Stockholm School of Economics
Department of Economics
5350 Master's thesis in economics
Academic year 2019–2020

# Grading practices and secondary school track choice: Evidence from a German policy reform

Leon Reich (41425)

Abstract: Many school systems across the world track their students by ability. The German school system tracks earlier than most and each track leads to very different academic degrees and labor market opportunities. This increases the relevance of educational policy at the elementary school level. This thesis exploits a policy reform from the 1970s and 1980s, in which a number of German states postponed the assignment of number grades to begin only with Grade 3. Using a difference-in-differences (DD) approach and data from the German Socio-Economic Panel, I analyse the effects of later grading on pupils' propensity to pursue either of the three secondary school tracks. I consistently fail to reject zero average effects on the degrees obtained. Considering treatment effect heterogeneity between genders, I find that later graded males are around 6 percentage points more likely to obtain either of the higher degrees relative to their earlier graded peers. I find no evidence of a treatment effect on females. Pupils from educated households are around 3.3 percentage points more likely to obtain the lowest degree, while pupils from low-educated households are around 3.6 percentage points less likely to obtain the lowest degree, compared to their earlier graded peers.

Keywords: school policy, grades, educational attainment, tracking, difference-in-differences

JEL: I21, I24, I28, J16, J24

| | |
|---|---|
| Supervisor: | Abhijeet Singh |
| Date submitted: | May 15, 2020 |
| Date discussed: | May 26, 2020 |
| Discussant: | Alexander Campbell |
| Examiner: | Andreea Enache |

**Acknowledgements**

I would like to thank my supervisor Abhijeet Singh for his invaluable guidance and input throughout the project. I would also like to thank my fellow students and seminar participants for their comments and insightful discussions. All of the support is greatly appreciated. All errors are my own.

# Contents

## List of Tables

## List of Figures

# 1 Introduction

Many public school systems around the world divide their students by ability, into different classrooms within a school or into different schools entirely. Proponents often argue that it is efficient to group students by ability, as this allows teachers to teach at the groups' appropriate level (cf. Hallinan, 1994; Duflo, 2001; Brunello and Giannini, 2004). Opponents instead focus on equity. They argue that tracking constrains pupils placed in the lower tracks to lower educational attainment and lower lifelong earnings, and thus aggravates socio-economic and educational inequality (Hanushek and Woessmann, 2005; Brunello and Checchi, 2007; Van Elk et al., 2011; van de Werfhorst, 2018). The risk of "misplacement" because of misjudged ability or even relative age effects is often considered to be particularly high if the tracking decision is made early in a pupil's school career (Jürges and Schneider, 2007; Betts, 2011; Borghans et al., 2020). Considering the large downstream effects of track choice, what track to pursue rightly matters greatly to pupils and their parents. If the tracking decision is made early in a students' education, getting this decision "right" is all the more important. Policy that affects the tracking decision should therefore be of particular interest to policy makers.

In international comparison, Germany is one of the education systems with the earliest tracking decision (Woessmann, 2009). Pupils are generally tracked into one of three secondary school types after Grade 4 (the basic *Hauptschule*, intermediate *Realschule*, or advanced *Gymnasium*), each of which leads to very different educational outcomes and subsequent labor market opportunities. Mobility between tracks is generally possible but is usually constrained to a downward-mobility (Jürges and Schneider, 2007). Given this early tracking decision, even limited variation in policy or pupils' month of birth can have a large effect on pupils' subsequent school career. The literature largely supports this hypothesis. Exploiting the variation in enrollment age around the enrollment-cutoff, Jürges and Schneider (2007) find that younger pupils are less often recommended to and actually attend the highest secondary school track (although the authors find that flexible enrollment and grade retention partly offset these inequalities). Analyzing a school reform in the state of Bavaria, Piopiunik (2014b) finds that earlier tracking decisions for students in low and middle track schools reduce the academic performance for both treated groups 5 years later. In a sibling study on the effect of preschool attendance on secondary school track choice, Schlotter (2011) finds no effect with respect to the propensity to enroll in the highest school track. In the German context, the literature is mostly concerned with changing the external parameters of elementary school education: whether pupils attend preschool and the timing of the tracking decision. The literature on the internal dimension of elementary school education (i.e. what happens *in the classrooms*) is more scarce. This is likely due to a lack of exploitable policy variation.

This thesis explores one such policy variation: whether the timing of first exposure to grades in elementary school affects pupils' track choice. Throughout the 1970s and 1980s, a number of West German states postponed the assignment of number grades to begin with Grade 3, as compared to Grades 1 or 2 (referred to throughout as *later grading*). This was meant to allot the students more time to acclimate to the school environment, and to shield them from the effects of overly competitive behavior. The effects of this policy are greatly understudied.

This thesis assesses the extent to which the later grading reforms affected pupils' secondary school track choice. In a first step I analyze the average effect of later grading on pupils' track choice. In a second step I explore the extent to which the effects of the policy differ across gender and parental educational backgrounds. This is motivated by a large literature in economics of education that has been concerned with the different effects of grades across genders and household educational backgrounds. For instance, Bonesrønning (2008) and Falch and Naper (2013) separately find in a Norwegian context that girls are exposed to easier grading than boys. Rangvid (2015) similarly finds that boys, pupils from low educated backgrounds, and migrants are systematically assessed lower by teacher scores than girls and pupils from educated backgrounds. Considering the role of parents' educational background, Dustmann (2004) report a strong relationship between parental background and secondary school track choice in a German context, as do Ermisch and Francesconi (2001) for a British context.

In order to estimate the effect of later grading on track choice in Germany, I apply the difference-in-differences (DD) framework, exploiting regional and temporal variation in policy implementation across German states. I use the degree obtained as a proxy for secondary school track choice and estimate the percentage change in pupils' propensity to pursue either of the three secondary school tracks. I use data from the German Socio-Economic Panel (SOEP), a longitudinal survey that began in 1984, which currently covers around 20,000 households and 34,000 individuals. Because of the structure of the SOEP, which regularly adds sample waves to refresh and extend the sample, I can effectively use more than 12,000 observations across six large West German states that account for about 86% of the West German population over the sample period. Four of these states introduced the later grading policy between 1977 and 1988. Two further states, which did not implement the policy, act as control groups throughout. A full-fledged theoretical analysis of the predicted effects of this policy is beyond the scope of this thesis. A naive prediction might be that later grading benefits pupils from low-educated households, as they are likely less acclimated to a school-like learning environment and might thus benefit from a longer transitory period.

