. Since the point estimates are not significant, we assume for the following analysis that the parallel trend assumption holds. With further robustness checks, we will again address the problem of a possible violation of the parallel trend assumption.

In addition, the DID estimator requires the stable unit treatment value assumption (SUTVA) for unbiased identification of the treatment effect. It is assumed that only the outcome of the treatment or the control group is observed and no exchange between the groups takes place. Spillovers between the treatment and the control group (e.g. internal migration into and out of Lower Saxony) would violate SUTVA. The share in Lower Saxony's total population of internal migration to and from Lower Saxony has fluctuated between 1.39 and 1.55 percentage points for inflows respectively 1.47 and 2.48 percentage points for outflows in the period from 2000 – 2012 (see table 19 in the appendix). Possible spillovers will not cause major biases. Therefore, we assume that SUTVA holds.

Another assumption concerns the anticipation of the reform by the treatment group and possible reactions of this group. Parents could try to avoid the reform and, for example, purposefully move away from Lower Saxony. The figures on migration from Lower Saxony give no indication of this. The percentage of children aged five to fifteen moving away from Lower Saxony decreases between 2000 and 2012, and the reform year 2004 does not stand out (see table 19 in the appendix). Nor do we see a switch to an alternative form of school without reform, such as Waldorf schools, based on total numbers of students. The number of students in alternative school types, such as Waldorf schools, does not increase rapidly in 2004 and 2005 (Niedersächsisches Kultusministerium 2006). However, there is another education reform, which could bring changes in educational behavior, by anticipating this reform. So far, only the education reform in Bremen was not considered (see above). Federal states implemented so called, G8 reforms at different points in time, from 2001 to 2007. The G8 reforms shortened the length of the *Gymnasium* by one year (see figure 3). Lower Saxony introduced its G8 reform in 2005, at a similar time to the abolition of the OS (Homuth 2017). Theoretically, the G8 reform could influence educational decisions (Homuth 2017). This may mean that treatment and control group are affected differently by the G8 reform and thus a possible effect of the OS reform on educational decisions cannot be clearly identified. Research finds no impact of the G8 reforms on educational decisions to attain the *Abitur* (Homuth 2017; Roth 2019). However, there is a tendency for students with a higher SES to switch to other school types without the G8 reform, such as the integrated comprehensive school, in order to obtain the *Abitur* (Roth 2019). Therefore, we perform robustness checks with a dichotomous variable indicating whether a student attends an integrated comprehensive school or not as well as an interaction between the integrated comprehensive school indicator and the indicator for lower half of ISEI.

3.5 Results

This section reports the results from the statistical analyses of preponing educational stratification in Lower Saxony. Table 9 shows the mean and standard deviation for the dependent variable and covariates in the sample for 2006 and 2009 for the full sample and separately for treatment and control group. From 2006 to 2009, the full sample and the control group show a slight increase in students attending *Gymnasium* of 3 percentage points, while *Gymnasium* attendance increases by 6 percentage points in Lower Saxony. Regarding parental background, the proportion of parents with *Abitur* or higher education is lower in Lower Saxony. The average ISEI is very similar between Lower Saxony and the control group, with almost no change between 2006 and 2009. In 2006, female students are slightly overrepresented. German as the first language at home is a little more frequent in Lower Saxony, compared to the control group.

Table 9: Descriptive statistics of *Gymnasium* attendance and covariates for the period 2006 and 2009

Variables		Full Sample		Lower Saxony (Treatment)		Rest of Germany (Bremen excluded)	
	Years	Mean	sd	Mean	sd	Mean	sd
Gymnasium	2006	0.418	0.493	0.426	0.495	0.417	0.493
	2009	0.448	0.497	0.485	0.500	0.444	0.497
Parental Education (*Abitur* or higher)	2006	0.558	0.497	0.542	0.498	0.560	0.496
	2009	0.405	0.491	0.358	0.480	0.411	0.492
ISEI	2006	51.903	15.913	50.444	15.771	52.065	15.921
	2009	51.164	15.833	50.354	16.169	51.269	15.787
Male	2006	0.463	0.499	0.442	0.497	0.466	0.499
	2009	0.515	0.500	0.520	0.500	0.515	0.500
German at Home	2006	0.941	0.235	0.957	0.202	0.940	0.238
	2009	0.945	0.228	0.974	0.158	0.941	0.235
N	2006	19,975		1,129		18,846	
	2009	13,992		785		13,207	

Notes: Data: PISA-E 2000, 2003, 2006 and IQB-LV 2009 and 2012.

Table 10 reports the results for the effect of preponing tracking on average *Gymnasium* attendance in Lower Saxony. According to *H1*, preponing tracking harms *Gymnasium* attendance. The results in table 10 do not support this hypothesis. While with the addition of control variables in model 2, the point estimate shows an increase by 3 percentage point in average *Gymnasium* attendance. However, the coefficient is not statistically significant. There is no signifi-

cant effect of preponing educational tracking by two years on the average *Gymnasium* attendance in Lower Saxony. This finding is in line with a study by Combet (2019) for Switzerland. In the analysis from Combet, the differences in the timing of tracking between Swiss cantons are similarly small as the differences between pre-and post-reform tracking in Lower Saxony. While research, which finds a positive effect of late tracking, shows a larger difference between early and late tracking (Berger & Combet 2017; Meghir & Palme 2005). This could be explained by the possibility that not just a shift in the timing of tracking affects educational decisions, but that the timing of tracking must be shifted for a certain amount to affect educational decisions significantly. However, this cannot be tested with the present data. The results in table 10 only consider the short-term effects of preponing tracking. Table 20 in the appendix utilizes all available periods to consider possible mid-term effects of preponing tracking on *Gymnasium* attendance in Lower Saxony. There is no significant mid-term effect of the educational reform on average *Gymnasium* attendance.

Table 10: DiD results for the effect of preponed tracking in Lower Saxony from 2006 to 2009

	Model 1	Model 2
Year 2009 X Treatment	0.032 (0.120)	0.030 (0.102)
Year 2009	0.027 (0.037)	0.058 (0.031)
Treatment	0.009 (0.072)	0.026 (0.061)
ISEI		0.010^{***} (0.000)
Parental Education (*Abitur* or higher)		0.155^{***} (0.010)
Male		-0.017 (0.010)
German at Home		0.073^{***} (0.018)
Intercept	0.417^{***} (0.024)	-0.238^{***} (0.025)
N	33,967	33,967
R^2	0.001	0.169
adj. R^2	0.001	0.169

Notes: Data: PISA-E 2006 and IQB-LV 2009. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

While there is no evidence of a negative effect of preponing educational tracking on *Gymnasium* attendance, students with a low SES background could be negatively affected to attend *Gymnasium* by the reform in Lower Saxony because this group could be very sensitive to changes in the timing of tracking. According to *H2*, preponing tracking harms *Gymnasium* attendance for socially disadvantaged students. As described above, an additional dichotomous variable is included in the models to indicate whether a student belongs to the lower half of the

ISEI distribution. Table 11 reports the results of the DDD estimator for the effect of preponed tracking on *Gymnasium* attendance for students, with a low SES background, in Lower Saxony. *H2* is not confirmed by the results in table 11. Although the point estimate of the interaction between year, treatment group, and low ISEI indicates a drop of about 6 percentage points for low SES students in gymnasium, the point estimate is not significant. In addition to the non-existent short-term effect, there is also no medium-term effect of the reform (see table 21 in the appendix).

Table 11: DDD results for the effect of preponed tracking in Lower Saxony from 2006 to 2009

	Model 1	Model 2
Year 2009 X Treatment X ISEI (lower half)	-0.060 (0.065)	-0.058 (0.060)
Year 2009 X Treatment	0.048 (0.116)	0.056 (0.109)
Year 2009 X ISEI (lower half)	-0.010 (0.023)	-0.026 (0.021)
Treatment X ISEI (lower half)	-0.004 (0.043)	-0.015 (0.041)
Year 2009	0.019 (0.039)	0.068 (0.035)
Treatment	0.029 (0.075)	0.034 (0.070)
ISEI (lower half)	-0.317*** (0.016)	-0.032 (0.017)
ISEI		0.009*** (0.000)
Parental Education (*Abitur* or higher)		0.156*** (0.010)
Male		-0.017 (0.010)
German at Home		0.072*** (0.018)
Intercept	0.527***(0.026)	-0.163*** (0.036)
N	33,967	33,967
R^2	0.107	0.170
adj. R^2	0.107	0.170

Notes: Data: PISA-E 2006 and IQB-LV 2009. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

Table 12: DiD results for the effect of preponed tracking for students with low SES and above-average reading performance in Lower Saxony from 2006 to 2009 (without control variables)

	Model 1.1	Model 1.2	Model 1.3	Model 1.4	Model 1.5
Year 2009 X	0.062	0.007	0.015	-0.007	0.085
Treatment	(0.143)	(0.153)	(0.145)	(0.148)	(0.140)
Year 2009	0.007	0.024	0.031	0.015	0.017
	(0.044)	(0.044)	(0.044)	(0.044)	(0.044)
Treatment	0.047	0.044	0.034	0.048	0.010
	(0.086)	(0.088)	(0.085)	(0.087)	(0.085)
Intercept	0.476***	0.469***	0.464***	0.469***	0.464***
	(0.030)	(0.030)	(0.029)	(0.029)	(0.029)
N	6,588	6,654	6,531	6,617	6,579
R^2	0.002	0.001	0.002	0.001	0.002
adj. R^2	0.002	0.001	0.001	0.001	0.001

Notes: Data: PISA-E 2006 and IQB-LV 2009. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

The following models test the hypothesis (*H3*) that especially high-performing students with low SES have a lower probability of attending a *Gymnasium* because of the reform in Lower Saxony. As already mentioned, performance is measured with plausible values and for each plausible value a separate model is estimated. As a result, the number of cases per model varies. Therefore, the models cannot be compared directly with each other. However, they generally indicate whether *H3* is confirmed or not. Table 12 shows the models without control variables and table 13 with control variables. Contrary to the hypothesis, the models with and without control variables show that the probability of attending a *Gymnasium* increased for high-performing students with low SES after the reform. However, the increase is not statistically significant in any model. This means that the models do not confirm *H3*.

Table 13: DiD results for the effect of preponed tracking for students with low SES and above-average reading performance in Lower Saxony from 2006 to 2009 (with control variables)

	Model 2.1	Model 2.2	Model 2.3	Model 2.4	Model 2.5
Year 2009 X Treatment	0.066	0.020	0.031	0.003	0.095
	(0.137)	(0.147)	(0.140)	(0.141)	(0.135)
Year 2009	0.049	0.063	0.068	0.052	0.058
	(0.043)	(0.042)	(0.042)	(0.042)	(0.042)
Treatment	0.048	0.049	0.035	0.052	0.007
	(0.083)	(0.085)	(0.082)	(0.083)	(0.083)
ISEI	0.009***	0.009***	0.010***	0.009***	0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Parental Education (*Abitur* or higher)	0.152***	0.144***	0.154***	0.152***	0.149***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Male	0.006	0.003	-0.002	0.007	-0.006
	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)
German at Home	-0.048	-0.078	-0.020	0.010	-0.020
	(0.046)	(0.043)	(0.044)	(0.044)	(0.043)
Intercept	0.085	0.099	0.025	0.031	0.041
	(0.068)	(0.072)	(0.066)	(0.063)	(0.064)
N	6,588	6,654	6,531	6,617	6,579
R^2	0.050	0.048	0.054	0.048	0.051
adj. R^2	0.049	0.047	0.053	0.047	0.050

Notes: Data: PISA-E 2006 and IQB-LV 2009. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

3.6 Robustness checks

Table 14: DiD results for alternative control group specification and nonparametric synthetic control for the effect of preponed tracking in Lower Saxony

	Alternative Control Group		Nonparametric Synthetic Control	
	Model 1	Model 2	Model 1	Model 2
Year 2009 X Treatment	0.041	0.045	-0.091	-0.106
	(0.131)	(0.111)	(0.111)	(0.097)
Year 2009	0.018	0.042	0.014	0.048
	(0.065)	(0.053)	(0.040)	(0.035)
Treatment	0.022	0.034	0.016	0.031
	(0.078)	(0.066)	(0.076)	(0.065)
ISEI		0.010***		0.009***
		(0.001)		(0.001)
Parental Education (*Abitur* or higher)		0.148***		0.128***
		(0.015)		(0.014)
Male		-0.022		-0.034*
		(0.014)		(0.013)
German at Home		0.046		0.055
		(0.032)		(0.031)
Intercept	0.404***	-0.233***	0.399***	-0.182***
	(0.039)	(0.044)	(0.027)	(0.040)
N	10,550	10,550	15,860	15,860
R^2	0.002	0.174	0.003	0.140
adj. R^2	0.002	0.173	0.003	0.140

Notes: Data: PISA-E 2006 and IQB-LV 2009. All estimations using population weights (alternative control group) or average unit weights (nonparametric synthetic control). Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

To check the robustness of the results, especially due to the possible problems of the main analysis, alternative model specifications and other methods were used for further analysis. The control group in the initial analysis only excludes Bremen, due to a similar educational reform as in Lower Saxony, and other reforms that occurred simultaneously. However, several other German states have also implemented educational reforms during the period under study. This may lead to a biased estimate of the reform effect. Therefore, an alternative control group is created using only the states of Baden-Wuerttemberg, Rhineland-Palatinate, Saxony, and Thuringia. In the alternative control group, there is no change in the binding nature of the transition recommendation, grade binding in the transition recommendation to the *Gymnasium*, entrance exams, or the timing of tracking (Büchler 2016; Helbig & Nikolai 2015). While other federal states had reforms of those characteristics. The coefficients of the placebo DiD are very

similar to the result of the placebo DiD analysis of the main analysis (see table 23 in the appendix). Therefore, it should be emphasized that although the coefficients are not significant, they are not close to zero and thus a time trend seems to be present. Table 14 reports the results of the DiD estimator for the alternative control group. The results confirm the findings of the main analysis. While an increase in the average *Gymnasium* attendance of about 4.5 percentage points after the education reform in Lower Saxony is detectable, it turns out that this increase is not statistically significant. Moreover, the results of the DDD analysis show also similar findings for both control group specifications. After the education reform in Lower Saxony, the average *Gymnasium* attendance of low-SES students' drops by 6.7 percentage points, but this drop is not statistically significant (see table 15).

For another robustness check, a synthetic control group was formed to increase the fit between Lower Saxony and the control group concerning confounders. The descriptive analysis shows some differences between the groups, for example, the share of parents with *Abitur* or a higher education is consistently lower in Lower Saxony (see table 9). This may bias the estimation of the reform effect. Additionally, synthetic control does not rely on the parallel trend assumption (Abadie 2021). Thus, any problems due to a violation of the parallel trend assumption can be avoided. The synthetic control method (Abadie et al. 2010) can be applied based on a federal-state panel constructed from PISA-E and IQB-LV data. Because of the small sample size on a federal state level, the nonparametric extension of the parametric synthetic control is used (Cerulli 2019). The nonparametric synthetic control imputes the counterfactual status of Lower Saxony without the educational reform as a weighted average of covariates and estimates an average unit weight for each state in the control group. Covariates for estimating the average unit weight are means on the federal state level for parental education, ISEI, male and German used at home. For the average unit weights, see table 22 in the appendix. Subsequently, the average unit weights for federal state are used in the analysis. Unlike the main analysis, the analyses with the nonparametric synthetic control show a decrease in the average attendance of the *Gymnasium* after the reform in Lower Saxony (see table 14) and an increase in the average attendance of students with low SES (see table 15). However, again the DiD and DDD coefficients are not significant.

Table 15: DDD results for alternative control group specifications and nonparametric synthetic control for the effect of preponed tracking in Lower Saxony

	Alternative Control Group		Nonparametric Synthetic Control	
	Model 1	Model 2	Model 1	Model 2
Year 2009 X Treatment	-0.072	-0.067	0.063	0.063
X ISEI (lower half)	(0.075)	(0.067)	(0.074)	(0.069)
Year 2009 X Treatment	0.065	0.075	-0.147	-0.140
	(0.130)	(0.120)	(0.124)	(0.116)
Year 2009 X ISEI	0.002	-0.015	0.001	-0.006
(lower half)	(0.043)	(0.038)	(0.027)	(0.025)
Treatment X ISEI	-0.007	-0.018	-0.044	-0.049
(lower half)	(0.049)	(0.046)	(0.043)	(0.040)
Year 2009	0.001	0.049	0.008	0.049
	(0.070)	(0.063)	(0.044)	(0.041)
Treatment	0.037	0.044	0.052	0.057
	(0.083)	(0.077)	(0.081)	(0.074)
ISEI (lower half)	-0.314***	-0.013	-0.281***	-0.024
	(0.028)	(0.028)	(0.018)	(0.022)
ISEI		0.009***		0.008***
		(0.001)		(0.001)
Parental Education (*Abi-*		0.150***		0.128***
tur or higher)		(0.015)		(0.014)
Male		-0.022		-0.034*
		(0.014)		(0.013)
German at Home		0.048		0.054
		(0.032)		(0.031)
Intercept	0.564***	-0.189***	0.542***	-0.120*
	(0.043)	(0.060)	(0.030)	(0.051)
N	10,550	10,550	15,860	15,860
R^2	0.109	0.175	0.090	0.141
adj. R^2	0.108	0.174	0.090	0.141

Notes: Data: PISA-E 2006 and IQB-LV 2009. All estimations using population weights (alternative control group) or average unit weights (nonparametric synthetic control). Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

Table 16: Pooled OLS on *Gymnasium* attendance for the full sample and by SES background

	Full sample	ISEI (lower half)	ISEI (upper half)
Year 2000 X Lower	-0.069	-0.031	-0.108
Saxony	(0.105)	(0.094)	(0.124)
Year 2003 X Lower	-0.051	-0.042	-0.058
Saxony	(0.107)	(0.094)	(0.127)
Year 2009 X Lower	0.061	0.035	0.077
Saxony	(0.134)	(0.126)	(0.147)
Year 2012 X Lower	0.018	0.037	-0.002
Saxony	(0.111)	(0.105)	(0.128)
Lower Saxony	0.051	0.050	0.058
	(0.078)	(0.073)	(0.091)
Year 2000	0.008	-0.025	0.032
	(0.073)	(0.063)	(0.088)
Year 2003	0.016	-0.003	0.021
	(0.074)	(0.062)	(0.088)
Year 2009	0.028	-0.000	0.051
	(0.092)	(0.082)	(0.106)
Year 2012	0.016	0.017	0.009
	(0.076)	(0.067)	(0.091)
ISEI	0.009***	0.007***	0.009***
	(0.000)	(0.000)	(0.000)
Parental Education	0.149***	0.126***	0.164***
(*Abitur* or higher)	(0.005)	(0.007)	(0.007)
Male	-0.042***	-0.048***	-0.036***
	(0.006)	(0.006)	(0.008)
German at Home	0.064***	0.049***	0.108***
	(0.010)	(0.010)	(0.018)
Intercept	0.418***	0.190***	0.381***
	(0.058)	(0.047)	(0.066)
N	102,540	48,458	54,082
R^2	0.178	0.062	0.099
adj. R^2	0.178	0.060	0.098

Notes: Data: PISA-E 2000, 2003, 2006 and IQB-LV 2009 and 2012. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Besides the indicator for Lower Saxony, federal state indicators and state-time interactions are not shown. Baden-Wuerttemberg is the reference category. Significance levels: * p < .05, ** p < .01, *** p < .001.

The previous analyses control for group-specific time trends between treatment and control groups. However, state-specific time trends may bias the results, for example, as a result of not having controlled for other state-specific characteristics that could influence educational decisions. Therefore, the effect of the educational reform on average *Gymnasium* attendance is also analyzed by using a pooled OLS model with state fixed effects and state-time interactions to control state-specific time trends. This approach is similar to the fixed effect with individual slopes (Rüttenauer & Ludwig 2020). Table 16 reports the results of the pooled OLS regressions with state-specific time trends for the full sample, as well as separate models for the upper and lower half of ISEI for Lower Saxony compared to Baden-Wuerttemberg. The focus here should lie on the state-time interaction for 2009. The model for the full sample shows a positive point estimate, which indicates an increase in average *Gymnasium* attendance of roughly 6 percentage point. The models for the upper and lower halves of the ISEI also show positive point estimates, suggesting that, unlike in the main analysis, average attendance at the *Gymnasium* increased regardless of social background. However, again, the point estimates relevant to the question are not statistically significant. The findings of the initial DiD and DDD analysis seem robust.

The final robustness check controls for whether a student attends an integrated comprehensive school. Most integrated comprehensive schools also offer the *Abitur*. Thus, this type of school offers an alternative way to obtain the *Abitur*. The integrated comprehensive school is a widespread school type in Germany, but it is also unequally distributed among the federal states (Autorengruppe Bildungsberichterstattung 2010). Thus, by controlling students at integrated comprehensive schools, some of the state differences in access to the *Abitur* can also be kept constant. With the introduction of the G8 reform, the *Abitur* at the *Gymnasium* was acquired after grade 12 instead of grade 13. As a result, there was an SES-specific evasion behavior. Students with a higher SES were increasingly sent to school types, such as the integrated comprehensive school, where pupils could still obtain the *Abitur* after the 13th grade (Roth 2019). Therefore, in addition to controlling for students in comprehensive schools, we include an interaction between comprehensive school and the indicator for belonging to the lower half of the ISEI. Table 17 presents the results of the DiD analysis. The results of the main analysis can be confirmed once again. The reform increases the average *Gymnasium* attendance by 2.6 percentage points. However, this increase is not statistically significant. All robustness tests confirm the results of the main analysis with different model specifications and methods. Again,

preponing the timing of tracking by two years has no significant effect on average Gymnasium attendance. There is also no significant effect for students with low SES.

Table 17: DiD results for the analysis with control of comprehensive school (CS) for the effect of preponed tracking in Lower Saxony

	Analysis with CS Control	
	Model 1	Model 2
Year 2009 X Treatment	0.017 (0.103)	0.026 (0.099)
Year 2009	0.029 (0.031)	0.068* (0.029)
Treatment	0.012 (0.063)	0.012 (0.059)
CS	-0.633*** (0.019)	-0.607*** (0.018)
ISEI (lower half of ISEI)	-0.341*** (0.011)	-0.075*** (0.015)
CS X ISEI (lower half of ISEI)	-0.344*** (0.011)	-0.324*** (0.011)
ISEI		0.008*** (0.000)
Parental Education (*Abitur* or higher)		0.148*** (0.009)
Male		-0.017 (0.010)
German at Home		0.063*** (0.018)
Intercept	0.619*** (0.024)	-0.075* (0.037)
N	33,967	33,967
R^2	0.181	0.238
adj. R^2	0.181	0.238

Notes: Data: PISA-E 2006 and IQB-LV 2009. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

3.7 Conclusion

Timing of tracking is considered to be crucial for the formation and persistence of educational inequalities. Countries with an early tracking system usually display higher educational inequalities in respect of performance development and educational decision-making in the educational system. In countries with a late tracking system, these inequalities are usually lower (Cordero et al. 2018; Hanushek & Wössmann 2011). This study aimed to examine the effects of an educational reform that preponed tracking in Lower Saxony on *Gymnasium* attendance. Through the reform in Lower Saxony, the placement into separate schools in secondary education was preponed by two years, from grade seven to grade five. The analysis utilizes the reform as a natural experiment and estimates the causal effect of the reform with difference-in-difference and difference-in-difference-in-difference estimators based on PISA-E 2006 and IQB-LV 2009 data.

Theoretically, the timing of tracking can influence educational decisions, based on the change in uncertainty about a student's performance development. With late tracking comes more certainty about the performance development, which in turn allows for a more rational decision and less SES bias. *H1* assumes that the abolishment of the OS negatively affects *Gymnasium* attendance on students in general. However, the results of a DiD analysis show a positive but not significant effect of preponing tracking for two years on average *Gymnasium* attendance in Lower Saxony, and do not support the hypothesis. While *H1* considers an equal effect of the reform on all students in Lower Saxony, *H2* assumes a differential effect of the reform among students with different SES backgrounds. *H2* states that the preponing of tracking has a negative effect on *Gymnasium* attendance for low SES students. Although DDD analyses show a negative effect of preponing tracking on average *Gymnasium* attendance of low SES students in Lower Saxony, however, this effect is not significant. Thus, the results do not support *H2*. An increase in uncertainty in a student's performance development should have a particular impact on students with low SES and above average performance. Therefore, *H3* assumes a performance-specific effect in addition to the SES-specific one. For the most part, the models show an increase in the average attendance of students in *Gymnasium* with low SES and above-average performance after the reform, but this increase is not statistically significant. The results do not confirm *H3*. In addition to the main analyses, various robustness checks were performed. The main analyses were checked with an alternative control group specification, a nonparametric synthetic control group, pooled OLS models controlling for state-specific time trends, and an analysis with control of students in integrated comprehensive schools. In general, the robustness checks confirmed the results of the main analyses. Although some of the point estimators point in a different direction, no significant result could be estimated with any of the robustness checks. In conclusion, the OS reform in Lower Saxony has widened performance gaps between high- and low-performing students, benefiting especially students with high educational backgrounds (Roller & Steinberg 2020). However, the reform does not affect the educational decision in favor of the *Gymnasium* in general, as well as for low SES students and low SES students with above average reading performance.

Those results are in line with findings, which show that small differences in tracking time between early and late tracking do not influence educational decisions (Combet 2019). While research with a higher difference between the timing of early and late tracking shows a significant reduction in educational inequalities (Berger & Combet 2017; Meghir & Palme 2005). This

suggests that not just a change in tracking time is necessary to affect educational inequalities. Rather, that the change in timing of tracking needs to reduce uncertainty significantly to affect educational inequalities. However, this cannot be tested with the present data.

Further research should further focus on the mechanisms behind the effects of timing of tracking. The influence on the decision-making process for SES groups and performance groups should be of major concern. The number of studies, especially on the influence of tracking on educational decision-making, is limited, and more empirical testing of theoretical predictions under different settings is crucial for an understanding of timing of the tracking on educational inequalities. In particular, the variation in the timing of tracking should be examined in more detail and the impact of this variation on uncertainty in the educational decision process.

3.A Appendix Chapter 3

Table 18: Placebo DiD for the period from 2000 to 2006

	Model 1	Model 2
Year 2003 X Treatment	-0.020 (0.088)	-0.003 (0.074)
Year 2006 X Treatment	0.031 (0.094)	0.040 (0.079)
Year 2003	0.028 (0.029)	0.030 (0.024)
Year 2006	0.044 (0.031)	0.020 (0.026)
Treatment	-0.021 (0.061)	-0.015 (0.050)
ISEI		0.009*** (0.000)
Parental Education (*Abitur* or higher)		0.161*** (0.006)
Male		-0.061*** (0.007)
German at Home		0.076*** (0.012)
Intercept	0.373*** (0.020)	-0.229*** (0.020)
N	66,168	66,168
R^2	0.002	0.173
adj. R^2	0.002	0.173

Notes: Data: PISA-E 2000, 2003, 2006. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

Table 19: Internal German migration in and out of Lower Saxony from 2000 to 2012

Year (Age)	In Migration from Germany	Percentage Share at Population	Out Migration into Germany	Percentage Share at Population	Population of Lower Saxony
2000	122,137	1.54	181,179	2.29	
(5-10)	14,377	0.18	12,591	0.16	7,926,193
(10-15)	15,800	0.20	12,898	0.16	
2001	122,806	1.54	197,159	2.48	
(5-10)	14,839	0.19	12,406	0.16	7,956,416
(10-15)	16,200	0.20	13,170	0.17	
2002	123,678	1.55	188,530	2.36	
(5-10)	13,732	0.17	11,661	0.15	7,980,472
(10-15)	14,487	0.18	12,142	0.15	
2003	122,336	1.53	173,126	2.17	
(5-10)	11,902	0.15	10,334	0.13	7,993,415
(10-15)	11,666	0.15	9,879	0.12	
2004	118,934	1.49	161,857	2.02	
(5-10)	10,307	0.13	9,147	0.11	8,000,909
(10-15)	9,963	0.12	8,662	0.11	
2005	112,133	1.40	143,384	1.79	
(5-10)	8,288	0.10	7,702	0.10	7,993,946
(10-15)	7,465	0.09	6,786	0.08	
2006	111,289	1.39	118,964	1.49	
(5-10)	6,184	0.08	5,683	0.07	7,982,685
(10-15)	5,138	0.06	4,818	0.06	
2007	113,768	1.43	119,590	1.50	
(5-10)	6,009	0.08	5,829	0.07	7,971,684
(10-15)	5,097	0.06	4,767	0.06	
2008	117,048	1.47	122,335	1.54	
(5-10)	6,061	0.08	5,956	0.07	7,947,244
(10-15)	5,055	0.06	4,958	0.06	
2009	117,460	1.48	120,251	1.52	
(5-10)	6,174	0.08	5,746	0.07	7,928,815
(10-15)	5,212	0.07	4,701	0.06	
2010	113,803	1.44	116,294	1.47	
(5-10)	6,087	0.08	5,058	0.06	7,918,293
(10-15)	5,068	0.06	4,235	0.05	
2011	119,384	1.54	123,818	1.59	
(5-10)	6,522	0.08	5,295	0.07	7,774,253
(10-15)	5,181	0.07	4,221	0.05	
2012	120,310	1,55	119,177	1.53	
(5-10)	7,044	0.09	5,096	0.07	7,778,995
(10-15)	5,725	0.07	4,070	0.05	

Notes: Based on the numbers of the State Statistical Office in Lower Saxony for the respective years.

Table 20: DiD results for the effect of preponed tracking in Lower Saxony from 2000 to 2012

	Model 1	Model 2
Year 2000 X Treatment	-0.031 (0.094)	-0.040 (0.079)
Year 2003 X Treatment	-0.051 (0.096)	-0.044 (0.082)
Year 2009 X Treatment	0.032 (0.120)	0.031 (0.103)
Year 2012 X Treatment	0.008 (0.103)	-0.015 (0.086)
Year 2000	-0.044 (0.031)	-0.020 (0.026)
Year 2003	-0.016 (0.032)	0.009 (0.027)
Year 2009	0.027 (0.037)	0.058 (0.031)
Year 2012	0.056 (0.034)	0.050 (0.028)
Treatment	0.009 (0.072)	0.025 (0.061)
ISEI		0.009*** (0.000)
Parental Education (*Abitur* or higher)		0.150*** (0.005)
Male		-0.042*** (0.006)
German at Home		0.063*** (0.010)
Intercept	0.417 *** (0.024)	-0.190*** (0.022)
N	102,540	102,540
R^2	0.006	0.174
adj. R^2	0.006	0.173

Notes: Data: PISA-E 2000, 2003, 2006 and IQB-LV 2009 and 2012. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

Table 21: DDD results for the effect of preponed tracking in Lower Saxony from 2000 to 2012

	Model 1	Model 2
Year 2000 X Treatment X ISEI (lower half)	0.021 (0.058)	0.036 (0.054)
Year 2003 X Treatment X ISEI (lower half)	-0.009 (0.062)	0.006 (0.058)
Year 2009 X Treatment X ISEI (lower half)	-0.060 (0.065)	-0.058 (0.060)
Year 2012 X Treatment X ISEI (lower half)	-0.007 (0.059)	0.001 (0.055)
Year 2000 X Treatment	-0.058 (0.101)	-0.059 (0.093)
Year 2003 X Treatment	-0.053 (0.104)	-0.048 (0.097)
Year 2009 X Treatment	0.048 (0.116)	0.056 (0.109)
Year 2012 X Treatment	-0.011 (0.104)	-0.018 (0.096)
Year 2000 X ISEI (lower half)	-0.001 (0.021)	-0.009 (0.019)
Year 2003 X ISEI (lower half)	0.005 (0.021)	-0.001 (0.019)
Year 2009 X ISEI (lower half)	-0.010 (0.023)	-0.025 (0.022)
Year 2012 X ISEI (lower half)	-0.022 (0.022)	0.054** (0.020)
Treatment X ISEI (lower half)	-0.004 (0.043)	-0.014 (0.041)
Year 2000	-0.047 (0.034)	-0.018 (0.031)
Year 2003	0.072* (0.033)	0.086** (0.033)
Year 2009	0.068 (0.036)	0.113** (0.035)
Year 2012	0.055 (0.035)	0.026 (0.032)
Treatment	0.029 (0.075)	0.033 (0.070)
ISEI (lower half)	-0.317*** (0.016)	-0.035* (0.015)
ISEI		-0.009*** (0.000)
Parental Education (*Abitur* or higher)		0.149*** (0.005)
Male		-0.042*** (0.006)
German at Home		0.064*** (0.010)
Intercept	0.572*** (0.026)	-0.138*** (0.028)
N	102,540	102,540
R^2	0.112	0.175
adj. R^2	0.112	0.175