I estimate the policy effects separately for the three secondary school tracks *Hauptschule*, *Realschule*, and *Gymnasium*, as well as for a grouped dependent variable that considers jointly whether pupils attended a *Realschule* or a *Gymnasium*. As pupils almost exclusively choose one of these tracks, changes in one dependent variable have to be equalized by opposite changes in another. Considering the average effect irrespective of gender or parental background, I fail to reject the null hypothesis of no effect for all four dependent variables. Allowing for the treatment effect to differ between genders, I find suggestive evidence that later graded males are between 4.1 and 6.7 percentage points less likely to obtain a *Hauptschule* degree compared to earlier graded boys and around 6 percentage points more likely to obtain a higher degree than the *Hauptschule* degree. I find no evidence of a treatment effect on females. Allowing the treatment effect to differ between pupils from educated households (who have at least one parent who holds the highest secondary school degree *Abitur*) and low-educated households, I find suggestive evidence that later grading decreases the propensity of pupils from educated households to obtain a higher degree than the *Hauptschule* degree by around 3.3 percentage points. Pupils from low-educated households are around 3.6 percentage points less likely to

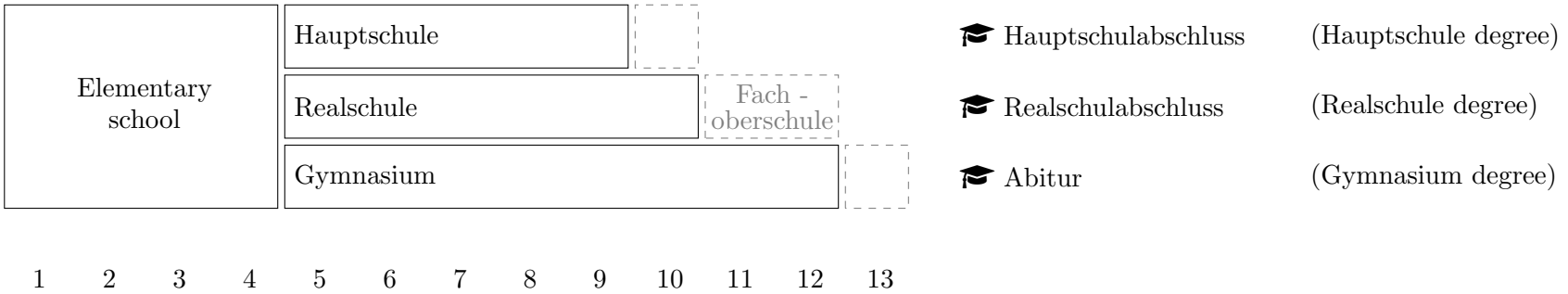
obtain a *Hauptschule* degree, compared to their earlier graded peers, but are not statistically more likely to obtain a higher degree. This almost puzzling result is likely due to a small share of pupils (2% of the sample) who drop out of school without a degree. The results for males are consistent across an alternative weighted estimation and a number of alternative variable definitions and sample constructions. The results for pupils from educated and low-educated households vary more across alternative specifications and weighing choice. Given this variation and since I can almost uniformly only reject the null hypotheses of zero effects at the 5% level, the results presented in this thesis should be considered suggestive.

The only comparable institutional set-up that I am aware of is a grading reform in Sweden between 1969 and 1982. The new national curriculum introduced in 1969 allowed each municipality to decide independently whether to grade its pupils before Grade 7 or not. Studies analysing the effects of this reform have yielded mixed evidence. Using a difference-in-differences strategy, Sjögren (2010) finds some evidence that being graded earlier increased girls' years of schooling, but finds no average effect for boys. Earlier grading increased the probability of graduation for both boys and girls from low-educated households while earlier graded sons of university graduates were found to earn less and to be less likely to obtain a university degree. Somewhat contrary to these findings, using a different Swedish dataset, and controlling for cognitive ability, Klapp et al. (2014) find no average effect of grading on achievement one year later but do find that earlier graded boys and earlier graded low-ability students in general obtain lower grades one year later compared to later graded students. Furthermore, the authors find no substantial effect heterogeneity across socio-economic backgrounds. In light of this inconsistent evidence, the effects of later grading on achievement and school track remain an open question.

This thesis extends the existing literature in three ways. First, it expands the literature on German educational policy by considering a previously unstudied policy change. The existing literature has mostly focused on policies with larger expected effects such as the extension of compulsory schooling in the 1960s and the shortening of the *Abitur* track education in the *Gymnasium* from 13 to 12 years in the early 2000s (see Section 2). Second, this thesis expands on this literature by also accounting for the enrollment decision around the two lower tracks: the *Hauptschule* and the *Realschule*. Most of the previous research has focused on the enrollment decisions into the highest school track. Finally, it enhances the limited literature on the effect of exposure to grades in elementary school by incorporating evidence from a highly stratified education system.

The rest of the thesis is structured as follows. Section 2 introduces the German school system and discusses the institutional background and the roll-out of the later grading reforms in the 1970s and 1980s. Section 3 discusses the data from the Socio-Economic Panel. Section 4 introduces the empirical strategy. Section 5 presents the results. Section 6 discusses robustness checks and further difference-in-differences analytics. Section 7 discusses and proposes potential mechanisms. Section 8 concludes.

**Figure 1** The German school system




*Notes*: The figure illustrates the German school system. Elementary school covers Grade 1 to Grade 4. In elementary school, all students learn together. At the end of Grade 4, students are tracked into one of three secondary school tracks: the basic *Hauptschule*, the intermediate *Realschule*, or the advanced *Gymnasium*. The basic *Hauptschule* covers Grades 5–9 (in some states 5–10). After the *Hauptschule*, graduates obtain a *Hauptschulabschluss* (or *Hauptschule* degree) and continue at vocational schools. The intermediate *Realschule* covers grades 5–10. *Realschule* graduates obtain a *Realschulabschluss* (or *Realschule* degree) and can, if their grades allow it, continue at a *Fachoberschule*, which usually covers Grades 11–12. At the end of the *Fachoberschule*, students can obtain the *Fachhochschulreife*, which allows entry to a university of applied sciences (*Fachhochschule*). This is indicated by a shaded rectangle. The advanced *Gymnasium* covers Grades 5–12 or 5–13. *Gymnasium* graduates obtain the *Abitur* or *Allgemeine Hochschulreife* (general university entrance qualification), which allows entry to all degree programs at university. The figure provides a general description. The shaded rectangles in the *Hauptschule* and *Gymnasium* track indicate variation in track-length across states. Across German states, some variation with respect to track names and study length persist.