Notes: Data: PISA-E 2000, 2003, 2006 and IQB-LV 2009 and 2012. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

Table 22: Average unit weights by federal state, estimated with the nonparametric synthetic control method

Federal State	Average Unit Weight
Saarland	0
Rhineland Palatinate	0.221
North Rhine Westphalia	0
Schleswig-Holstein	0
Hamburg	0
Mecklenburg Western Pomerania	0.151
Brandenburg	0
Berlin	0
Saxony	0
Bavaria	0.182
Baden Württemberg	0
Hesse	0.113
Thuringia	0.302
Saxony-Anhalt	0.031

Table 23: Placebo DiD for the alternative control group specification and for the analysis with control of integrated comprehensive school (CS)

	Alternative Control Group		Analysis with CS Control	
	Model 1	Model 2	Model 1	Model 2
Year 2003 X Treatment	-0.031	-0.004	-0.008	-0.002
	(0.096)	(0.080)	(0.078)	(0.072)
Year 2006 X Treatment	0.029	0.039	0.052	0.045
	(0.102)	(0.085)	(0.083)	(0.077)
Year 2003	0.038	0.032	0.018	0.033
	(0.047)	(0.039)	(0.025)	(0.023)
Year 2006	0.046	0.021	0.039	0.016
	(0.050)	(0.042)	(0.027)	(0.025)
Treatment	-0.007	-0.006	-0.039	-0.033
	(0.065)	(0.054)	(0.054)	(0.049)
CS			-0.588***	-0.564***
			(0.013)	(0.013)
ISEI (lower half of ISEI)			-0.334***	-0.062***
			(0.008)	(0.010)
CS X ISEI (lower half of ISEI)			0.333***	0.311***
			(0.008)	(0.008)
ISEI		0.009***		0.008***
		(0.000)		(0.000)
Parental Education (*Abi-tur* or higher)		0.164***		0.159***
		(0.010)		(0.006)
Male		-0.084***		-0.056***
		(0.009)		(0.007)
German at Home		0.083***		0.071***
		(0.018)		(0.012)
Intercept	0.358***	-0.228***	0.573***	-0.095***
	(0.032)	(0.031)	(0.019)	(0.000)
N	22,590	22,590	66,168	66,168
R^2	0.003	0.175	0.173	0.236
adj. R^2	0.002	0.175	0.172	0.236

Notes: Data: PISA-E 2000, 2003, 2006. All estimations using population weights. Clustered standard errors are in parentheses. They are obtained by using schools as clusters. Significance levels: * p < .05, ** p < .01, *** p < .001.

4 Effects on performance of binding teacher recommendations for the transition to tracked secondary education

Abstract

Early division of students into different types of secondary education is a key part of the German education system. The division seeks to homogenize classes according to students' performance and thereby to promote performance development. The performance-specific division at the end of elementary school relies on teacher recommendations, which are binding or non-binding depending on the federal state. These recommendations determine students' secondary educational trajectories. However, this practice remains under debate, both theoretically and empirically. The empirical evidence relating teacher recommendations to students' performance is limited. Using difference-in-difference estimators on PISA data, this study seeks to investigate binding teacher recommendations' causal effects on educational performance, exploiting the quasi-experimental structure of an educational reform in the federal state of Bremen, which in 2003 made previously non-binding teacher recommendations binding. Our results show that making recommendations binding has no effect on students' performance, either in general or in certain school types, with point estimates being small (some close to zero) and not significant. This suggests that binding recommendations are not an appropriate implement for improving student performance.

4.1 Introduction

Early division (at the end of primary education) of students into different school types based on teacher recommendations differentiates the educational system in Germany from those of other Western societies. This division should assign students to more or less demanding secondary educational tracks as appropriate and lead to homogenization of performance within educational tracks (Betts 2011). Class teachers or conferences of class teachers determine these recommendations. However, federal states in Germany vary as to the binding character of teacher recommendations: Some states have binding teacher recommendations, which usually require families to follow teachers' assessments as to students' academic abilities and the associated educational paths. Other states provide more freedom of choice, treating recommendations as advisory, leaving educational decisions up to families (Büchler 2016; Helbig & Nikolai 2015).

Different views regarding the homogenization of performance in secondary education have emerged in the literature. Some scholars advocate early homogenization, arguing that homogenization of students' performance within classrooms may positively impact students' performance development through classroom composition and ability-appropriate teaching methods, making learning more efficient, benefiting all students and raising overall performance (Duflo et al. 2011; Esser & Hoenig 2018; Betts 2011). Others question the benefits of homogenization for students' performance development, pointing out that promoting homogenization could amplify educational inequalities tied to social background, especially if tracking begins early in children's educational careers and characteristics other than prior performance determine track placement (Betts 2011). Further, students in lower tracks in homogenous classrooms lose their high-ability peers, whose presence would have benefited their learning (Sacerdote 2011). Separating students according to ability can increase the performance gap between students in lower and in higher tracks and may lead to a decrease in average performance (Betts 2011). Empirical analyses of binding teacher recommendations and academic performance in secondary education have previously estimated the causal effect of recommendations using multilevel models (results were mixed) (Esser & Seuring 2020; Heisig & Matthewes 2021) Other results for elementary school performance show positive effects of binding recommendations (Bach & Fischer 2020). However, a change in teacher recommendations through legislative reform and the possible effects on performance in secondary education await specific analysis.

This study examines a 2003 educational reform in Bremen to investigate the effect of binding teacher recommendations on students' performance, utilizing data from PISA-E. The reform in Bremen changed teacher recommendations from non-binding to binding while leaving the basis for teacher recommendations unchanged (in Bremen, teachers could use grades and additional characteristics for their recommendations). This created an opportunity to investigate the effect of binding teacher recommendations on average performance, as well as possible heterogeneous effects by school type. We estimate the effect of the reform on performance using difference-in-difference estimators, a standard method for policy evaluation (Athey & Imbens 2017).

The structure of the article is as follows. First, we explain teacher recommendations and their impact on performance, including a discussion of findings from prior research. The second section presents the hypothesis. After that, we present the data used in our analysis of the reform of teacher recommendations in Bremen, which is the focus of this article. Then follows a presentation of the variables used in the analysis, as well as the analytic strategy and method, followed by the results of the analysis and a robustness check. The article ends with a discussion of the results, the limitations of its analysis, and a conclusion.

4.2 Teacher recommendation and performance

Before discussing different aspects of teacher recommendations and their implications for performance, we describe the basic structure of the German education system and certain deviations from it in Bremen. Children begin attending primary school (*Grundschule*) around the age of six. In most federal states (also in Bremen), primary school covers the first four years. The transition from primary to secondary education happens when students are approximately ten years old. Students split into two or three different secondary educational school tracks. The number of tracks in secondary education varies from one federal state to another. States with three tracks in secondary education have the lower secondary school (*Hauptschule*) as their lowest secondary track and the intermediate secondary school (*Realschule*) as their intermediate secondary track. The *Hauptschule* offers a degree after the ninth or tenth grade and the *Realschule* after the tenth. States with two tracks in secondary education have a multitrack combination of *Hauptschule* and *Realschule* — called schools with two educational programs (*Schularten mit zwei Bildungsgängen*) — offer both degrees. In Germany, lower and interme-

diate secondary education are vocationally oriented and lead into apprenticeships. The academic track of upper secondary education — called *Gymnasium* — offers the highest education entrance qualification (*Abitur*). This degree is the most common way to enter university in Germany (Eckhardt 2017; Maaz et al. 2008).[9] In the period of analysis, Bremen has three tracks in secondary education. Unlike in most other German states, however, students do not transfer directly to one of the three secondary school types after completing elementary education, but rather enter a two-year orientation stage independent of school type. Teachers make recommendations at the end of the orientation stage (Schuchart & Weishaupt 2004).

For the transition to the different secondary school tracks, all students in Germany receive teacher recommendations indicating which type of secondary education or track most suit them. Federal states vary as to the binding character of teacher recommendations. Some federal states permit less freedom of choice in the transition from primary to secondary education. In those states, teacher recommendations highly determine secondary school types. This should limit the influence of parental aspirations upon the transition process regardless of students' ability (Dollmann 2016; Roth & Siegert 2016). Other states offer more freedom of choice in the transition process: parents can use teacher recommendations as guidance in their own decision-making processes. Thus parents are in fact able fully to realize their aspirations regardless of their offspring's actual abilities (Dollmann 2016; Roth & Siegert 2016). In this context, previous research in Germany has found that students from families of higher socioeconomic status (SES) are more likely to attend *Gymnasia*. This association is stronger in federal states without binding teacher recommendations (Gresch et al. 2010).

In an institutional setting with binding teacher recommendations, parents cannot easily incorporate their educational aspirations into their children's educational transitions. This, in turn, could reduce SES bias in track placement, in that teacher recommendations have a stronger foundation in students' prior performance compared with parental aspiration (Dollmann 2016; Ditton et al. 2005). This should also lead to less track misplacement of students in terms of ability and stronger homogenization of students' performance within secondary educational tracks. As mentioned above, theoretical views on the effects of classroom homogenization are

[9] The *Abitur* was for some time the only way to enter university. In the 1960s, educational reforms introduced other possibilities (e.g. through professional qualification). However, these new paths, unlike the *Abitur*, ensured only limited access to university (Schindler 2017). Although the status of the *Abitur* as the only way to enter university changed, it remains the most usual way to enroll in university (Müller et al. 2011)..

ambiguous. In both views, however, the variance of performance within classes is central to the role of teacher recommendations. However, due to persistent SES bias, it is questionable whether teacher recommendations alone can sufficiently change classroom composition in order to homogenize classes by performance (Gresch et al. 2010).

The bases of teacher recommendations (which vary from one federal state to another) may impact this situation. In all states, teachers must consider students' grades in their recommendations. In some federal states, grades are the only criterion for recommendations. Other states allow teachers to base their recommendations on additional student characteristics: alongside grades, teachers may also rely on performance development as well as study and work habits. These characteristics also constitute part of grading in general (Helbig & Nikolai 2015), which means they may indirectly influence recommendations even in settings where grades are the sole basis for recommendations. However, in states where teachers may additionally rely on performance development and study and work habits, the weight of these additional characteristics is much greater in that they influence recommendations both directly and indirectly (through grades). Moreover, these additional criteria are unequally distributed among students from families of differing SES. Socially advantaged students typically display better learning and work habits (Helbig & Morar 2017), while socially disadvantaged students actually tend to receive lower grades for the same performance (Maaz & Nagy 2010). Therefore, in an environment where teachers can rely on grades along with other characteristics, recommendations should associate more strongly with SES and performance homogenization within classes should be less pronounced.

Research on freedom of choice in educational transitions in Germany has focused on educational participation, with mostly consistent results. While some studies find reduction of educational inequalities by lowering SES-specific educational decisions for secondary schools in transition settings with less freedom of choice (Dollmann 2016), most could not replicate this finding and find no significant effect of freedom of choice on SES-specific educational decisions across different federal states (Jähnen & Helbig 2015; Neugebauer 2010; Roth & Siegert 2016). These results thus fit with previous findings (Gresch et al. 2010). With regard to the relation of freedom of choice to educational performance, some results indicate increased performance with binding teacher recommendations in Germany (Esser & Relikowski 2015) and that performance equity within the transition process to secondary education increases (Esser & Hoenig 2018). This positive effect on educational performance, in a strict transition setting,

seems especially prominent for students in less ambitious educational tracks (Esser & Seuring 2020). Yet, a reanalysis and extension of the analysis from Esser and Seuring casts doubt on those results. The replication finds no effect of classroom homogeneity on students' performance (Heisig & Matthewes 2021). While these results refer to secondary education, results for performance in elementary school show that binding recommendations affect performance in different fields positively. However, with the trade-off that students feel more pressure and show lower intrinsic motivation to learn (Bach & Fischer 2020). In particular, lower levels of intrinsic motivation could lead to stagnation in performance development once the transition to the desired educational track has been achieved. Other research examining the impact of performance heterogeneity in classes on student learning in reading and mathematics finds no significant differences in learning outcomes between homogeneous and heterogeneous classes (Gröhlich et al. 2009).

For other results within Germany, the focus is not on teacher recommendations but on the timing of separating students into tracks, comparing a comprehensive and a tracked educational system. However, the mechanism driving the effect of classroom composition on performance is similar. This research shows that students benefit, to some extent, from homogenous classes, although the efficiency gains have limits in lower tracks (Matthewes 2021) and students with higher performances, who are usually in higher tracks, profit in their performance development from an early differentiation (associated with performance homogeneity in classes) (Roller & Steinberg 2020). International research in this area also finds varied effects of performance homogeneity in classes (i.e. that homogeneity disadvantages lower-performing students while benefitting higher-performing students (Van de Werfhorst, Herman G. & Mijs 2010). Results from South Africa, in accord with this, show that lower-performing students respond more strongly to peer group composition than do higher-performing students (Garlick 2018).

In a strict transition setting, the teacher recommendation is binding for secondary education. As discussed above, this could lead to homogenization in students' abilities in secondary education, since teacher recommendations should be stronger based on students' abilities, compared to parental assessment. However, prior research does not find a reduction in SES-specific educational decisions and a persistent SES-bias in *Gymnasium* attendance. Two factors may attenuate the relationship between past performance and teacher recommendation: First, teachers base their recommendation not only on students' past performance, but also on SES and

parental aspiration (Becker & Birkelbach 2013). While students' SES also bias teacher recommendations, recommendations correlate more strongly with prior performance than with parental aspiration (Ditton et al. 2005). However, this leads to SES-biased recommendations, not based fully on prior performance, due perhaps to contact between teachers and parents, which influence teacher recommendation (Barg 2013) (higher SES parents also have more contact with teachers) (Barg 2019).

Second, even in a less free transition setting, parents and students still have the option to deviate from teacher recommendations at the end of primary school. Upward deviation, while possible, typically faces barriers (e.g. students must pass a trial phase or an entrance examination) (Helbig & Nikolai 2015). Generally it is higher-SES students who exercise the possibility of upward deviation from teacher recommendations (Lohmann & Groh-Samberg 2010; Usslepp 2019). It is also possible to deviate downward if a teacher recommends a more demanding school track (downward deviation faces no barriers). Mostly lower-SES students employ downward deviation against teacher recommendations (Lohmann & Groh-Samberg 2010; Usslepp 2019). SES-specific deviations from teacher recommendation, as well as SES-specific behaviors in contact with teachers, may weaken the association between past performance and recommendations, perhaps reducing in turn the effect of teacher recommendations on homogenization of class composition by performance in secondary education. This, in turn, may hamper a possible effect of binding teacher recommendations on students' performance development.

4.3 Hypotheses

Based on the above theoretical arguments and empirical research, it seems unlikely that binding teacher recommendations would create performance-homogenous classrooms, as freedom of choice has no effect on SES-specific educational decisions and some form of SES-based bias remains for the transition to secondary education. Because teacher recommendations already partially reflect students' SES and because parents may influence recommendations through contact with teachers, as well due to an SES-specific behavior to bypass recommendations, the first hypothesis is therefore:

H1: Binding teacher recommendations have no effect on average performance.

Regarding the effect of the reform upon individual secondary tracks, the theoretical considerations and empirical findings suggest varying effects on the different educational tracks. Classrooms may be more performance-homogenous in lower secondary education because of SES-

bias in teacher recommendations, SES-specific contact, and behavior bypassing those recommendations in relation to binding teacher recommendations, which should allow lower-performing students with higher SES in particular to enter *Gymnasium* more frequently. In addition, for lower-performing students, who are particularly present in lower secondary education, the composition of the peer group is more important for the development of performance. This would suggest negative effects of binding teacher recommendations in lower secondary education and positive effects in upper secondary education. However, as noted above, we assume that binding teacher recommendations alone will not be able to effect significant changes in class composition. Therefore, we also assume no effects of binding teacher recommendations on performance in the individual school types.

H2: Binding teacher recommendations have no effect on the average performance in each of the three school types at the secondary level.

4.4 Data and educational reforms of teacher recommendation

To investigate the hypotheses, this study will use data from the Program for International Student Assessment[10] (PISA) by the Organization for Economic Co-operation and Development (OECD). Since its introduction in 2000 (Baumert et al. 2009), the PISA survey, which gathers information about the competencies of students in reading, mathematics, and science, as well as data on attitudes, demographics, social origin, and other information, has taken place every three years in several countries. In Germany, the samples consist of fifteen-year-old students, who are usually in the ninth grade. In addition, PISA conducts interviews with parents, teachers, and school management for other relevant information (Jude & Klieme 2010). Thus, the PISA data allow analysis of the effects of teacher recommendations on educational performance.

[10] PISA 2000 was designed in Germany as a national research program by the German PISA Consortium (Juergen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, Manfred Weiß). It was lead-managed by Professor Dr. Juergen Baumert, Max-Planck Institute for Human Development, Berlin. Primary research results have been published, e.g., in Baumert et al. (2001); Baumert et al. (2002, 2003)). Survey questionnaires have been documented in Kunter et al. (2002)). We thank the German PISA Consortium and the Research Data Center (*Forschungsdatenzentrum*, FDZ) in Berlin for granting permission to conduct this secondary analysis and for their support.

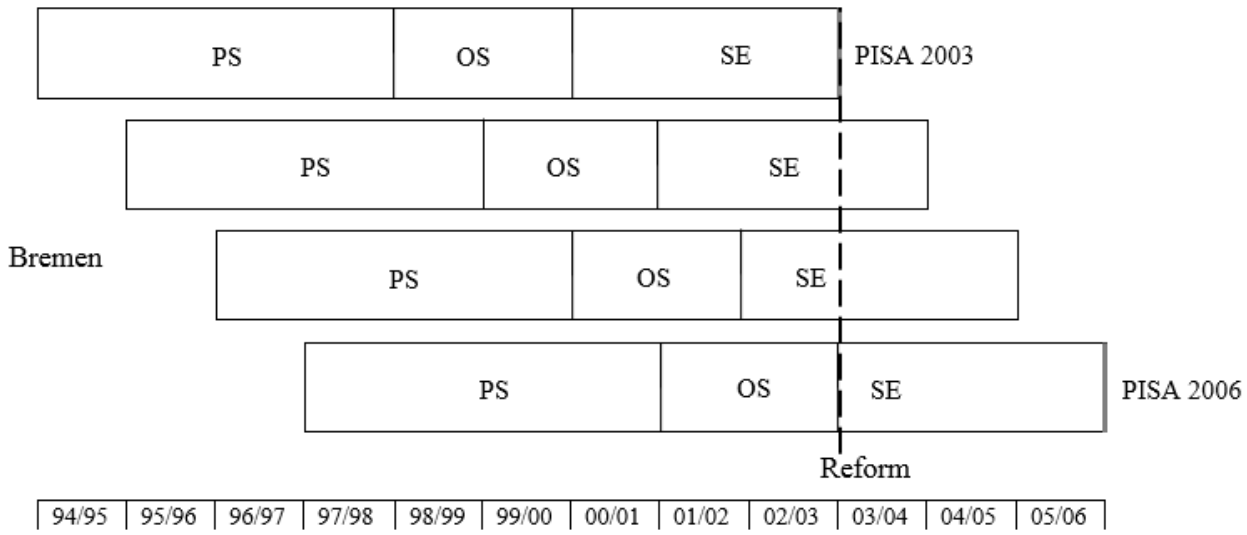
Because the focus of this study lies at the level of federal states in Germany, it is necessary to use the federal state extension to PISA, called PISA-E (replaced in 2009 by IQB-LV, created by the *Institut für Qualitätssicherung im Bildungswesen* (IQB)), which is similar to PISA-E but does not collect data for all competencies. For example, PISA-E collects data for competencies in reading comprehension, mathematics and natural sciences, while IQB-LV only collects data for a certain field of competencies (e.g. IQB-LV 2008-9: languages; IQB-LV 2012: mathematics and natural sciences). Hence, PISA-E and IQB-LV together provide continuous measures for language skills between 2000 and 2009 and for mathematics skills between 2000 and 2006.

It is important to consider the transition from PISA-E to IQB-LV for the identification of reforms suitable for analysis. As the analysis requires two measurements prior to the data reflecting each reform, these reforms' implementations must have occurred in 2005 at the latest. In the appendix, table 30 shows a number of different reforms of the binding character of teacher recommendations in Germany between 1996 and 2006 (for additional reforms and a broader period, see (Büchler 2016; Helbig & Nikolai 2015). The binding character of teacher recommendations falls into three different modes. Parents being free to decide when teacher recommendations are advisory characterize the first mode, *parental decision*. *Limited parental decision* is the second mode, in which parents are free to decide but students must pass certain thresholds of performance after specified times (e.g. the first school year) similar to a trial period. Students who do not meet their thresholds are demoted. In the third mode, *teacher decision*, the educational transitions require teacher recommendations (Büchler 2016). For analysis of the effects of teacher recommendations on educational performance, we consider only reforms between *parental decision* and *teacher decision* for a clear distinction between binding and non-binding recommendations. This restricts the number of reforms suitable to this analysis to three federal states: Bremen, North Rhine Westphalia, and Saarland. However, for North Rhine Westphalia we have only one data point collected prior to the educational reform of 1996; this federal state is therefore unsuitable for analysis. Saarland restructured its educational system, during the period under investigation, into a system with two school types in secondary education and implemented a G8 reform, shortening the time to *Abitur* by one year, and is therefore also unsuitable for analysis. That leaves only Bremen.

In Bremen in 2003, educational reform replaced free *parental decision* about secondary education at the end of sixth grade with binding teacher recommendations (although families could

still deviate from teacher recommendations if students passed a qualification test) (Helbig & Nikolai 2015). At that time, Bremen had a biennial orientation stage (*Orientierungsstufe*) between primary education (first to fourth grades) and the tracked secondary education (as of seventh grade). Transition to the tracked secondary education occurred at the end of sixth grade. Tracked secondary education comprises three school types: This means that teacher recommendations are relevant for transitions to *Realschule* and *Gymnasium*. The basis for teacher recommendations remained unaffected by the 2003 educational reform. For recommendations in Bremen, teachers could take into account other characteristics beyond grades (Bremische Bürgerschaft 2003). Another set of educational reforms in Bremen followed that of 2003, in 2004 and 2005, abolishing the orientation stage and shortening the time to the *Abitur* by one year (the so-called G8 reform (2004)) and integrating *Hauptschule* and *Realschule* into a multitrack school (2005) (Büchler 2016; Homuth 2017). However, these reforms did not affect the student population in PISA 2006; these reforms only affected students in Bremen from PISA 2009.

Figure 5: Treatment over student cohorts in Bremen



Notes: PS = primary education, OS = orientation stage (*Orientierungsstufe*), SE = tracked secondary education

With PISA-E data, it is possible to analyze the reform in Bremen. Figure 5 gives an overview of the treated (post-reform) and untreated (pre-reform) student cohorts in Bremen. For our analysis, PISA-E 2003 (Prenzel et al. 2007) and 2006 (Prenzel et al. 2010) are of special interest. Students participating in PISA 2003 (then in the ninth grade) were the last student cohort

surveyed by PISA unaffected by the reform in Bremen, as their transition to secondary education had already happened. In contrast, students surveyed in PISA 2006 were in sixth grade at that time. PISA 2006 surveys the first student cohort affected by the reform of the binding character of teacher recommendation in their transition to secondary education. Analysis of PISA data of before and after the educational reform in Bremen can therefore aid our understanding of how an educational system's institutional regulations may affect educational performance.

4.5 Analytical strategy, method, and variables

The reform in Bremen amounts to a natural experiment. The treatment is the change regarding binding teacher recommendations. The treatment group consists of students in Bremen. Because no other federal states experienced the Bremen reform, those possessing a secondary education system with three tracks, which underwent no reform of teacher recommendations or other educational reforms during the analysis period, may serve as the control group. Baden-Wuerttemberg and Schleswig Holstein fulfill these requirements and are thus the control group in the analysis. Assignment to the treatment or control group depends on places of residence and years of birth (assumed to be "as good as random").

The following section describes the variables for the analysis. We measure the outcome variable, educational performance, using students' reading and math performance. To analyze binding teacher recommendations' causal effect on educational performance, separate models will address reading and math competencies. PISA measures educational competencies for every student using five plausible values (not a direct measure of individual-level performance but an estimate of population-level performance) normally distributed, with a mean of 500 points and a standard deviation of 100. For descriptive and multivariate analysis, we need to employ plausible values 1 to 5 separately and then to average them. To calculate the standard error, we use the recommended procedure by the OECD (OECD 2009).

Other variables in the models include the specific federal state, the highest ISEI and highest ISCED in the family, students' sex, the language used at home, and an indicator for the type of school visited. The federal state wherein a student attends school is the variable for assignment into the treatment or control group. For the social background, we measure SES using the highest ISEI (International Socio-Economic Index of Occupational Status) in the family. In addi-

tion, we include a measurement for educational background, using the highest ISCED (International Standard Classification of Education) in the family. The ISCED is recoded as a dichotomous variable, indicating whether the highest ISCED in the family is lower than *Abitur*. Control variables in the models include the language spoken at home (Ruhose & Schwerdt 2016), students' sex (Legewie & DiPrete 2012; Pekkarinen 2008), and school type. Language spoken at home indicates whether German serves as the main language in students' home. School type is a set of dichotomous variables indicating whether a student visits a *Gymnasium*, *Realschule*, or *Hauptschule*. Students in comprehensive schools, special schools, Waldorf schools, and vocational schools were excluded from the analysis.

Table 24 shows the mean and standard deviation for the covariates in the analyses for the years 2000, 2003, and 2006.[11] The social background variables, parental education, and ISEI in table 24 show that Bremen has more students of lower educational background and a slightly lower average ISEI compared to the control group for 2000 and 2003. In 2006, Bremen continues to have more students of lower educational background, but the average ISEI remains the same between Bremen and the control group. Compared to the control group, Bremen has a lower share of students with German as the language used at home for all years. In Bremen and the control group, female students are overrepresented in all years. Regarding the individual school types, it appears that, in Bremen, the proportion of students in the *Gymnasium* is higher, the proportion of students in the *Realschule* does not differ between Bremen and the control group, while the proportion of students in the *Hauptschule* is higher in the control group. Looking at the covariates, some differences between Bremen and the control group become apparent. These differences, however, probably reflect the fact Bremen is a city-state, unlike the states in the control group. However, these differences are fairly constant and the following analysis controls for them.

[11] In PISA 2000 the population weights for reading and math performance are not identical. Therefore, we show separate samples based on reading and based on math, with weights bigger than zero, for the covariates in 2000.

Table 24: Descriptive statistics of reading and math performance and covariates by year

Variables	Year	Full Sample		Bremen		Control Group	
		Mean	sd	Mean	sd	Mean	sd
Parental Education (below *Abitur*)	2000 (r)	0.438	0.496	0.459	0.499	0.431	0.495
	2000 (m)	0.431	0.495	0.449	0.498	0.425	0.495
	2003	0.552	0.497	0.591	0.492	0.529	0.499
	2006	0.412	0.495	0.466	0.499	0.402	0.490
Highest ISEI in Family	2000 (r)	51.858	16.248	49.316	16.295	52.700	16.148
	2000 (m)	51.653	16.582	49.402	16.335	52.448	16.602
	2003	50.851	16.341	49.685	16.500	51.522	16.214
	2006	53.319	16.080	53.075	16.902	53.364	15.927
German at Home	2000 (r)	0.950	0.217	0.906	0.292	0.965	0.184
	2000 (m)	0.951	0.216	0.902	0.298	0.968	0.176
	2003	0.889	0.314	0.862	0.345	0.905	0.294
	2006	0.930	0.255	0.902	0.298	0.935	0.246
Male	2000 (r)	0.485	0.500	0.480	0.500	0.487	0.500
	2000 (m)	0.484	0.500	0.473	0.500	0.488	0.500
	2003	0.469	0.499	0.462	0.499	0.473	0.499
	2006	0.462	0.499	0.464	0.499	0.462	0.499
Gymnasium	2000 (r)	0.482	0.500	0.543	0.499	0.462	0.499
	2000 (m)	0.479	0.500	0.527	0.500	0.462	0.499
	2003	0.460	0.498	0.528	0.499	0.422	0.494
	2006	0.489	0.500	0.589	0.492	0.471	0.499
Realschule	2000 (r)	0.328	0.470	0.305	0.461	0.336	0.473
	2000 (m)	0.337	0.473	0.314	0.465	0.345	0.475
	2003	0.358	0.480	0.319	0.466	0.380	0.486
	2006	0.345	0.475	0.251	0.434	0.362	0.481
Hauptschule	2000 (r)	0.189	0.392	0.152	0.359	0.202	0.401
	2000 (m)	0.184	0.388	0.159	0.366	0.193	0.395
	2003	0.181	0.385	0.153	0.360	0.198	0.398
	2006	0.166	0.372	0.159	0.366	0.167	0.373
N	2000 (r)	2962		737		2225	
	2000 (m)	1563		408		1155	
	2003	4024		1469		2555	
	2006	3277		509		2768	

Notes: Data: PISA-E 2000, 2003 and 2006. 2000 (r) is the sample based on reading performance and 2000 (m) is the sample based on math performance.

To identify the reform's causal effect on performance, we use difference-in-difference (DiD) OLS-regression estimators (often used for policy evaluation) (Athey & Imbens 2017). Using DiD requires a natural experiment with at least two time periods as well as a treatment and a control group. Assignment to the treatment or control group must be essentially accidental. The basic idea of DiD is to compare the treatment group before and after an intervention to identify that intervention's effect. The control group enables differencing out pre-existing time trends. One advantage of DiD is that it does not require individual-level panel data, but works with aggregate-level data or repeated cross-sectional data such as PISA-E (Angrist & Pischke 2009; Gangl 2010; Wooldridge 2010). The following equation represents the model:

$$y_i = \beta_0 + \beta_1 S_i + \beta_2 T_i + \beta_3 (S_i * T_i) + \beta_C X_i + u_i \quad (1)$$

Equation 1 gives the DiD estimator for the effect of the reform for individual i on the educational performance y, measured by reading or mathematics competencies. The dummy variable S_i is the state indicator: it equals one for the treatment group (Bremen) and zero for the control group (Baden-Wuerttemberg and Schleswig Holstein). This variable accounts for possible time-invariant differences between both groups in the model. T_i is a dummy for the time, which equals one for the post-reform period and zero otherwise. The time dummy captures possible changes over time, which could lead to changes in y even without the reform. The interaction term between S_i and T_i equals one for Bremen in the post-reform period. β_3, the regression coefficient for this interaction effect, indicates the effect of the reform on educational performance y. X_i is a vector with control variables and u_i is the error term.

The DiD estimator requires some assumptions for identification of the causal effect. The following discusses central assumptions in relation to the analysis. The causal interpretation of the DiD estimator rests on, among other things, the parallel (or common) trend assumption, which states that, had no reform (i.e. no treatment) occurred, educational performance in both groups would have developed in parallel (Angrist & Pischke 2009; Gangl 2010). Pre-treatment data are necessary for testing this assumption: If the control and treatment groups show parallel trends in the outcome variable before the treatment, this serves as an indication the assumption holds. To test the parallel trend assumption, we perform placebo DiD analysis (Gertler et al. 2016) of the pre-reform (no reform in the treatment group) period, 2000–2003. If the parallel trend assumption holds, then the interaction term between year and treatment group should be not significant and close to zero. Table 31 in the appendix shows the results of the placebo DiD, separate for reading and mathematics competencies. Based on effect sizes and statistical

significance, they suggest that the parallel trend assumption holds for Bremen in reading and math, because both models, with and without covariates, do not display significantly different pre-reform trends for 2003 compared to 2000. Thus, the parallel trend assumption seems to hold for both competencies. However, while the point estimates for math in models 1 and 2 are close to zero, the point estimate for reading in model 2 shows a slightly more positive pre-treatment trend of 7.8 points in Bremen compared to the control group. Although this difference may indicate violation of the parallel trend assumption, it is so small (7.8 points on a scale with an SD of 100) that any biases are likely negligible in magnitude.