## 2 Institutional background

In Germany, education is the domain of the *Länder* (states). German states may design their educational system according to their preferences, within the parameters set by the *Düsseldorf Accord* (1955) and the *Hamburg Accord* (1964) (Helbig and Nikolai, 2015b). The *Düsseldorf Accord* aligned the broad parameters of the school year, the grading scale, and the structure of the secondary schools (Konferenz der Ministerpräsidenten, 1955). The *Hamburg Accord* aligned the beginning of the school year across the states and extended compulsory education to nine years (Konferenz der Ministerpräsidenten, 1964).[1]

Despite its considerable regional variation, the German school system is aligned in its three-stage system, dividing schooling into primary, lower secondary, and upper secondary education. Pupils usually enter school in the year they turn 6 years old and learn in shared classes throughout their primary education at elementary school, which consists of the first four years of school.[2] After elementary school, the German education system is highly stratified. After Grade 4, pupils enter one of three tracks according to their ability, their grades, and their parents' decision. Each track leads to a different secondary school degree, and subsequent educational and vocational prospects. Ordered from least academic to most academic, the three tracks are (i) *Hauptschule*, (ii) *Realschule*, and (iii) *Gymnasium*. Figure 1 provides a graphical illustration of the German school system. The different school types in lower secondary education can be organised either as wholly separate schools, independently teaching towards either of the degrees; or as integrated schools with multiple tracks. In the latter case, the education is administered either in degree-specific classes, or, if there is no clear distinction by class, students from different tracks are taught separately in some core subjects. Out of these types, wholly separate schools were the most common school type. Integrated schools are a more recent phenomenon. In 1980, 39% of pupils in lower secondary school studied at a *Hauptschule*, 27% studied at at *Realschule*, 30% studied at a *Gymnasium*, and only 4% studied at a so-called *integrated general school* (Federal Statistical Office of Germany (Destatis), 2019, own calculations). Compulsory schooling in Germany consists of at least 9 years of school education (primary and lower secondary) plus, in the case of the lower secondary school degrees, some form of mandatory part-time vocational education (Konferenz der Ministerpräsidenten, 1964; Kultusministerkonferenz, 2019b).

The *Hauptschule* is the most basic type of school at lower secondary level and provides a basic general education, usually comprising Grades 5–9. It culminates with the first lower-secondary degree, the *Hauptschule* degree, which entitles the holder to pursue a dual vocational education. The *Realschule* is an intermediate type of school at lower secondary level, usually comprising Grades 5–10. It culminates in the second lower-secondary degree, the *Realschule* degree. This provides pupils with a more extensive general education and the opportunity to go on to courses of education at upper secondary level that lead to vocational or higher education entrance qualifications. The *Gymnasium* is a type of school covering both lower and upper secondary level (Grades 5–13 or 5–12) and provides an in-depth general education. It

[1] The school year was uniformly moved to begin after the summer holidays instead of beginning in the spring, as was the case in some states prior to the Accord. The implementation of this policy brought with itself a period of shortened school years to facilitate the transition from the start of the school year in the spring to the fall.

[2] In Berlin and Brandenburg, primary school comprises six Grades (Kultusministerkonferenz, 2019a).

culminates in the *Allgemeine Hochschulreife* which is obtained by passing the *Abitur* in either 12th or 13th Grade. The *Allgemeine Hochschulreife* is a general university entrance qualification, which entitles a holder to admission to all subjects at all higher education institutions. It should be noted that a fourth type of school and degree is nested in between the *Realschule* and the *Gymnasium*. This is the *Fachoberschule*, which culminates in the *Fachhochschulreife* after usually 2 additional years of schooling. Entry to the *Fachoberschule* requires a *Realschule* school degree. The *Fachhochschulreife* entitles a holder to study at a *Fachhochschule* or university of applied sciences. In the context of this thesis, pupils who have obtained a *Fachoberschule* degree are treated as having obtained a *Realschule* degree, as this is the most common path into the *Fachoberschule*.

On a national scale, a number of policies have shaped the German education system since the Second World War. These policies have attracted the majority of the attention in the literature. First to note are the abolition of school fees for Gymnasiums in the 1950s and 1960s (Reinhold and Jürges, 2010; Riphahn, 2012), followed by the extension of compulsory schooling to nine years (Pischke and Von Wachter, 2008; Piopiunik, 2014a; Cygan-Rehm, 2018), and the short school-years following the alignment of the beginning of the school year (Pischke, 2007). Other important more recent policy changes were the move to an 8-year Gymnasium (from the previously standard 9 years) (Huebener and Marcus, 2017; Marcus and Zambre, 2019; Meyer et al., 2019) and the introduction of centralized exit examinations for the *Abitur* (Jürges et al., 2012; Piopiunik et al., 2013).

While these policy changes have been adopted by most states, their implementation has differed across states as cross-state policy differences permeate the German educational system. Of particular interest in the context of this thesis are policies that affect the transition from elementary school to secondary education. Helbig and Nikolai (2015b) outline five dimensions along which policies regulating the transition from primary to secondary education have differed across states: (i) whether the school's recommendation is pegged to a specific grade average; (ii) whether the school's recommendation is binding; (iii) whether, in the case of a binding recommendation, it is possible to conduct entry exams; (iv) whether an entry exam is prescribed for all students wishing to enter *Gymnasium*; and (v) whether there is a trial period after transitioning from elementary school to the *Gymnasium*. The majority of the policy reforms that have changed how states managed pupils' transition into secondary school changed before or after the time period considered in this thesis. One exception is the relaxation of the degree to which the schools' recommendation for a child's secondary school track was binding. Bremen relaxed the recommendation in 1977, Lower-Saxony in 1978, and Rhineland-Palatinate in 1984. The policy change in Lower-Saxony happened only one year after its late grading reform. Thus, this policy change potentially confounds my treatment estimates. I show in section 6 that the results presented in this thesis are robust to excluding Lower Saxony from the sample, which indicates that this almost simultaneous policy reform does not materially confound my estimates.

## 2.1 The late grading reforms

The question, whether and when children should be graded in elementary school is a a topic of recurring argument in German education policy. The *German Elementary School Association* (Grundschulverband), for example, advocates for an abolishment of grades throughout elementary school, pointing to what it says is "questionable evidence of its effect on performance" (Grundschulverband, 2019).[3] On the other hand, the *German Association of Philologists* (Deutscher Philologenverband, an advocacy group of Gymnasium teachers) endorses number grades as a tool to introduce pupils to the comparative power of formal assessments, which it contends is important for success in later life (Deutscher Philologenverband, 2016).

In 1970, the *German Education Committee* (Deutscher Bildungsrat) recommended that adolescents and children should be shielded from the principles of competition that underpin most social and economic interaction. Rather, children should be introduced to the competitive aspect of society in a manner appropriate for their age and free of the threat of life-long disadvantage or social downgrading that is often associated with number grades (Deutscher Bildungsrat, 1970, cited in Urabe (2009)). In response to this, the *Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany* (KMK) recommended that "given the objective of the first stage of education in elementary school, in first and second Grade a general assessment of the child's performance is more important than the precise grading of the achievement in each individual subject" (Kultusministerkonferenz, 1970, p. 35, own translation).[4] Effectively, the Conference recommended the abolishment of the formal number grades in Grades 1 and 2.

In the following years, many German states followed this recommendation.[5] In line with the KMK's recommendation, written assessments of pupils' strengths, weaknesses, and interests took the place of the formal number grades. Other states have maintained their policy of formally grading students even before Grade 3, leading to an ongoing diversity of grading practices in elementary school. The debate has not subsided. Journalists regularly comment on the alleged benefits of postponed or abolished grading and chronicle parents' stories of the psychological burden that the grading culture is placing on their young children.[6]

In the sample used in this thesis, the first state to postpone grading in elementary school was Lower Saxony for the school-year 1977/1978 (Lower Saxony, 1977; Helbig and Nikolai, 2015a). North Rhine-Westphalia followed in 1979/80, Hesse in 1981/82, and Rhineland-Palatinate in 1988/89 (North Rhine-Westphalia, 1979; Hesse, 1980; Rhineland-Palatinate, 1988). The small German city states of Bremen and Hamburg also enacted the postponed grading reforms in

[3] The "questionable evidence" cited by the German Elementary School Association amounts to high-level analyses of cross-country differences in performance (cf. Grundschulverband, 2018), but does not include rigorous analyses of the performance effect of abolished or introduced tests and grading.