For the identification of the treatment effect of binding teacher recommendations on performance, it is also necessary that no other reform occur in the treatment and the control groups during the same period. Other reforms might influence the outcome otherwise and clear identification of the causal effect of binding teacher recommendations on performance would thus not be possible. As mentioned above, other reforms were implemented shortly after the reform examined here. However, subsequent reforms have not yet affected students in PISA 2006.

The DiD estimator requires the stable unit treatment value assumption (SUTVA) for unbiased identification of the treatment effect. SUTVA assumes observation of only the treated or untreated outcome. Spillovers between the treatment and the control group (e.g. internal migration into and out of Bremen) would violate SUTVA. The percentage share of the total population of Bremen of internal German out- and in-migration from and to Bremen averages approximately 3.2 and 3.3 percentage points, respectively, for the years 2002–2006 (see table 32 in the appendix). Based on these values, we assume that SUTVA holds.

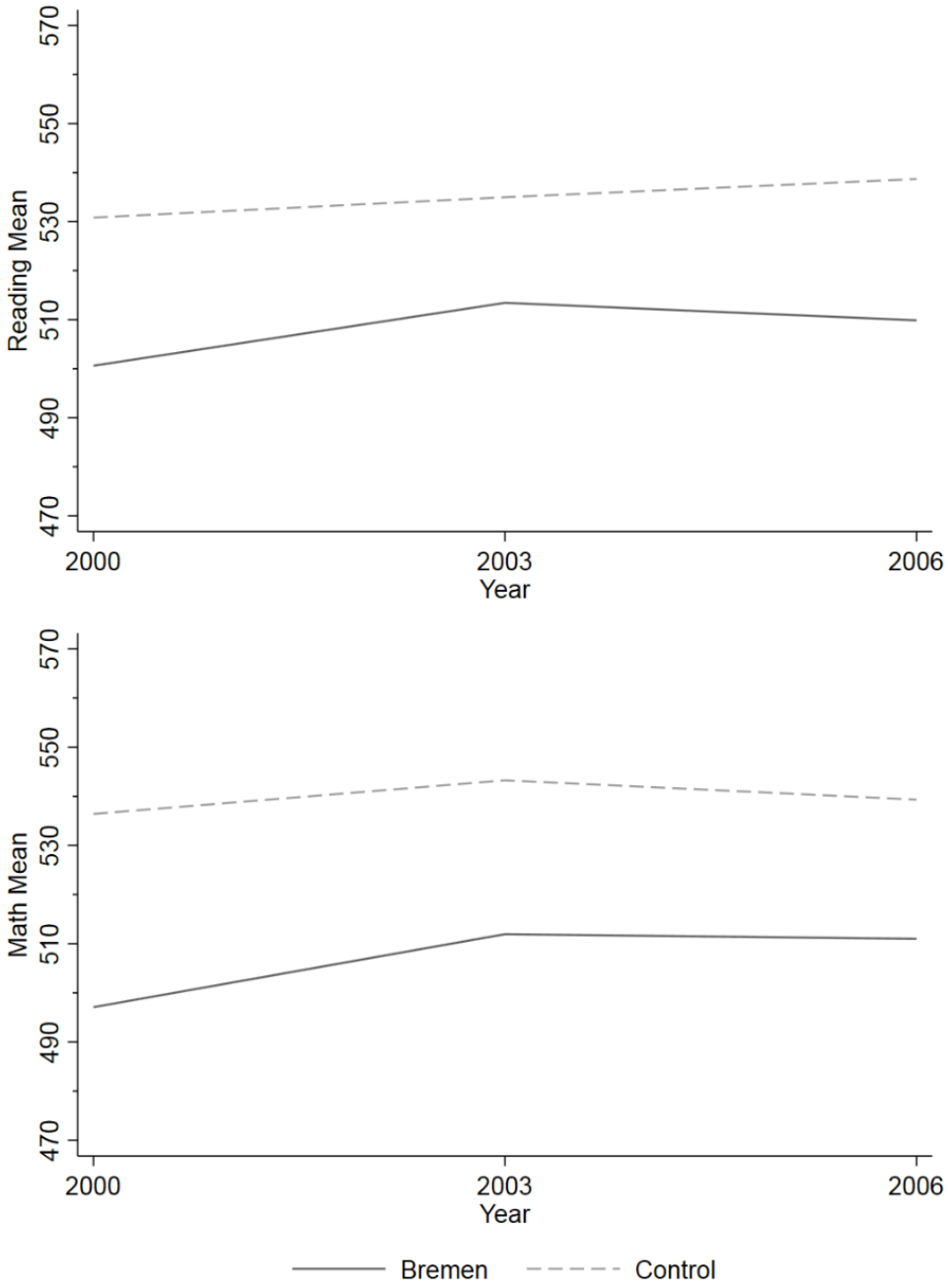
The last assumption discussed here refers to a possible reaction of the treatment group due to anticipation of the reform. We can rule out a change in behavior of the pre-reform population since those students have already made the transition to secondary education by the time of survey. However, there may yet have been a reaction to the reform in the post-reform population. As discussed earlier, parents can influence teachers through contact and there may additionally be an upward deviation from teacher recommendations. In response to the reform, this behavior could intensify, weakening the effect of binding teacher recommendations on performance. The available data do not enable controlling for these behaviors. Thus, it becomes part of the estimated reform effect. However, the time for such behavioral change is quite limited, as post-reform students are simply the first cohort affected by the reform. This means that upward deviation from teacher recommendations could especially bias the results, in that the time

to intensify contact with teachers is limited. Another possible strategy for avoiding binding teacher recommendations could be deliberate out-migration of students with recommendations incompatible with parental aspirations. While the reform year 2003 sees the highest out-migration for the period of 2002–2006, the general trend shows decreased out-migration. The largest drop in out-migration takes place in 2004–2005 (see table 32 in the appendix). Based on these numbers, it is unclear how big this problem is, but we cannot rule it out. Both of these possible biases come from one subgroup: the low-performing high SES students. That is why we assume that possible biases are small and negligible in magnitude.

4.6 Results

The following section reports the results from the analyses on the effects of binding teacher recommendations on students' average reading and math performance in Bremen. Figure 6 presents separately the mean for reading and math performance for Bremen and the control group for the years 2000–2006. In figure 6, we can see that reading performance increases for the full sample from 2000 to 2003, while reading performance decreases slightly from 2003 to 2006 in Bremen and only increases in the control group in that time period. Similarly, overall performance in math increases from 2000 to 2006. However, in Bremen it remains stable from 2003 to 2006, while decreasing slightly in the control group. In addition, the results of the placebo DiD analysis are once again graphically clear. For the time prior to treatment, reading and math performance show parallel trends between Bremen and the control group.

Figure 6: Mean of reading and math in Bremen and the control group from 2000 to 2006



Notes: Data: PISA-E 2000, 2003 and 2006.

Table 25: DiD results for the effect of binding teacher recommendations on reading and math performance from 2003 to 2006 in Bremen

	Reading		Math	
	Model 1	Model 2	Model 1	Model 2
Year 2006 X Bremen	6.213	-4.653	18.029	6.645
	(15.613)	(6.776)	(15.742)	(6.449)
Year 2006	-4.343	-2.771	-13.882	-12.022***
	(8.711)	(3.426)	(9.185)	(2.920)
Bremen	-21.239*	-30.180***	-29.970**	-40.576***
	(9.9572)	(3.933)	(10.151)	(3.639)
Parental Education (below *Abitur*)		-7.371**		-7.390**
		(3.075)		(2.577)
Highest ISEI in Family		0.316***		0.217**
		(0.104)		(0.085)
German at Home		47.219***		34.305***
		(6.263)		(5.683)
Male		-19.801***		31.418***
		(2.274)		(2.272)
Gymnasium		131.693***		145.684***
		(5.489)		(4.017)
Realschule		80.051***		80.092***
		(5.413)		(4.104)
Intercept	538.229***	407.399***	545.353***	401.910***
	(5.861)	(8.912)	(6.135)	(6.523)
N	7301	7301	7301	7301
R^2	0.002	0.485	0.008	0.533
adj. R^2	0.001	0.485	0.007	0.532

Notes: Data: PISA-E 2003, 2006. All estimations using weights. Clustered standard errors (obtained using schools as clusters) are in parentheses. Significance levels: + p < .1, * p < .05, ** p < .01, *** p < .001.

Table 25 reports the results for the effect of binding teacher recommendations on average reading and math performance. The coefficient of interest is the interaction between year and Bremen. It shows the effect of the reform on educational performance. *H1* states that the reform to binding teacher recommendations has no effect on average performance. The point estimates for reading and math go in different directions. Although there is initially a positive effect of 6.2 points for reading in model 1, with the inclusion of covariates in model 2 the point estimate switches signs and the effect becomes negative, with -4.7 points. This result suggests that binding teacher recommendations slightly decreased reading performance in Bremen. In both models, however, effects are not significant and are small, given the standard deviation of 100 of the plausible values. This confirms *H1* because binding teacher recommendations do not affect

average reading performance. The point estimates for math performance show positive effects of binding teacher recommendations in models 1 (18.0 points) and 2 (6.6 points). Neither effect is significant. While in both models we find an increase in the average math performance, with the inclusion of covariates in model 2 the increase in math performance is small, thus also confirming *H1* for mathematics performance. Binding teacher recommendations do not increase average math performance.

Table 26: DiD results for the effect of binding teacher recommendations in *Hauptschule* on reading and math performance from 2003 to 2006 in Bremen

	Reading		Math	
	Model 1	Model 2	Model 1	Model 2
Year 2006 X Bremen	-18.955	-15.662	4.692	7.531
	(14.827)	(13.479)	(12.605)	(11.735)
Year 2006	3.517	2.124	-11.875^{+}	-9.580
	(9.068)	(8.563)	(6.169)	(5.672)
Bremen	-48.299***	-49.423***	-42.948***	-41.375***
	(9.141)	(8.520)	(8.252)	(7.785)
Parental Education (below *Abitur*)		2.689		-1.282
		(6.680)		(6.700)
Highest ISEI in Family		0.375		-0.040
		(0.257)		(0.189)
German at Home		54.828***		35.583***
		(9.642)		(7.902)
Male		-21.325***		34.299***
		(5.792)		(5.358)
Intercept	441.367***	390.283***	451.101***	404.793***
	(6.004)	(17.127)	(4.227)	(13.254)
N	1274	1274	1274	1274
R^2	0.011	0.128	0.014	0.121
adj. R^2	0.009	0.123	0.012	0.116

Notes: Data: PISA-E 2003, 2006. All estimations using weights. Clustered standard errors (obtained using schools as clusters) are in parentheses. Significance levels: + p < .1, * p < .05, ** p < .01, *** p < .001.

The remaining hypothesis *H2* states that binding teacher recommendations should not affect the average performance in each of the school types at the secondary level. To test this hypothesis, we estimate individual models for each school type in secondary education. Table 26 shows the results of the models for the *Hauptschule* by competency. The results for the *Hauptschule* confirm this hypothesis. While the effect for math performance shows a small positive effect, with an increase by 7.5 points in model 2, the point estimate for reading performance displays a negative effect of -15.7 points in model 2. None of the DiD-coefficients is

significant and the increase of math and decrease of reading are each relatively small compared to the standard deviation of 100 of the plausible values. This suggests that binding teacher recommendations do not affect performance development in the *Hauptschule*.

Table 27 presents the results for the *Realschule* by competency. For the *Realschule*, we observe positive effects of binding teacher recommendations on average reading and math performance. Students in the *Realschule* increase reading performance by 8.9 points (model 2) and in math performance by 13.1 points (model 2). As with the *Hauptschule*, the DiD-coefficients for the *Realschule* are not significant and are relatively small in relation to the standard deviation of the dependent variable. These results confirm *H2*, that binding teacher recommendations do not affect students' performance in the *Realschule*.

Table 27: DiD results for the effect of binding teacher recommendations in *Realschule* on reading and math performance from 2003 to 2006 in Bremen

	Reading		Math	
	Model 1	Model 2	Model 1	Model 2
Year 2006 X Bremen	8.626	8.886	14.692	13.104
	(12.537)	(11.979)	(11.341)	(10.841)
Year 2006	-4.958	-5.605	-11.769*	-12.227*
	(6.329)	(5.731)	(5.864)	(5.427)
Bremen	-50.680***	-46.133***	-59.232***	-55.008***
	(6.012)	(5.254)	(5.613)	(5.264)
Parental Education (below *Abitur*)		-9.791**		-7.346*
		(4.054)		(3.942)
Highest ISEI in Family		0.378**		0.288**
		(0.123)		(0.115)
German at Home		46.888***		33.275**
		(8.846)		(12.180)
Male		-24.168***		32.355***
		(4.112)		(3.411)
Intercept	535.680***	489.723***	535.052***	479.409***
	(3.768)	(12.425)	(3.038)	(12.918)
N	2570	2570	2570	2570
R^2	0.013	0.108	0.024	0.126
adj. R^2	0.012	0.106	0.022	0.123

Notes: Data: PISA-E 2003, 2006. All estimations using weights. Clustered standard errors (obtained using schools as clusters) are in parentheses. Significance levels: + p < .1, * p < .05, ** p < .01, *** p < .001.

Table 28 presents the DiD results by competency for the *Gymnasium*. The results are similar to the results for the *Realschule* and *Hauptschule*, confirming *H2*. Again, the point estimates for both competencies are not significant. The point estimates for math performance in models 1 (1.4 points) and 2 (0.9 points) are quite close to zero. This shows that binding teacher recommendations have no effect on the average math performance in the *Gymnasium*. For reading performance, the point estimates in models 1 (-9.4 points) and 2 (-11.2 points) indicate a decrease in average reading performance. In relation to the standard deviation of the dependent variable, however, the DiD-coefficients are relatively small. Binding teacher recommendations seem to have no effect on average reading performance in the *Gymnasium*. All results taken together, binding teacher recommendations have no effect on student performance in general or among the different school types of secondary education.

Table 28: DiD results for the effect of binding teacher recommendations in *Gymnasium* on reading and math performance from 2003 to 2006 in Bremen

	Reading		Math	
	Model 1	Model 2	Model 1	Model 2
Year 2006 X Bremen	-9.414	-11.170	1.384	0.925
	(9.092)	(8.447)	(9.406)	(8.820)
Year 2006	-2.820	-3.821	-11.607*	-13.139**
	(4.565)	(4.669)	(4.819)	(4.535)
Bremen	-19.258***	-15.660***	-33.834***	-31.201***
	(5.142)	(4.653)	(5.844)	(5.109)
Parental Education (below *Abitur*)		-12.069**		-11.283**
		(4.949)		(4.096)
Highest ISEI in Family		0.191		0.260*
		(0.155)		(0.137)
German at Home		31.170**		37.318***
		(12.759)		(9.294)
Male		-14.538***		28.936***
		(3.471)		(3.319)
Intercept	592.530***	561.178***	605.001***	544.606***
	(2.668)	(14.670)	(3.182)	(10.783)
N	3457	3457	3457	3457
R^2	0.007	0.045	0.019	0.111
adj. R^2	0.006	0.044	0.018	0.109

Notes: Data: PISA-E 2003, 2006. All estimations using weights. Clustered standard errors (obtained using schools as clusters) are in parentheses. Significance levels: + p < .1, * p < .05, ** p < .01, *** p < .001.

4.7 Robustness check

A robustness check employing a different method serves to validate the results of the analysis further. In the initial analysis, we controlled for group-specific time trends between the treatment and the control groups. However, we did not control for state-specific time trends, which may bias the results, for example because we did not control for all possible education reforms, or other state-specific characteristics, which might influence school performance, not included in the previous analyses. To control for possible state-specific time trends, we also analyze the effect of binding teacher recommendations on average performance in reading and math using a pooled OLS model with state fixed effects and state-time interactions. This analysis approach is similar to the fixed effect with individual slopes method (with the advantage of not requiring a parallel trend assumption) (Rüttenauer & Ludwig 2020).