[4] In the German original: "In der 1. und 2. Klasse ist eine allgemeine Aussage über die Leistungen eines Kindes im Hinblick auf das Ziel dieser Schulstufe bedeutsamer als die vorgeblich genaue Benotung der Leistungen in den einzelnen Teilgebieten des Unterrichts. In diesen beiden Klassen ist daher jeweils am Ende eines Schuljahres eine allgemeine Beurteilung des Kindes in freier Form im Zeugnis zu erteilen." (Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland; Kultusministerkonferenz, KMK).

[5] See Table A.1 in the Appendix for a full list of states and implementation years.

[6] See for instance Susanne Klein, "Grades are unnecessary in elementary school." (Noten sind in der Grundschule unnötig.) Sueddeutsche Zeitung, December 1, 2017; Rainer Werner, "Without grades looms the teachers' secret code." (Ohne Noten droht der Geheimcode der Lehrer.) Die Welt, February 9, 2019.

1971/72 and 1979/80, respectively, but are not part of the sample. The Saarland ultimately enacted the reform in 1994/95, before moving back to grading before Grade 3 in 1999/2000 (Helbig and Nikolai, 2015a).[7]

# 3 Data

The results in this thesis are based on the 34th wave of the German Socio-Economic Panel (SOEP) (covering years 1984–2017). The SOEP is a large annual household survey that has been conducted in West Germany since 1984 and is representative of the resident population. It includes detailed questionnaires on demographic and household characteristics, retrospective biographical information and (parental) educational outcomes. Following the first wave starting in 1984, subsequent waves have added observations in order to keep the sample representative and to expand its scope.[8] It is comparable in structure and scope to the Panel Study of Income Dynamics (PSID) in the United States.

The ideal data-set for this study would contain a cross-section of individuals living in West Germany who started elementary school between a few years before the first policy change (Lower Saxony in 1977/78) and a few years after the last policy change (Rhineland-Palatinate in 1988/89). It would include their state of school attendance, their subsequent track choice, their parents educational attainment, and other covariates such as whether their parents had a migrant background and their household income during their school-years. Except for historical income data, the SOEP either contains information on all of these variables or their value can be imputed based on reasonable assumptions. The following paragraphs discuss the relevant imputations and sample restrictions. Given the SOEP's longitudinal study design, each individual in the SOEP appears in multiple years. In order to maximize the informational content and accuracy of the data, I restrict the sample to only include the last available information on each individual in the sample.[9]

**States in the sample** I restrict the sample to six West German states that together account for 86% of the population of West Germany in 1970 (Federal Statistical Office of Germany (Destatis), 1970, own calculations). I do this to exclude observation from East Germany and to ensure a sufficient number of observations for each state and year. Since the policy was only enacted by West German states and the SOEP only contains information for West German residents for any time before the reunification, I drop all observations that are first registered in any of the states of the former East Germany (including Berlin). I furthermore drop all

[7] Hesse also aborted its late-grading policy in 1998/99.

[8] See SOEP (2019). The SOEP deploys random probability sampling. General population samples are drawn in a nation-wide two-stage stratified sampling procedure, first sampling nation-wide sampling points by federal state and municipality size, secondly, within each sampling point, sampling households using a random-walk procedure. Within each household, all residents are included in the sample. If members of an originally sampled household leave that household, both the original and the split-household are interviewed. See Goebel et al. (2019) for an outline of the various waves and sample sizes. All questionnaires in German and English are available at `http://panel.gsoep.de/soepinfo2017/`.

[9] This is possible since I am not interested in any variable for a particular year (such as *income in year t*, for instance). The information on each individuals' highest degree obtained and on their parents education should be most accurate in the latest observation available.

observations that indicate that the respondents resided in Eastern Germany including Berlin at the time of the fall of the Berlin wall in 1989. Given the SOEP's relatively small sample size on a per-cohort level, I am concerned with precisely estimating the policy effect. This requires sufficiently large sample sizes for each state-cohort cell, such as to limit the effect of outliers on the sample distribution. Observations per state-cohort cell are particularly low in the smallest states. I thus drop observations for the two West German city states Bremen and Hamburg and for the small West German states Saarland and Schleswig-Holstein. I may be concerned that, having dropped 4 of 10 West German states (plus Berlin), my sample may be overly restricted in terms of population. However, since the states I drop are either city states or very small, the remaining states still account for roughly 86% of the population of West Germany in 1970. This leaves six states in the sample: Baden-Württemberg, Bavaria, Hesse, Lower Saxony, North Rhine-Westphalia, and Rhineland-Palatinate. Baden Württemberg and Bavaria serve as the control states throughout the analysis as they did not change their grading policy.

**State of school attendance** I use the first observed state of residence for each individual as a proxy for the state in which an individual enrolled in school. Ideally, the SOEP would include the state of first entry into school or the state of secondary school graduation. This information is only available for a small subset of individuals in the sample. The SOEP does, however, track each individuals' geographic location (on a state level) and this information is available for every year the individual participates in the survey.[10]

**Cohorts in the sample** I restrict the sample to include individuals born between 1964 and 1987. The first state to postpone grading in elementary school was Lower Saxony in 1977/78. The last state to postpone grading was Rhineland-Palatinate in 1988/89. Individuals born in 1964 entered school at the latest in 1971, 5 years before the first affected cohort entered school in 1976. Since the treatment states did not distribute grades in Grade 1, the first treated cohort for each state is the cohort entering Grade 2 in the year the policy took effect. Individuals born in 1987 entered school at the earliest in 1993, 6 years after the first affected cohort in Rhineland-Palatinate. Thus, as children in Germany usually enter school in the year after they turn six years old, this allows for a sufficient number of pre-treatment cohorts for the first treated state, Lower Saxony, and for a sufficient number of post-treatment cohorts for the last treated state, Rhineland-Palatinate.[11]

**Year of school entry** I use month of birth information to impute the year of school entry, if available. If the information is not available, I assume the child entered school 7 years after its birth. I have to impute this information, since the SOEP does not in general contain information

[10] Tables A.2 and A.3 in the Appendix show that the last known state of residence corresponds to the state of school enrollment and school graduation for 92–93% of individuals for whom this information is available.

[11] Determining the school starting cohort for each individual is a separate issue that I discuss below. The marginal cohorts to include/exclude are somewhat arbitrary. Results are robust to excluding the entry cohorts after 1986, see section 6. I exclude entry cohorts at the margin that do not consist of full entry cohorts. Children born early in 1964 entered school in 1970 and make up one half of the 1970 entry cohort. The other half consists of children born late in 1963. In order to only include full entry cohorts, I exclude children born early in 1964 and thus exclude the entry cohort of 1970.

on the year in which an individual enters school. Until 1997, a child was obliged to attend school starting with the year they turned 6. The effective cut-off was uniformly on June 30th, meaning that a child born before June 30th would begin school in the calendar year they turned 6, while a child born after June 30th would begin school in the following year (Konferenz der Ministerpräsidenten, 1964).[12] Of the 12,898 individuals in the final sample, information on the month of birth is available for 11,390 individuals.

**School track** I use each individual's ultimately obtained secondary school degree as a proxy for the attended school track. I do this because the SOEP does not contain full information on an individuals' first secondary school track. I drop all observations that have incomplete or non-usable information on the school degree obtained.[13] Individuals who are coded as having obtained a *Fachoberschule* degree (or *Fachhochschulreife*) are coded as *Realschule* graduates, because attending a *Fachoberschule* requires at least a secondary school degree, usually a *Realschule* degree.[14]

**Parental education** Throughout the empirical analysis the primary parental educational control variable is whether either parent attended the *Gymnasium* and obtained the highest secondary school degree *Abitur*. This is true for roughly 16% of individuals in the sample. I drop all observations with missing information on parental educational background.[15] This reduces my effective sample size to 12,898 individuals.

Summary statistics for this sample are presented in Table 1. Panel A reports the distribution of individuals in the sample across the school types. Panel B reports the gender distribution and information on parental background for the sample and separately for each school type. A plurality of individuals (42%) obtained a *Realschule* degree, compared with 33% who obtained a *Hauptschule* degree, 29% who obtained a *Gymnasium* degree, and 2% who failed to obtain any degree. Of the individuals in the sample, 53% are female. However, only 45% of pupils who obtain a *Hauptschule* degree are female compared with 59% for the *Realschule* and 54% for the *Gymnasium*. With respect to their parental households, 13% have parents with a migrant background and a majority of pupils (69%) has at least one parent with a *Hauptschule* degree. 28% of individuals have at least one parent with a *Realschule* degree, and 16% of individuals have at least one parent who attended a *Gymnasium*. 7% of individuals have at least one parent

[12]The timing of school enrollment is subject to some parental and institutional discretion. Every year, some parents may choose to postpone or to prepone the enrollment of their children but recent data suggest that this is not common. While most students tend to be enrolled on time (89%), of those who are not enrolled, the majority postpones enrollment (8%). Only a minor fraction enrols prematurely (3%) (Federal Statistical Office of Germany (Destatis), 2018). In a robustness check in section 6, I impute the year of school entry as birth year plus *6*. The results with this alternative specification are similar.

[13]This contains observations marked as *no information (keine Angabe)*, *other (anderer Schulabschluss)*, and *not yet graduated (noch kein Schulabschluss)*.

[14]The decision to consider *Realschule* degree and *Fachhochschulreife* jointly is the same as the approach of Piopiunik et al. (2013).

[15]For some observations, the information is completely missing, while for some observations the reason for missing is provided. See Appendix B for a discussion of the observations with missing information and potential issues. The primary background control throughout the empirical analysis is whether either parent holds a *Gymnasium* degree. An alternative would be to code 'missing' information as 'non-*Gymnasium*'. Non-randomly missing information on parental education is likely a good proxy for neither parent having a *Gymnasium* degree. The results using this alternative specification are qualitatively similar. See section 6.

**Table 1** Summary statistics: Sample means and background characteristics

| | Sample | Hauptschule | Realschule | Gymnasium |
|---|---|---|---|---|
| **Panel A: Degree distribution** | | | | |
| Hauptschule | 0.27 | | | |
| Realschule | 0.42 | | | |
| Gymnasium | 0.29 | | | |
| No degree obtained | 0.02 | | | |
| **Panel B: Pupils' and household characteristics by school type attended** | | | | |
| Female | 0.53 | 0.45 | 0.59 | 0.54 |
| Male | 0.47 | 0.55 | 0.41 | 0.46 |
| At least one parent with Hauptschule | 0.69 | 0.79 | 0.75 | 0.51 |
| Both parents with Hauptschule | 0.45 | 0.60 | 0.49 | 0.25 |
| At least one parent with Realschule | 0.28 | 0.11 | 0.30 | 0.42 |
| Both parents with Realschule | 0.07 | 0.02 | 0.07 | 0.10 |
| At least one parent with Gymnasium | 0.16 | 0.03 | 0.11 | 0.38 |
| Both parents with Gymnasium | 0.05 | 0.00 | 0.02 | 0.14 |
| At least one parent without a secondary school degree | 0.07 | 0.12 | 0.06 | 0.04 |
| Both parents without a secondary school degree | 0.02 | 0.04 | 0.02 | 0.01 |
| Parents with a migrant background | 0.13 | 0.17 | 0.12 | 0.11 |
| Father's occupational prestige score (MPS) | 60.00 | 45.39 | 55.39 | 79.73 |
| Mother's occupational prestige score (MPS) | 57.08 | 46.73 | 54.38 | 70.20 |

*Notes:* The table reports summary statistics for the main sample and for the dependent variables. Panel A reports information on the degrees obtained by individuals born between 1964 and 1983. Panel B reports information on their gender, household characteristics, and parental background. "Both parents with <school type>" reports the share of individuals for whom both parents attended precisely <school type>. "At least one parent with <school type>" reports the share of individuals for whom at least one parent attended the respective <school type> without considering the other parent's educational background. "Parents with a migrant background" reports the share of parents who have immigrated to Germany. The magnitude prestige score (MPS) is a measure of the prestige ascribed to different professions, ranging from 30 (helping farmworkers) to 216 (dentists).

without a secondary school degree. Noticeably, the share of pupils with at least one parent with a *Hauptschule* degree decreases with pupils' secondary school degrees. While 79% of pupils in the *Hauptschule* have at least one parent with a *Hauptschule* degree, the same is true for only 51% of pupils in the *Gymnasium*. The opposite effect can be observed when looking at parents who attended a *Gymnasium*. Only 3% of pupils in the *Hauptschule* have at least one parent who attended a *Gymnasium* and no one in the *Hauptschule* has two parents who attended a *Gymnasium*, while 38% of pupils in the *Gymnasium* have a parent who attended the same school type and 14% have two parents who graduated from the *Gymnasium*.

**Table 2** Summary statistics: Number of observations per state

| State | Year of policy change | Treatment share | Observations | Observations (weighted) | Complete observations | Complete observations (weighted) |
|---|---|---|---|---|---|---|
| Baden-Wuerttemberg (C) | – | – | 2840 | 2746.7 | 2401 | 2453.5 |
| Bayern (C) | – | – | 3511 | 3142.8 | 2977 | 2803.1 |
| Lower Saxony | 1977/78 | 0.78 | 1998 | 1906.3 | 1679 | 1659.2 |
| North Rhine-Westphalia | 1979/80 | 0.70 | 4524 | 4361.8 | 3759 | 3847.9 |
| Hesse | 1981/82 | 0.61 | 1528 | 1484.5 | 1259 | 1321.8 |
| Rhineland-Palatinate | 1988/89 | 0.30 | 973 | 916.9 | 823 | 812.5 |
| Total | | | 15374 | 14559 | 12898 | 12898 |

*Notes*: *Year of policy change* indicates the school year that the policy first took effect. The *treatment share* is the share of a state's cohorts in the sample that are "treated". *Observations* are the number of valid observations per state, notwithstanding missing information on additional controls. *Observations (weighted)* are the weighted number of observations net of zero-weight observations. *Complete observations (weighted)* are weighted observations with complete information on covariates (i.e. parental educational background). (C) indicates control states.