Table 29 shows the results of the pooled OLS regressions with state-specific time trends separately for reading and math performance for Bremen compared to Baden-Wuerttemberg. The focus here should lie on the state-time interaction for 2006. The results show no significant time trends before or after the reform for Bremen compared with Baden-Wuerttemberg. The state-time interaction for 2006 is very similar to the DiD coefficients in model 2 for reading and math in table 25. The point estimates for both competencies are small. They are negative for reading performance and positive for math performance. Thus, the robustness check confirms the results of the initial analysis. The reform to binding recommendations do not affect academic performance.

Table 29: Pooled OLS with state-specific time trends on the average performance in reading and math for Bremen

	Reading	Math
Year 2000 X Bremen	-9.679	-0.707
	(6.671)	(8.026)
Year 2006 X Bremen	-6.366	6.361
	(6.993)	(9.761)
Bremen	-31.951***	-40.951***
	(4.092)	(3.848)
Year 2000	-6.331*	-14.228***
	(3.164)	(4.142)
Year 2006	-0.908	-11.390***
	(3.888)	(3.350)
Parental Education (below *Abitur*)	-5.421*	-8.075***
	(2.331)	(2.259)
Highest ISEI in Family	0.340***	0.217**
	(0.079)	(0.071)
German at Home	48.405***	40.624***
	(5.671)	(5.596)
Male	-17.102***	31.262***
	(1.920)	(2.070)
Gymnasium	134.597***	140.017***
	(4.055)	(3.833)
Realschule	83.185***	78.909***
	(4.043)	(3.671)
Intercept	402.030***	400.422***
	(7.066)	(6.157)
N	10263	8864
R^2	0.498	0.508
adj. R^2	0.497	0.507

Notes: Data: PISA-E 2000, 2003 and 2006. All estimations using weights. Clustered standard errors (obtained using schools as clusters) are in parentheses. Besides the indicators and interactions for Bremen, federal state indicators and state-time interactions not shown. The reference for Bremen is Baden-Wuerttemberg. Significance levels: + p < .1, * p < .05, ** p < .01, *** p < .001.

4.8 Discussion

The aim of this analysis is to examine the effect of a reform to binding teacher recommendations on educational performance. This analysis, a DiD-analysis based on PISA data, therefore focuses on an educational reform in Bremen that changed non-binding recommendations to binding ones. Since other characteristics besides grades may inform recommendations, the recommendation may have a stronger SES-specific bias. Additionally, parents' behavior may influence and bypass recommendations. Previous research finds no effect of binding teacher recommendations on SES-specific educational decisions. Therefore, there should be only minor changes in classroom composition due to the reform. This leads to hypothesis 1, which assumes no effect of the reform on performance. The results confirm *H1*, with small, non-significant point estimates. Class composition influences lower-performing students more than high-performing students, suggesting school type-specific hypotheses. However, since we assume, that the reform should not change class composition, hypothesis 2 assumes no effects of the reform on performance in the different school types of secondary education. The results for the *Hauptschule*, *Realschule*, and *Gymnasium* all confirm *H2*. For these school types in secondary education, point estimates for reading and math are small or close to zero and are not significant. Overall, therefore, we conclude that simply changing non-binding recommendations for secondary education to binding ones is not an appropriate way to influence student performance in secondary education. All point estimates are relatively small, considering the standard deviation of 100 from the dependent variable, and not significant. In this respect, the findings are similar to those of (Heisig & Matthewes 2021).

The present analysis has several limitations. First, the varying characteristics of the federal states (e.g. size, population structure, and urbanity). These differences should be particularly large between city-states and territorial states. Bremen is a city-state and the two states in the control group are territorial states. The descriptive analysis revealed some differences (e.g. in family SES and German spoken at home). However, these differences were relatively small. In addition, the robustness check, which considered state-specific time trends, yielded similar results to the DiD analysis. This suggests that Bremen and the control group are still suitable for comparison. Second, in addition to the characteristics of the state, the structure of the educational system may also pose problems for analysis. Although the education systems of the states in the analysis are similar, the education system in Bremen differs from those in the states in the control group in the orientation stage. The orientation stage in Bremen is independent of

school type. This means that students in Bremen move on to the biennial orientation stage after elementary school and only afterward separate among the school types in secondary education. The placement into separate school types in secondary education thus takes place two years later than in the control states. However, this difference remains constant over the analysis period and should therefore be controlled by the DiD analysis. Third, because performance measurement by plausible values in PISA is at the population level rather than the student level, analyses at the class level are not possible. It follows that we cannot verify the hypothesized effect of the recommendation reform on class composition. It therefore remains unclear whether binding teacher recommendations homogenize classes by ability. This also means that we cannot investigate the mechanism of class composition held responsible for positive or negative changes in performance. However, the analysis by (Heisig & Matthewes 2021) shows that classroom homogeneity does not mediate the relationship between a strict transition and students' performance. Additionally, research on the effect of a strict transition setting on educational decisions did not find a significant effect (Jähnen & Helbig 2015; Neugebauer 2010; Roth & Siegert 2016), indicating that binding recommendations do not affect classroom performance composition. This makes the proposed mechanism for the effect of binding teacher recommendations on performance seem questionable. Finally, we examine only a change in the recommendation here. It would also be interesting to see the effects of different variations, for example also changing the basis of recommendations to purely grade-based.

Further research should therefore address the question of whether mandatory teacher recommendations can change the composition of performance in classes at all, as well as different designs and combinations with other aspects of recommendations, such as the basis of recommendations. Parental behavior, like contact with teachers and the SES-specific bypassing of the recommendations, also requires consideration. However, binding recommendations on their own do not seem to influence students' performance development.

4.A Appendix Chapter 4

Table 30: Reforms of teacher recommendations (P = *parental decision*, p = *limited parental decision*, T = *teacher decision*) and basis for teacher recommendations (G = only grades, O = grades and other characteristics) between 1996 and 2006 and the relation to PISA

	Reform (year)	PISA-E 00	PISA-E 03	PISA-E 06	IQB-LV 09
Baden-Wuerttem-	/	T	T	T	T
berg	/	G	G	G	G
Bavaria	/	T	T	T	T
	/	G	G	G	G
Berlin	/	p	p	p	p
	O-G (01)	O	O	G	G
Brandenburg	P-p (02)	P	P	p	p
	/	O	O	O	O
Bremen	P-T (03), T-P (05)	P	P	T	P
	/	O	O	O	O
Hamburg	/	P	P	P	P
	G-O (03)	G	G	G	O
Hesse	P-p (99)	P	P	p	p
	/	O	O	O	O
Mecklenburg	P-p (06)	P	P	P	P
Western Pomerania	/	O	O	O	O
Lower Saxony	p-P (02)	p	p	P	P
	/	O	O	O	O
North Rhine West-	T-P (96), P-T(05)	T	P	P	P
phalia	/	O	O	O	O
Rhineland Palati-	/	P	P	P	P
nate	/	G	G	G	G
Saarland	P-T(00)	P	P	T	T
	O-G(00)	O	O	G	G
Saxony	/	T	T	T	T
	/	G	G	G	G
Saxony-Anhalt	p-P (96), P-p (03), p-T(05)	p	P	P	p
	/	O	O	O	O
Schleswig Holstein	/	p	p	p	p
	/	O	O	O	O
Thuringia	/	T	T	T	T
	/	G	G	G	G

Notes: Table based on Büchler (2016) and Helbig & Nikolai (2015).

Table 31: Placebo DiD for reading and math for the period 2000–2003 in Bremen

	Reading		Math	
	Model 1	Model 2	Model 1	Model 2
Year 2003 X Bremen	2.755	7.771	-1.703	-0.014
	(15.512)	(6.490)	(15.906)	(7.786)
Year 2003	10.751	8.070**	15.323+	14.637***
	(8.404)	(2.880)	(8.426)	(3.539)
Bremen	-23.994*	-38.242***	-28.267*	-39.185***
	(11.902)	(5.179)	(12.255)	(6.877)
Parental Education (below *Abitur*)		-4.554+		-6.548*
		(2.561)		(2.799)
Highest ISEI in Family		0.417***		0.320***
		(0.082)		(0.092)
German at Home		42.693***		42.336***
		(7.564)		(6.419)
Male		-15.146***		32.103***
		(2.428)		(2.780)
Gymnasium		135.868***		135.656***
		(4.297)		(5.033)
Realschule		85.534***		77.501***
		(4.751)		(4.671)
Intercept	527.478***	390.909***	530.030***	378.801***
	(6.038)	(7.483)	(5.802)	(8.333)
N	6986	6986	5587	5587
R^2	0.006	0.494	0.011	0.485
adj. R^2	0.005	0.493	0.011	0.484

Notes: Data: PISA-E 2000, 2003. All estimations using weights. Clustered standard errors (obtained using schools as clusters) are in parentheses. Significance levels: + p < .1, * p < .05, ** p < .01, *** p < .001.

Table 32: Internal German migration in and out of Bremen from 2002 to 2006

Year	In Migration from Germany	Percentage Share at Population	Out Migration into Germany	Percentage Share at Population	Population of Bremen
2002	22552	3.41	21367	3.23	662098
2003	22295	3.36	21622	3.26	663129
2004	21980	3.31	21243	3.20	663213
2005	21586	3.25	20776	3.13	663467
2006	21988	3.31	20617	3.11	663979

Notes: Based on the numbers of the State Statistical Office in Bremen for the respective years.

5 Final discussion and conclusion

5.1 Introduction

Institutional characteristics of educational systems can influence educational inequalities in several ways. One highly debated characteristic is tracking. Different aspects of tracking, such as timing and strictness, are discussed and researched in terms of their influences on educational attainment or decision-making, as well as school performance. Some of the theoretical arguments and empirical findings on the aspects of tracking are inconclusive. Therefore, this contribution dealt with the timing and strictness of tracking in order to extend the state of research on the influences of tracking on educational attainment, decision-making, and performance in contexts where specific aspects have not been studied before or where research is outdated.

5.2 Summary of results

5.2.1 Study 1

The first study examines the relationship between comprehensive schools and *Abitur* attainment. Comprehensive schools are an alternative late-tracking school type, alongside the early-tracking schools in Germany. The analysis of this relationship uses NEPS data, which offer the advantage of oversampling comprehensive schools. As students self-select into comprehensive schools, these students differ systematically from the overall student population. It is therefore necessary to control for this selection; for this purpose, we created a sample with propensity score matching for the analysis. Propensity score matching finds comparable students from both student populations based on observed variables. Because of the matched sample, it can be analyzed whether attending a comprehensive school influences the probability of students achieving an *Abitur* and whether there are differences in this respect according to the social origin and ability of the students.

The results show that students at comprehensive schools in general have no change in the likelihood of receiving an *Abitur* compared to students at other schools. However, separate analyses for students from higher and lower social backgrounds, respectively, indicate that students from lower social backgrounds in comprehensive schools are significantly more likely compared to students from lower social backgrounds in other schools to obtain the *Abitur*. However, the probability of obtaining the *Abitur* is not different for students with a higher social background at comprehensive schools compared to other schools. Thus, comprehensive

schools can help reduce educational inequalities among students with different social origins. For students without a recommendation for the *Gymnasium*, in other words those with a lower ability, the probability of attaining the *Abitur* also increases at comprehensive schools. Additionally to social origin, comprehensive schools can also reduce inequality for different ability groups.

However, the study has some limitations. First, propensity score matching tries to deal with the problem of self-selection to allow a comparison between two different student populations, but it cannot solve this problem. Second, the validity of the results is therefore limited to students in comprehensive schools and cannot be generalized to the entire student population in Germany. Further limitations are due to the data. On the one hand, there are many missing values, especially due to non-response by parents, but they are the only source of data for some background information. Thus, out of around 9000 observations of the initial sample, about 4000 observations are lost for the analysis due to non-response. Particularly, due to the number of dropouts, statistical power suffers as a result. On the other hand, certain important information is not or is insufficiently recorded, such as performance at the beginning of lower secondary school or the distance to school.

5.2.2 Study 2

Study 2 examines an educational reform in Lower Saxony that abolished a two-year school-type-independent orientation stage in 2004, thereby prepone placement into separate school types in secondary education by two years. The aim of the study is to examine the effect of the reform, which changed the timing of tracking, on placement in a *Gymnasium* in the ninth grade. The study investigates the effect for the overall student population in Lower Saxony, by social origin, as well as for students with a low social origin and an above-average performance. The data basis are the federal state extensions of PISA, PISA-E and IQB-LV. The effect of the reform on attendance at a *Gymnasium* is estimated using difference-in-difference and difference-in-difference-in-difference models.

The analysis of the reform effect for the entire student population in Lower Saxony shows no effect of the reform on average ninth grade *Gymnasium* attendance. Also, for students with a low social origin, no effect of the reform on average *Gymnasium* attendance in the ninth grade can be found. Students with a low social origin and an above-average performance are also not affected by the reform in *Gymnasium* attendance. The results indicate that prepone tracking by

two years has no effect on average *Gymnasium* attendance. Various robustness checks, based on different methods and model specifications, also come to a similar conclusion.

In contrast to *Study 1*, there should be less of a problem of self-selection here, since the whole student population of regular schools was affected by the reform. However, unlike *Study 1*, *Study 2* cannot analyze the actual educational attainment, but only the school type students' attend in the ninth grade, i.e., at the end of lower secondary education. Thus, it remains open whether the abolition of the orientation level influenced the educational attainment. Problems with the assumptions of the difference-in-difference estimator are resolved by various robustness checks, which yielded results similar to those of the initial analysis.

5.2.3 Study 3

Study 3 has a somewhat different focus compared to the previous two studies. Here, the focus is not on the timing of tracking, but on the effect of the strictness of tracking on academic performance. Thus, this study examines another aspect of tracking. For this purpose, this study uses an educational reform in Bremen for the analysis, where in 2003 non-binding teacher recommendations were transformed into binding recommendations. The federal state extension of PISA, called PISA-E forms the data basis for this analysis, which used difference-in-difference models. The goal of the study is to examine the effect of the change from a non-binding to a binding transition recommendation on students' reading and math performance in ninth grade.

After the strictness of tracking in Bremen increased as a result of the reform, the findings for students in general show no effect on average reading and math performance. When the analyses are performed separately by school type, the same pattern emerges. The point estimates are relatively small and, in some cases, close to zero. This indicates that the change from non-binding recommendations to binding recommendations have no effect on student performance in the ninth grade.

The limitations of this study relate to the selection of the treatment and control groups and to the level of measurement of the dependent variable. First, Bremen is a city-state, unlike the states in the control group. Therefore, they may differ systematically. There are no major differences in the variables in the analysis, but these differences may be in unobserved variables. Second, the performance measure is at the population level, which means that the performance

composition of classes cannot be taken into account. However, empirical research results question the mediation of the effect of strictness of tracking on performance by class composition. The last is a restriction on the scope of validity. Only the binding character of recommendations was changed in the reform. Therefore, the results cannot be generalized to contexts where the basis of the recommendation is also changed.

5.3 Discussion, conclusion, and implications for further research

The final section of this paper discusses the results of the three studies and draws conclusions, places them in the context of previous research, and provides an outlook for further research. First discussing timing of tracking and educational attainment or decision-making and afterwards strictness of tracking and performance.

5.3.1 Timing of tracking, educational decisions, and attainment

The first and second study of this contribution both address the relationship of timing of tracking and educational decisions. Similar to previous empirical research on timing of tracking and educational decisions, different studies, namely *Study 1* and *2* produce different findings.

Study 1 finds similar results to older previous findings on comprehensive schools and educational inequality (Tillmann 1988). Comprehensive schools can still increase the likelihood of *Abitur* attainment, especially for students with a lower social background. This finding is also consistent with other research on the effect of timing of tracking. It is repeatedly shown that later tracking is particularly beneficial for students with a low social background and late tracking can reduce educational inequality (Meghir & Palme 2005). In addition, *Study 1* can also demonstrate that students with a higher social origin have no disadvantage in reaching the *Abitur* by attending a comprehensive school. This is important to show because it illustrates that a school type that is beneficial for students with a low social origin does not automatically become a disadvantage for students with a higher social background. The analyses can also show that students with low social origin and low initial performance are more likely to attain the *Abitur*. These two results also fit previous experimental findings (Tillmann 1988). *Study 1* uses recent data and modern statistical methods and shows that some of the goals associated with comprehensive schools since their inception, such as reducing educational inequality, can be achieved.