Table 2 depicts the distribution of the observations across the treatment and control states. Column 1 (*Year of policy change*) in Table 2 presents the school year in which each respective policy change took effect. Column 2 (*Treatment share*) indicates the share of a states cohort that are treated throughout the sample. To illustrate, 78% of the cohorts from Lower Saxony are in the later graded group, compared to only 30% of cohorts from Rhineland-Palatinate. Column 3 (*Observations*) indicates the number of unweighted observations per state.[16] Column 4 (*Observations (weighted)*) indicates the (weighted) number of observations (with non-zero weights). The difference between columns 3 and 4 is due to observations that have zero weights, see Appendix B for a discussion. Column 5 (*Complete observations*) presents the number of observations for which all information is available and Column 6 (*Complete observations (weighted)*) presents the weighted distribution of complete observations across the states. The observations according to Column 5 are the basis for the baseline analysis in this thesis. The observations according to Column 6 are the basis for the weighted analysis in section 6.

## 3.1 Weights

The SOEP aims to allow inference about the population based on a relatively small sample. Drawing of the target households (and, consequently, individuals) is designed to be representative. However, some drawn households/individuals do not actually participate in the survey.

[16]The SOEP includes weights that relate the sampled observations to the resident population. Section 3.1 further discusses the issue of weighting the estimation. The baseline regression is estimated using unweighted observations but only includes observations for which the weights are non-zero. See Appendix B for a discussion of the zero-weight observations and for the estimates for the analogous weighted estimations.

Furthermore, while some of the sampling waves are representative of the German population, other sub-samples over-sample households according to certain characteristics (e.g. family structure or income group). To account for this non-response and oversampling, the SOEP includes a weighting factor that is estimated based on the drawn gross sample and aligned with the known distribution of these characteristics within the German population. The "known" parameters are based on the Micro Census, an annual sample of the persons and households in Germany. On the household level, these take into account the state of residence, the size of the municipality, home-ownership, and household size. On the individual level, these take into account age, sex, and nationality (Pischner, 2007; Goebel et al., 2019).

The question of how to account for sample weights in studies aimed at causal inference is not trivial (cf. Angrist and Pischke, 2008). Solon et al. (2015) distinguish three potential motives for weighting when estimating causal effects. One reason is to achieve more precise estimates by correcting for heteroskedasticity. In my analysis, heteroskedasticity of standard errors is accounted for separately by clustering standard errors on the state level and by Wild t bootstrapping the p-values.[17] A second common reason for weighting is to identify average partial effects in the presence of heterogenous effects. Group weights are then used to average out heterogeneity of the treatment effects across groups. I explore potential heterogeneity of the treatment effects separately and in more detail in section 5.2. The most pertinent case for weighting the estimation is to achieve consistent estimates by correcting for endogenous sampling. This issue arises if the probability of selection varies with the dependent variable even after conditioning on the explanatory variables (Solon et al., 2015). Since the SOEP oversamples along certain population-characteristics in sub-samples, this may be an issue. I use the unweighted estimates for the main analysis. The sensitivity of the results to the weighting-choice is discussed in section 6. The results are broadly consistent across the unweighted and weighted estimates, although the weighted estimates tend to indicate a slightly higher level of significance.

# 4 Empirical strategy

In order to estimate the causal impact of postponed grading on track choice, ideally pupils would be randomly assigned to either the treatment group (later grading) or the control group (earlier grading). Then, the causal effect of later grading could be estimated by the coefficient on the treatment dummy.[18] I follow Kahn-Lang and Lang (2019) in their discussion of the

[17] The reasoning to apply weighting to correct for heteroskedasticity follows the observation, that, if the error term across groups is independent and identically distributed and if the groups vary significantly in size, then the group-average error term will be highly heteroskedastic. Weighting by group size might then be applied to correct the standard error. However, Solon et al. (2015) note that the assumption that the individual-level error terms are independent is often wrong and that instead individual-level error terms tend to be correlated with each other because of unobserved, group-specific factors. That is, the error terms tend to be clustered. In order to account for heteroskedasticity and cluster error terms, I cluster the error terms by state and calculate bootstrapped p-values (see section 4.3).

[18] This is a simplification. For an extensive discussion of what randomization can and cannot achieve, see Deaton and Cartwright (2018).

difference-in-differences methodology with respect to the potential outcomes framework.[19] In the potential outcomes framework, let $E(Y_{ic}(D_1))$ be the potential outcome of individual $i$ belonging to cohort $c$ if they are treated $(D_1)$ and $E(Y_{ic}(D_0))$ if they are not treated $(D_0)$. A standard experiment with randomized assignment to treatment and control group would then ensure that the average outcomes for treated and untreated individuals would be equal in the absence of treatment (subject to sample variation and assuming no attrition). To illustrate, let $T_i = 1$ denote individuals who belong to the treatment group. Then

$$E(Y_{ic}(D_1)|T_i = 1) = E(Y_{ic}(D_1)|T_i = 0) \tag{4.1}$$

and

$$E(Y_{ic}(D_0)|T_i = 1) = E(Y_{ic}(D_0)|T_i = 0) \tag{4.2}$$

where the right hand term in equation 4.1 and the left hand term in equation 4.2 are counterfactuals that cannot be observed. This set-up fails when there are systematic differences between treatment and control groups that are also correlated with the outcome variable. This may be the case in the type of quasi-natural experiment that I propose to exploit. Given the regional character of education policy in Germany, different states may differ in the distribution of degrees, resulting in different levels for the outcome variables in the pre-treatment periods. The key to identifying the causal effect in the difference-in-differences framework is then that, in the absence of treatment, outcomes between the treatment and the control group would have moved in *parallel* in the treatment period. Consider a two-period model ($c \in \{0, 1\}$), where treatment is introduced between period 0 (*pre-treatment period*) and period 1 (*post-treatment period*):

$$\begin{aligned} &E(Y_{i1}(D_0)|T_i = 1) - E(Y_{i0}(D_0)|T_i = 1) = \\ &E(Y_{i1}(D_0)|T_i = 0) - E(Y_{i0}(D_0)|T_i = 0) \end{aligned}$$

which is to say that the difference between the expected value of the treatment group's hypothetical post-treatment outcome under no treatment and its actual expected pre-treatment outcome is equal to the difference between the control group's expected post-treatment outcome under no treatment and its expected pre-treatment outcome under no treatment.

This is the identifying assumption within the difference-in-differences framework: in the absence of treatment, the outcome variable for treatment and control group would have moved in parallel in the treatment periods. It is not possible to evaluate this counterfactual, as we only observe the realized outcome for the treatment group under treatment and the control group under no treatment and not their respective counterfactual potential outcomes. The common practice in the difference-in-difference literature is to evaluate whether the common trends assumption holds in the pre-treatment period and to use leading and lagged treatment variables to partially verify the sensibility of this assumption in an event study (Angrist and Pischke, 2008).

[19] Rubin D. B (1974) introduced what is now often called the "Rubin Causal Model" to formalize the identifying assumption underlying observational (as opposed to randomized) studies.

### 4.1 Event study analysis of pre-treatment trends

In the context of an event study, the leading treatment variables are a set of dummies $D_{s,c+\tau}$ (where $s$ indicates *state*, $c$ indicates the first treated *cohort*, and $\tau$ indicates a cohort's time distance to $c$) that are equal to 1 if for a specific cohort the onset of the treatment is $\tau$ periods in the future. Similarly, the lagging treatment variables are a set of dummies $D_{s,c-\tau}$ that are equal to 1 if for a specific cohort the onset of the treatment was $\tau$ periods in the past. For each individual belonging to a specific cohort, only one of these variables is equal to 1, all other variables are equal to 0. The intuition behind the event study analysis is that if the explanatory variable $D_{sc}$ causally determines the dependent variable, then leads of the policy variable $D_{sc}$ should not matter in the regression (Angrist and Pischke, 2008, p.177), i.e. they should be statistically indistinguishable from zero. This is to say that the treatment should not have a statistically significant effect before it is initiated. In the context of this thesis, the later grading reforms should not statistically matter for pupils' propensity to pursue a specific school track *before* the policy is implemented.

To analyze this hypothesis, I follow Goodman-Bacon (2019) and assign each observation in the treatment sample to a cohort relative to treatment.[20] In the estimation, I group all relative event times greater than +12 in the lagging dummy $\tau = 12$ and all relative event times smaller than -6 in the leading dummy $\tau = -6$. Methodologically, the event study estimates only provide sensible estimations for the balanced event times, i.e. for the time periods relative to treatment that are observed for all treatment groups. Therefore, I only present relative event times within the $[-5, 6]$ interval. To obtain the event study estimates, I then regress the outcome variable (i.e. separate regressions for *Gymnasium*, *Realschule*, and *Hauptschule*) on a set of leading and lagging treatment dummies, cohort and state fixed effects, and a female dummy.

Formally, the regression to obtain the event study plots is represented by the following model:

$$y_{isc} = \gamma_s + \lambda_c + \delta_1 \cdot female_i + \sum_{\tau=0}^{m} \beta_{-\tau} \cdot D_{s,c-\tau} + \sum_{\tau=2}^{q} \beta_{+\tau} \cdot D_{s,c+\tau} + \varepsilon_{isc} \tag{4.3}$$

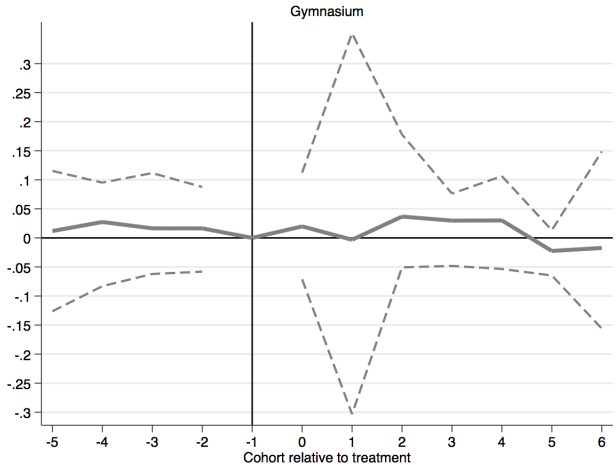
where $\gamma_s$ is a state fixed effect, $\lambda_c$ is a cohort fixed effect, $female_i$ is a female dummy (equal to 1 if an individual is female, 0 otherwise), and the sums on the right-hand side allow for $m = 12$ lags $(\beta_{-1}, \beta_{-2}, \ldots, \beta_{-12})$ or post-treatment effects and $q = 6$ leads $(\beta_{+2}, \ldots, \beta_{+6})$ or anticipatory effects. $D_{s,c-\tau}$ is a dummy variable that equals 1 if for a given cohort the policy was implemented $\tau$ periods ago. Consider the track choice $y$ of a student in Lower Saxony ($s = LowerSaxony$) of the cohort starting school in 1975. The first cohort affected by the policy change $c$ is the cohort starting school in 1976 ($c = 1976$). Relative to the 1975 cohort, the policy was thus implemented 1 year in the future ($1975 + 1 = 1976$). Thus, the active treatment dummy for this cohort in Lower Saxony is the leading treatment dummy $D_{s,c+1}$.

Figure 2 plots the event study estimates for each possible degree for the leads and lags around the policy implementation. The x-axis depicts the cohort relative to treatment. $X = 0$ is the first

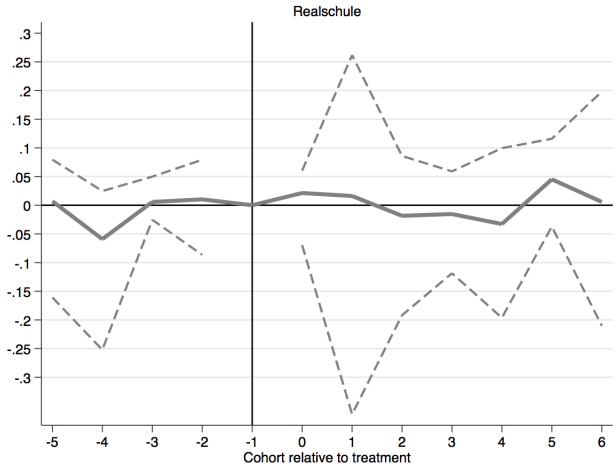
[20] For example, an observation that started school a year after the first treated cohort is coded as `cohort_relative=1`. Observations from the control states are also coded as `cohort_relative=-1`. The relative cohort immediately prior to the policy implementation serves as the effective reference group. Therefore, the dummy `relative_cohort=-1` is omitted from the regression.