In contrast, the results of S*tudy 2* demonstrate that earlier timing of tracking due to the termination of the orientation stage in Lower Saxony does not lead to a lower probability of attending a *Gymnasium*, but rather the change in the timing of tracking has no effect on attending a *Gymnasium* in the ninth grade. This result is not consistent with the results from *Study 1*, but it is important to note that two different populations were being studied. Moreover, the results of *Study 2* do not confirm the theoretical arguments about the effects of the timing of tracking on educational decisions (Berger & Combet 2017). However, other research findings come to similar conclusions as *Study 2*. For example, in Finland, no effect of later timing of tracking was found (Pekkarinen 2008). Similarly, there are results for Switzerland that find no difference in the transition behavior of students between early and late tracking cantons (Combet 2019). What the results of *Study 2* and Switzerland have in common is that the difference in the timing of tracking is relatively small compared to other studies that find a positive effect of later tracking on transition behavior. Thus, it may be that the change in timing of tracking is not sufficient to substantially alter the uncertainties in the assessment of performance development. This is an argument that can be derived from the mechanism proposed by Berger and Combet (2017). Still, this cannot be tested here and offers possibilities for further research. However, the findings on the introduction of the orientation stage in Lower Saxony contradict this argument. Here, an increase in the educational attainment of students with a lower educational background and a decrease in the educational attainment of students with a higher educational background due to the orientation stage was found (Lange & Werder 2017).

In summary, the findings of *Study 1* and *2* on the timing of tracking are inconsistent. These studies in different contexts do not find a clear result on the effect of timing of tracking on educational decision-making or attainment. From the results of *Study 1*, it can be concluded that for students in comprehensive schools educational inequality in the attainment of the *Abitur* can be reduced without at the same time representing a disadvantage for students with a higher social background. However, it is unclear whether comprehensive schools can change the relationship between social background and *Abitur* attainment for all students. *Study 2* shows that preponing the timing of tracking by two years has no effect on ninth grade *Gymnasium* attendance.

Further research should consider the following. First, it is important to continue examining the effects of education reforms that change the timing of tracking, particularly in the wake of the introduction of new school types in secondary education in Germany. This can help to identify

missing patterns in the findings, based on which the theory can be adjusted to improve the theoretical predictions regarding the effect of timing of tracking. Secondly, the testing of the theoretical mechanisms should be more in focus of research. *Study 1* and *2* do not test these mechanisms directly, but report effects or correlations. Many other studies also do not test the mechanisms directly. In this context, further research should also examine how timing of tracking affects different parameters in the process of educational decision, for example, the assessment of future performance trajectories. Based on this, thirdly, different lengths in the change of timing of tracking should also be tested systematically for their effect on educational decisions. This offers an opportunity to refine the theoretical mechanisms and to improve the understanding of tracking.

5.3.2 Strictness of tracking and performance

The final study examines the effect of changing the strictness of tracking on students' performance. When changing from non-binding recommendations to binding ones, the study neither finds an effect on the average reading and math performance of ninth grade students in Bremen, nor finds an effect when the analysis is conducted separately by school type. The results suggest that a mere change in recommendations to stricter tracking is not an appropriate tool to improve students' performance. These results are consistent with the findings of Heisig and Matthewes (2021), who also find no effect of strict tracking on performance. Other results, however, find a positive effect of strict tracking on performance. For elementary school, a positive effect of stricter tracking on student performance is found. Here, mandatory recommendations function as an incentive for better performance (Bach & Fischer 2020). Further studies on the effect of strict tracking on performance in secondary education find a positive effect (Esser & Relikowski 2015; Esser & Seuring 2020). However, some of these studies seem to have methodological flaws (Heisig & Matthewes 2021).

For further research, the proposed mechanism of the effect of strict tracking on performance in secondary education, namely classroom performance composition, should be investigated in more detail. The mechanism is not directly tested in *Study 3*. While Heisig and Matthewes (2021) doubt this mechanism, it should be clarified whether strict tracking can indeed homogenize performance in classes at all, as hypothesized by Esser and others (Esser & Relikowski 2015; Esser & Hoenig 2018; Esser & Seuring 2020). For this, it is important that different conditions are also considered. For example, as the results from *Study 3* show, stricter tracking

with recommendations based on grades and other characteristics may have no effect on performance, but this could change if the basis for recommendation is also changed to purely grade-based. In this context, the influence of parents on teachers should not be forgotten. Further research should investigate whether parents change their behavior regarding teachers or the use of options to bypass recommendations after a reform of the strictness of tracking. Parental influence on teachers or the use of those options could counteract possible effects of binding recommendations on performance composition within classrooms and educational performance.

References

Abadie, A., 2021: Using Synthetic Controls. Feasibility, Data Requirements, and Methodological Aspects. Journal of Economic Literature 59: 391–425.

Abadie, A., A. Diamond & J. Hainmueller, 2010: Synthetic Control Methods for Comparative Case Studies. Estimating the Effect of California's Tobacco Control Program. Journal of the American Statistical Association 105: 493–505.

Allmendinger, J., 1989: Educational systems and labor market outcomes. European Sociological Review 5: 231–250.

Angrist, J.D. & V. Lavy, 1999: Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. The Quarterly Journal of Economics 114: 533–575.

Angrist, J.D. & J.-S. Pischke, 2009: Mostly harmless econometrics. An empiricist's companion. Princeton, NJ: Princeton Univ. Press.

Athey, S. & G.W. Imbens, 2017: The State of Applied Econometrics. Causality and Policy Evaluation. Journal of Economic Perspectives 31: 3–32.

Autorengruppe Bildungsberichterstattung, 2010: Bildung in Deutschland 2010. Ein indikatorengestützter Bericht mit einer Analyse zu Perspektiven des Bildungswesens im demografischen Wandel. Bielefeld.

Autorengruppe Bildungsberichterstattung, 2018: Bildung in Deutschland 2018. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration. Bielefeld.

Autorengruppe Bildungsberichterstattung, 2020: Bildung in Deutschland 2020. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung in einer digitalisierten Welt. Bielefeld: wbv Publikation.

Bach, M. & M. Fischer, 2020: Understanding the Response to High-Stakes Incentives in Primary Education. ZEW Discussion Papers 20-066. Mannheim.

Baier, T., V. Lang, M. Grätz, K.J. Barclay, D.C. Conley, C.T. Dawes, T. Laidley & T.H. Lyngstad, 2022: Genetic Influences on Educational Achievement in Cross-National Perspective. European Sociological Review.

Barg, K., 2013: The Influence of Students' Social Background and Parental Involvement on Teachers' School Track Choices. Reasons and Consequences. European Sociological Review 29: 565–579.

Barg, K., 2019: Why are middle-class parents more involved in school than working-class parents? Research in Social Stratification and Mobility 59: 14–24.

Bauer, P. & R.T. 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 & M. Weiß, 2002: PISA 2000 — Die Länder der Bundesrepublik Deutschland im Vergleich. Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J., C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiß, 2003: PISA 2000 — Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Baumert, J., E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß, 2001: PISA 2000. Wiesbaden: VS Verlag für Sozialwissenschaften.

Becker, D. & K. Birkelbach, 2013: Lehrer als Gatekeeper? Eine theoriegeleitete Annäherung an Determinanten und Folgen prognostischer Lehrerurteile. S. 207–237 in: R. Becker & A. Schulze (Hrsg.), Bildungskontexte. Strukturelle Voraussetzungen und Ursachen ungleicher Bildungschancen. Wiesbaden, Wiesbaden: Springer VS.

Becker, M., M. Neumann & H. Dumont, 2017: Recent Developments in School Tracking Practices in Germany. An Overview and Outlook on Future Trends. ORBIS SCHOLAE 10: 9–25.

Below, S., 2002: Bildungssysteme und soziale Ungleichheit. Das Beispiel der neuen Bundesländer. Wiesbaden: VS Verlag für Sozialwissenschaften.

Berger, J. & B. Combet, 2017: Late School Tracking, Less Class Bias in Educational Decision-Making? The Uncertainty Reduction Mechanism and Its Experimental Testing. European Sociological Review 33: 124–136.

Betts, J., 2011: The Economics of Tracking in Education. S. 341–381 in: E.A. Hanushek, S. Machin & L. Wössmann (Hrsg.), Handbook of the economics of education. Amsterdam: North-Holland.

Birkelbach, K., 2010: Lehrerurteile in der Leistungsgesellschaft. Ergebnisse einer Längsschnittstudie (1969-1997). S. 107–125 in: K. Birkelbach, A. Bolder & K. Düsseldorff (Hrsg.), Berufliche Bildung in Zeiten des Wandels. Festschrift für Rolf Dobischat zum 60. Geburtstag. Baltmannsweiler: Schneider-Verl. Hohengehren.

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

Boudon, R., 1974: Education, opportunity, and social inequality. Changing prospects in Western society. New York, NY: Wiley.

Breen, R. & J.H. Goldthorpe, 1997: Explaining Educational Differentials. Towards A Formal Rational Action Theory. Rationality and Society 9: 275–305.

Breen, R. & J.O. Jonsson, 2005: Inequality of Opportunity in Comparative Perspective. Recent Research on Educational Attainment and Social Mobility. Annual Review of Sociology 31: 223–243.

Breen, R., R. Luijkx, W. Muller & R. Pollak, 2010: Long-term Trends in Educational Inequality in Europe: Class Inequalities and Gender Differences. European Sociological Review 26: 31–48.

Bremische Bürgerschaft, 2003: Gesetz zur Änderung des Bremischen Schulgesetzes. Drucksache 15/1427.

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

Büchler, T., 2016: Schulstruktur und Bildungsungleichheit. Die Bedeutung von bundeslandspezifischen Unterschieden beim UEbergang in die Sekundarstufe I für den Bildungserfolg. Koeln Z Soziol (KZfSS Koelner Zeitschrift für Soziologie und Sozialpsychologie) 68: 53–87.

Bygren, M., 2016: Ability Grouping's Effects on Grades and the Attainment of Higher Education. Sociology of Education 89: 118–136.

Caliendo, M. & S. Kopeinig, 2008: Some Practical Guidance for the Implementation of Propensity Score Matching. Journal of Economic Surveys 22: 31–72.

Card, D., 1999: The Causal Effect of Education on Earnings. S. 1801–1863 in: O.C. Ashenfelter & D. Card (Hrsg.), Handbook of Labor Economics. Volume 3A: Elsevier.

Cerulli, G., 2019: A flexible Synthetic Control Method for modeling policy evaluation. Economics Letters 182: 40–44.

Combet, B., 2019: The Institutional Dimension of Class-based Educational Decision-making. Evidence from Regional Variation in Switzerland. Zeitschrift für Soziologie 48: 301–320.

Cordero, J.M., V. Cristobal & D. Santin, 2018: Causal Inference on Education Policies. A Survey of Empirical Studies Using PISA, TIMSS and PIRLS. Journal of Economic Surveys 32: 878–915.

Deutscher Bildungsrat, 1969: Einrichtung von Schulversuchen mit Gesamtschulen: Verabgeschiedet auf des 19. Sitzung der Bildunskommission am 30-31 Januar 1969: Bundesdruckerei.

Ditton, H., J. Krüsken & M. Schauenberg, 2005: Bildungsungleichheit — der Beitrag von Familie und Schule. Zeitschrift für Erziehungswissenschaft 8: 285–304.

Dollmann, J., 2016: Less Choice, Less Inequality? A Natural Experiment on Social and Ethnic Differences in Educational Decision-Making. European Sociological Review 32: 203–215.

Duflo, E., P. Dupas & M. Kremer, 2011: Peer Effects, Teacher Incentives, and the Impact of Tracking. Evidence from a Randomized Evaluation in Kenya. American Economic Review 101: 1739–1774.

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

Eckhardt, T., 2017: The Education System in the Federal Republic of Germany 2014/2015. A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe. Bonn.

Erikson, R. & J.O. Jonsson, 1996: Explaining Class Inequality in Education: The Swedish Test Case. S. 1–63 in: R. Erikson & J.O. Jonsson (Hrsg.), Can education be equalized? The Swedish case in comparative perspective. Boulder, Colo.: Westview Press.

Esser, H., 2016: Bildungssysteme und ethnische Bildungsungleichheiten. S. 331–396 in: C. Diehl, C. Hunkler & C. Kristen (Hrsg.), Ethnische Ungleichheiten im Bildungsverlauf. Wiesbaden: Springer Fachmedien Wiesbaden.

Esser, H. & K. Hoenig, 2018: Leistungsgerechtigkeit und Bildungsungleichheit. KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 70: 419–447.

Esser, H. & I. Relikowski, 2015: Is Ability Tracking (Really) Responsible for Educational Inequalities in Achievement? A Comparison between the Country States Bavaria and Hesse in Germany. IZA DP 9082. Bonn.

Esser, H. & J. Seuring, 2020: Kognitive Homogenisierung, schulische Leistungen und soziale Bildungsungleichheit. Zeitschrift für Soziologie 49: 277–301.

Fan, X. & M. Chen, 2001: Parental Involvement and Students' Academic Achievement. A Meta-Analysis. Educational Psychology Review 13: 1–22.

Figlio, D.N. & M.E. Page, 2002: School Choice and the Distributional Effects of Ability Tracking. Does Separation Increase Inequality? Journal of Urban Economics 51: 497–514.

Gamoran, A., M. Nystrand, M. Berends & P.C. LePore, 1995: An Organizational Analysis of the Effects of Ability Grouping. American Educational Research Journal 32: 687.

Gangl, M., 2010: Causal Inference in Sociological Research. Annual Review of Sociology 36: 21–47.

Gangl, M., 2015: Matching estimators for treatment effects. S. 251–276 in: H. Best & C. Wolf (Hrsg.), The Sage handbook of regression analysis and causal inference. Los Angeles, Calif., London, New Delhi, Singapore, Washington DC: Sage Reference.

Garlick, R., 2018: Academic Peer Effects with Different Group Assignment Policies. Residential Tracking versus Random Assignment. American Economic Journal: Applied Economics 10: 345–369.

Gertler, P.J., S. Martinez, P. Premand, L.B. Rawlings & C.M.J. Vermeersch, 2016: Impact Evaluation in Practice, Second Edition: Washington, DC: Inter-American Development Bank and World Bank.

Grätz, M. & R. Pollak, 2016: Legacies of the past: social origin, educational attainment and labour-market outcomes in Germany. S. 34–48 in: F. Bernardi & G. Ballarino (Hrsg.), Education, Occupation and Social Origin: Edward Elgar Publishing.

Grätz, M. & Ø.N. Wiborg, 2020: Reinforcing at the Top or Compensating at the Bottom? Family Background and Academic Performance in Germany, Norway, and the United States. European Sociological Review 36: 381–394.

Gresch, C., J. Baumert & K. Maaz, 2010: Empfehlungsstatus, Übergangsempfehlung und der Wechsel in die Sekundarstufe I. Bildungsentscheidungen und soziale Ungleichheit. S. 230–256 in: J. Baumert, K. Maaz & U. Trautwein (Hrsg.), Bildungsentscheidungen. Zeitschrift für Erziehungswissenschaft Sonderheft 12. Wiesbaden: VS Verlag für Sozialwissenschaften / Springer Fachmedien Wiesbaden GmbH Wiesbaden.

Gröhlich, C., K. Scharenberg & W. Bos, 2009: Wirkt sich Leistungsheterogenität in Schulklassen auf den individuellen Lernerfolg in der Sekundarstufe aus? Journal for Educational Research Online 1: 86–105.

Hanushek, E.A. & L. Wössmann, 2006: Does Educational Tracking Affect Performance and Inequality? Differences- in-Differences Evidence Across Countries. The Economic Journal 116: C63-C76.

Hanushek, E.A. & L. Wössmann, 2011: The Economics of International Differences in Educational Achievement. S. 89–200 in: E.A. Hanushek, S. Machin & L. Wössmann (Hrsg.), Handbook of the economics of education. Volume 3. Amsterdam: North-Holland.

Heisig, J.P., B. Elbers & H. Solga, 2020: Cross-national differences in social background effects on educational attainment and achievement: absolute vs. relative inequalities and the role of education systems. Compare: A Journal of Comparative and International Education 50: 165–184.

Heisig, J.P. & S.H. Matthewes, 2021: No evidence for positive effects of strict tracking and cognitive homogenization on student performance: A critical reanalysis of Esser and Seuring (2020).

Helbig, M. & T. Morar, 2017: Warum Lehrkräfte sozial ungleich bewerten: Ein Plädoyer für die Etablierung tertiärer Herkunftseffekte im werterwartungstheoretischen Standardmodell der Bildungsforschung. http://hdl.handle.net/10419/173280 (30.4.2021).

Helbig, M. & R. Nikolai, 2015: Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundeslaendern seit 1949. Bad Heilbrunn: Verlag Julius Klinkhardt.

Henninges, M., C. Traini & C. Kleinert, 2019: Tracking and Sorting in the German Educational System. LIfBi Working Paper 83. Bamberg.

Holland, P.W., 1986: Statistics and Causal Inference. Journal of the American Statistical Association 81: 945–960.

Homuth, C., 2017: Die G8-Reform in Deutschland. Wiesbaden: Springer Fachmedien Wiesbaden.

Horn, D., 2013: Diverging performances. The detrimental effects of early educational selection on equality of opportunity in Hungary. Research in Social Stratification and Mobility 32: 25–43.

IEA Data Processing and Research Center, 2010: Methodenbericht. NEPS Startkohort 4. Haupterhebung - Herbst/Winter 2010.

Jackson, M., R. Erikson, J.H. Goldthorpe & M. Yaish, 2007: Primary and Secondary Effects in Class Differentials in Educational Attainment. Acta Sociologica 50: 211–229.

Jähnen, S. & M. Helbig, 2015: Der Einfluss schulrechtlicher Reformen auf Bildungsungleichheiten zwischen den deutschen Bundesländern. KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 67: 539–571.

Jakubowski, M., H.A. Patrinos, E.E. Porta & J. Wisniewski, 2016: The effects of delaying tracking in secondary school. Evidence from the 1999 education reform in Poland. Education Economics 24: 557–572.

Jude, N. & E. Klieme, 2010: Das Programme for International Student Assessment (PISA). S. 11–21 in: E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Koeller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), PISA 2009. Bilanz nach einem Jahrzehnt. Muenster: Waxmann.

Kahneman, D. & A. Tversky, 1979: Prospect Theory: An Analysis of Decision under Risk. Econometrica 47: 263.

Klein, M., K. Barg & M. Kühhirt, 2019: Inequality of Educational Opportunity in East and West Germany: Convergence or Continued Differences? Sociological Science 6: 1–26.

Köller, O., 2008: Gesamtschule. Erweiterung statt Alternative. S. 437–465 in: K.S. Cortina, J. Baumert, A. Leschinsky, K.U. Mayer & L. Trommer (Hrsg.), Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick ; [der neue Bericht des Max-Planck-Instituts für Bildungsforschung]. Reinbek bei Hamburg: Rowohlt-Taschenbuch-Verl.

Köller, O., M. Knigge & B. Tesch, 2011: IQB Laendervergleich Sprachen 2008/2009 (IQB-LV 2008-9). http://doi.org/10.5159/IQB_LV_2008_v1 (18.12.2017).

Korthals, R.A. & J. Dronkers, 2016: Selection on performance and tracking. Applied Economics 48: 2836–2851.

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

Lange, S. & M. von Werder, 2017: Tracking and the intergenerational transmission of education. Evidence from a natural experiment. Economics of Education Review 61: 59–78.

Lavrijsen, J. & I. Nicaise, 2015: New empirical evidence on the effect of educational tracking on social inequalities in reading achievement. European Educational Research Journal 14: 206–221.

Lavrijsen, J. & I. Nicaise, 2016: Educational tracking, inequality and performance: New evidence from a differences-in-differences technique. Research in Comparative and International Education 11: 334–349.

Le Donné, N., 2014: European Variations in Socioeconomic Inequalities in Students' Cognitive Achievement. The Role of Educational Policies. European Sociological Review 30: 329–343.

Lechner, M., 2011: The Estimation of Causal Effects by Difference-in-Difference Methods. FNT in Econometrics (Foundations and Trends in Econometrics) 4: 165–224.

Lee, B., 2014: The influence of school tracking systems on educational expectations. A comparative study of Austria and Italy. Comparative Education 50: 206–228.

Legewie, J. & T.A. DiPrete, 2012: School Context and the Gender Gap in Educational Achievement. American Sociological Review 77: 463–485.

Leibniz Institute for Educational Trajectories (LIfBi), 2017: Starting Cohort 4: Grade 9 (SC4), Study Overview, Waves 1 to 9. Bamberg.

Lenski, A.E., M. Hecht, C. Penk, F. Milles, M. Mezger, P. Heitmann, P. Stanat & H.A. Pant, 2016: IQB-Ländervergleich 2012. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. http://dx.doi.org/10.18452/3125.

Leschinsky, A. & K.U. Mayer, 1999: Comprehensive Schools and Inequality of Opportunity in the Feder Republic of Germany. S. 13–39 in: A. Leschinsky & K.U. Mayer (Hrsg.), The comprehensive school experiment revisited. Evidence from Western Europe. Frankfurt am Main, New York: P. Lang.

Leuven, E., H. Oosterbeek & M. Rønning, 2008: Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway*. Scandinavian Journal of Economics 110: 663–693.

Leuven, E., E. Plug & M. Rønning, 2016: Education and cancer risk. Labour Economics 43: 106–121.

Leuven, E. & B. Sianesi, 2003: PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.

Lohmann, H. & O. Groh-Samberg, 2010: Akzeptanz von Grundschulempfehlungen und Auswirkungen auf den weiteren Bildungsverlauf / Acceptance of Secondary School Track Recommendations and Their Effects on Educational Achievement. Zeitschrift für Soziologie 39.

Luplow, N. & T. Schneider, 2014: Nutzung und Effektivitaet privat bezahlter Nachhilfe im Primarbereich / Social Selectivity and Effectiveness of Private Tutoring among Elementary School Children in Germany. Zeitschrift für Soziologie 43: 466.

Maaz, K., 2006: Soziale Herkunft und Hochschulzugang. Wiesbaden: VS Verlag für Sozialwissenschaften.

Maaz, K. & G. Nagy, 2010: Der Übergang von der Grundschule in die weiterführenden Schulen des Sekundarschulsystems: Definitionen, Spezifikation und Quantifizierung primärer und sekundärer Herkunftseffekte. S. 153–182 in: K. Maaz, J. Baumert, C. Gresch & N.

McElvany (Hrsg.), Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten. Bonn: Bundesministerium für Bildung und Forschung (BMBF) Referat Bildungsforschung.

Maaz, K., U. Trautwein, O. Luedtke & J. Baumert, 2008: Educational Transitions and Differential Learning Environments. How Explicit Between-School Tracking Contributes to Social Inequality in Educational Outcomes. Child Development Perspectives 2: 99–106.

Malecki, A., C. Schneider, S. Vogel & M. Wolters, 2014: Schulen auf einen Blick. Wiesbaden.

Matthewes, S.H., 2021: Better Together? Heterogeneous Effects of Tracking on Student Achievement. The Economic Journal 131: 1269–1307.

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

Mühlenweg, A., 2008: Educational Effects of Alternative Secondary School Tracking Regimes in Germany. Schmollers Jahrbuch 128: 351–379.

Müller, W., R. Pollak, D. Reimer & S. Schindler, 2011: Hochschulbildung und soziale Ungleichheit. S. 289–327 in: R. Becker (Hrsg.), Lehrbuch der Bildungssoziologie. Wiesbaden: VS Verlag für Sozialwissenschaften.

Negri, F. de, R. Galiezz, P. Miranda, P. Koeller, G. Zucoloto, J. Costa, C.M. Farias, G.H. Travassos & R.A. Medronho, 2021: Socioeconomic factors and the probability of death by Covid-19 in Brazil. Journal of public health (Oxford, England) 43: 493–498.

Neugebauer, M., 2010: Bildungsungleichheit und Grundschulempfehlung beim Übergang auf das Gymnasium. Eine Dekomposition primärer und sekundärer Herkunftseffekte / Educational Inequality and Teacher Recommendations at the Transition to Upper Secondary School: A Decomposition of Primary and Secondary Effects of Social Origin. Zeitschrift für Soziologie 39: 445.

Neugebauer, M., D. Reimer, S. Schindler & V. Stocké, 2013: Inequality in Transitions to Secondary School and Tertiary Education in Germany. S. 56–88 in: M. Jackson (Hrsg.), Determined to Succeed? Performance versus Choice in Educational Attainment. Palo Alto: Stanford University Press.

Niedersächsisches Kultusministerium, 2006: Die niedersächsischen allgemein bildenden Schulen in Zahlen. Stand: Schuljahr 2005/2006. Hannover.

OECD, 2009: PISA data analysis manual. SPSS. Paris: OECD.

OECD, 2012: PISA 2009 Technical Report. Paris: OECD Publishing.

Pant, H.A., P. Stanat, M. Hecht, P. Heitmann, M. Jansen, A.E. Lenski, C. Penk, C. Poehlmann, A. Roppelt, U. Schroeders & T. Siegle, 2015: IQB-Ländervergleich Mathematik und Naturwissenschaften 2012 (IQB-LV 2012). Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2012_v4.

Pekkarinen, T., 2008: Gender Differences in Educational Attainment. Evidence on the Role of Tracking from a Finnish Quasi-experiment. Scandinavian Journal of Economics 110: 807–825.

Pfeffer, F.T., 2008: Persistent Inequality in Educational Attainment and its Institutional Context. European Sociological Review 24: 543–565.

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

Prenzel, M., C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun, 2010: Programme for International Student Assessment 2006 (PISA 2006). Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2006_v1 (18.12.2017).

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

Psacharopoulos, G. & H.A. Patrinos, 2004: Returns to investment in education: a further update. Education Economics 12: 111–134.

Raudenbush, S.W. & R.D. Eschmann, 2015: Does Schooling Increase or Reduce Social Inequality? Annual Review of Sociology 41: 443–470.

Roller, M. & D. Steinberg, 2020: The distributional effects of early school stratification - nonparametric evidence from Germany. European Economic Review 125: 103422.

Rosenbaum, P.R. & D.B. Rubin, 1985: Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician 39: 33–38.

Roth, T., 2019: Welchen Einfluss hat die Schulzeitverkürzung am Gymnasium (G8) auf das Ausmaß der sozialen Ungleichheit beim Besuch der gymnasialen Oberstufe? Zeitschrift für Erziehungswissenschaft 22: 1247–1265.

Roth, T. & M. Siegert, 2016: Does the Selectivity of an Educational System Affect Social Inequality in Educational Attainment? Empirical Findings for the Transition from Primary to Secondary Level in Germany. European Sociological Review 32: 779–791.

Rubin, D.B., 1974: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66: 688–701.

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

Rüttenauer, T. & V. Ludwig, 2020: Fixed Effects Individual Slopes. Accounting and Testing for Heterogeneous Effects in Panel Data or Other Multilevel Models. Sociological Methods & Research: 004912412092621.

Sacerdote, B., 2011: Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? S. 249–277 in: E.A. Hanushek, S. Machin & L. Wössmann (Hrsg.), Handbook of the economics of education. Amsterdam: North-Holland.

Sachse, K., J. Kretschmann, A. Kocaj, O. Koeller, M. Knigge & B. Tesch, 2012: IQB-Laendervergleich 2008/2009. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente.

Scharenberg, K., 2012: Leistungsheterogenität und Kompetenzentwicklung. Zur Relevanz klassenbezogener Kompositionsmerkmale im Rahmen der KESS-Studie. Münster, New York, NY, München, Berlin: Waxmann.

Scharf, J., M. Becker, S.E. Stallasch, M. Neumann & K. Maaz, 2020: Primäre und sekundäre Herkunftseffekte über den Verlauf der Sekundarstufe: Eine Dekomposition an drei Bildungsübergängen. Zeitschrift für Erziehungswissenschaft 23: 1251–1282.

Schindler, S., 2017: School tracking, educational mobility and inequality in German secondary education. Developments across cohorts. European Societies 19: 28–48.

Schindler, S., E. Bar-Haim, C. Barone, J.F. Birkelund, V. Boliver, Q. Capsada-Munsech, J. Erola, M. Facchini, Y. Feniger, L. Heiskala, E. Herbaut, M. Ichou, K.B. Karlson, C. Kleinert, D. Reimer, C. Traini, M. Triventi & L.-A. Vallet, 2021: Educational tracking and long-term outcomes by social origin: Seven countries in comparison. DIAL Working Paper Series 02.

Schuchart, C., 2003: Die Bedeutung der Niedersaechsichen Orientierungsstufe für den Ausgleich sozialer Disparitaeten in der Bildungsbeteiligung. Zeitschrift für Erziehungswissenschaft 6: 403–420.

Schuchart, C., 2006: Orientierungsstufe und Bildungschancen. Eine Evaluationsstudie. Münster, New York, München, Berlin: Waxmann.

Schuchart, C. & H. Weishaupt, 2004: Die prognostische Qualität der Übergangsempfehlungen der niedersächsischen Orientierungsstufe. Zeitschrift für Pädagogik 50: 882–902.

Schütz, G., H.W. Ursprung & L. Wössmann, 2008: Education Policy and Equality of Opportunity. Kyklos 61: 279–308.

Stocké, V., 2007: Explaining Educational Decision and Effects of Families' Social Class Position: An Empirical Test of the Breen Goldthorpe Model of Educational Attainment. European Sociological Review 23: 505–519.

Terrin, É. & M. Triventi, 2022: The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis. Review of Educational Research: 003465432211008.

Tillmann, K.-J., 1988: Comprehensive schools and traditional education in the Federal Republic of Germany. International Journal of Educational Research 12: 471–480.

Usslepp, N., 2019: Individuelles Entscheidungsverhalten mit Blick auf Bildungsentscheidungen im Kontext des deutschsprachigen Bildungssystems. Dissertation. Tübingen.

Van de Werfhorst, Herman G., 2019: Early Tracking and Social Inequality in Educational Attainment: Educational Reforms in 21 European Countries. American Journal of Education 126: 65–99.

Van de Werfhorst, Herman G. & J.J. Mijs, 2010: Achievement Inequality and the Institutional Structure of Educational Systems. A Comparative Perspective. Annual Review of Sociology 36: 407–428.

van Elk, R., M. van der Steeg & D. Webbink, 2011: Does the timing of tracking affect higher education completion? Economics of Education Review 30: 1009–1021.

Webbink, D., 2005: Causal Effects in Education. Journal of Economic Surveys 19: 535–560.

Wenzler, I., 2003: Bundesrepublik Deutschland. Die Gesamtschule: Kräfte und Gegenkräfte im bildungspolitischen Konflikt. S. 65–86 in: H.-G. Herrlitz, D. Weiland & K. Winkel (Hrsg.), Die Gesamtschule. Geschichte, internationale Vergleiche, pädagogische Konzepte und politische Perspektiven. Weinheim: Juventa-Verl.

Wooldridge, J.M., 2010: Econometric analysis of cross section and panel data. Cambridge, Mass.: MIT Press.

Curriculum Vitae

Contact

Department of Social Sciences
TU Kaiserslautern
Erwin-Schroedinger-Str. 57
D-67663 Kaiserslautern
Phone: +49 (0) 631 205 5778
Email: kolbe@sowi.uni-kl.de

Academic Positions

Apr 2016 - Sep 2022	Research Assistant, Chair of Sociology and Social Stratification (Prof. Dr. H. Best). TU Kaiserslautern

Education

Aug 2017	Causal Inference. Summer School in Social Science Data Analysis. University of Essex
2016	Master of Arts, Sociology. Leipzig University
2012	Bachelor of Arts, Sociology. University of Mannheim
Oct 2011 - Feb 2012	Semester abroad. University of Ljubljana
2008	Abitur. Waldorf School Maschsee Hannover