**Figure 2** Event study graphs by degree

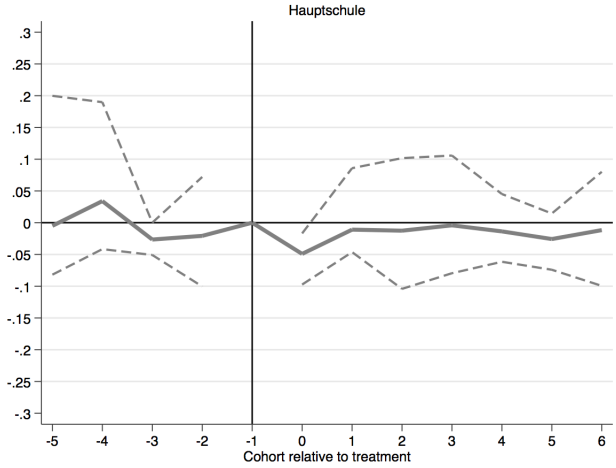
**(a)** Gymnasium (general university qualification)




**(b)** Realschule (intermediate degree)




**(c)** Hauptschule (basic degree)




*Note:* Event studies are plotted with bootstrapped 95% confidence intervals. Point estimates for the event study and plots for the weighted estimation can be found in Appendix C.

affected cohort. To the right are further treated cohorts relative to the policy implementation. To the left are pre-treatment or untreated cohorts. For the event study to support the validity of the identifying assumption, the leads (i.e. the coefficients towards the left of the $x = 0$) should not be statistically significantly different from zero as indicated by their bootstrapped confidence intervals. This is true for all of the leads, i.e. none of the leads are significantly different from zero. One close exception is the three period leading treatment dummy in the *Hauptschule* model (i.e. the confidence interval for $x = -3$). Here, the upper bound of the bootstrapped confidence interval is very close to zero. In general, the evidence presented in the event study plots is supportive of the hypothesis that the policy implementation did not result in anticipatory effects. The results are robust to the weighted estimation. Event study plots for the weighted estimation and for sub samples containing only males and pupils from low-educated households as well as their coefficients and confidence intervals for the leads and lags can be found in Appendix C.

An important concern with an observational study such as this is that the treatment and control groups differ materially across potentially relevant covariates. In order to assess the extent to which differences persist across treatment and control groups, I estimate the deviations from the mean of the control group for the pre-treatment and post-treatment treatment group for the share of males, the share of households in which the parents have at least some *Gymnasium*, and father's and mother's scores on the magnitude prestige scale (MPS), a measure of the prestige ascribed to different professions.[21] The *pre-treatment* treatment group collects all observations enrolling in school in one of the treatment states before the policy is implemented in that state. The *post-treatment* treatment group, conversely, groups individuals in the treatment states that enter school *after* the policy has been implemented. Thus, I effectively estimate the following model:

$$y_{sc} = \beta_1 \cdot PreTreatment_{sc} + \beta_2 \cdot PostTreatment_{sc} + \gamma_s + \lambda_c + \varepsilon_{sc} \tag{4.4}$$

where $y_{sc}$ is the respective dependent variable for cohort $c$ in state $s$, $\beta_1$ is the coefficient on the *PreTreatment* treatment group, $\beta_2$ is the coefficient on the *PostTreatment* treatment group, $\gamma_s$ is a set of state fixed effects, and $\lambda_c$ is a set of cohort fixed effects. Since I control for state and cohort fixed effects, the coefficients $\beta_1$ and $\beta_2$ effectively partial out any residual differences across the control group and the two treatment groups. The magnitude prestige scale scores could proxy as a further indicator for pupils' household background. However, they are not included in the regressions in section 5 because the information is missing for a substantial share of the observations. It cannot be ruled out that missing information on the MPS is correlated with educational outcomes, which would bias the estimates.

The results for the regressions from equation 4.4 are reported in Table 3. In the control states, 51.4% of individuals are male, with the pre-treatment and post-treatment treatment groups not significantly different.[22] 12.1% of households in the control group have at least one parent with

[21] The magnitude prestige scale is based on surveys in which representative cross sections of the German population are surveyed about the prestige they ascribe to a number of professions. See Wegener (1985). Scores on the magnitude prestige scale range from 30 (helping farmworkers) to 216 (dentists).

[22] As in the main estimations, model 4.4 is estimated with standard errors clustered at the state level. As the

**Table 3** Identification: Deviations of covariates relative to control group mean

| | Share of males | Share of households w/ some *Gymnasium* | Father's MPS | Mother's MPS |
|---|---|---|---|---|
| Mean control | 0.514 | 0.121 | 57.281 | 52.038 |
| | (0.009) | (0.014) | (0.944) | (1.378) |
| Pre-treatment | 0.001 | -0.004 | 1.295 | 1.539 |
| | (0.008) | (0.006) | (0.731) | (0.596) |
| Post-treatment | 0.003 | 0.001 | 0.051 | 0.754 |
| | (0.005) | (0.004) | (0.501) | (0.369) |
| P-val 'Pre-treatment' | 0.904 | 0.639 | 0.529 | 0.128 |
| P-val 'Post-treatment' | 0.565 | 0.931 | 0.914 | 0.472 |
| P-val 'Joint´ | 0.304 | 0.669 | 0.452 | 0.083 |
| *N* | 12898 | 12898 | 11422 | 7674 |

*Notes*: The reported standard errors are clustered at the state level. Bootstrapped significance levels are reported at the bottom of the table. "P-val 'Pre-treatment"' reports the Wild t cluster bootstrapped p-value for the 'Pre-treatment' group. "P-val 'Post-treatment"' similarly reports the bootstrapped p-value for the 'Post-treatment' group. "P-val 'Joint"' reports the bootstrapped p-value for a test of joint significance.

some *Gymnasium* and again, the treatment groups are not significantly different either. The average magnitude prestige score for the fathers' profession is 57.3 in the control group and roughly 58.6 and 57.8 in the pre-treatment and post-treatment treatment states respectively. The difference is not statistically significant. Mother's average magnitude prestige score in the control group is 52.0, compared with roughly 53.6 in the pre-treatment treatment states and 52.8 in the post-treatment treatment states. This difference is also not statistically significant.

## 4.2 Estimation

Turning to the main analysis, the hypothesis that later grading affects secondary school track and pupils degree is estimated within the following difference-in-differences model:

$$y_{isc} = \beta_1 \cdot D_{sc} + \gamma_s + \lambda_c + \beta_2 \cdot female_i + \delta \mathbf{X}_{isc} + \varepsilon_{isc} \tag{4.5}$$

where $y_{isc}$ denotes the dependent variable (*Gymnasium*, *Realschule* degree, or *Hauptschule* degree), $D_{sc}$ denotes the treatment and equals 1 if state $s$ graded students from cohort $c$ from 3rd grade and 0 otherwise, $\gamma_s$ denotes a set of state fixed effects, $\lambda_c$ denotes a set of cohort fixed effects, $female_i$ is a female dummy that equals 1 if an individual is female and 0 otherwise, and $\mathbf{X}_{isc}$ is a set of further control variables such as households' educational background and parental migration background. The likely role of parental educational background was discussed in the introduction. Apart from parents' education, parental migration background likely also affects pupils track choice. Migrant parents may be on average more or less educated than the general population or subject their children to unique expectations with respect to their educational attainment.[23]

I use a linear probability model for the main specification in equation 4.5 and for all sub-

number of clusters (6) is too low to provide sufficient asymptotic approximation, which leads to over-rejection, I report Wild t bootstrapped p-values at the bottom of Table 3. See discussion in section 4.3.

[23]For instance, Siahaan et al. (2014) find that in the United States, pupils with an immigrant background have higher educational attainment than natives. Given the geographical divergence of migration patterns and immigration policy, it is not clear that these findings transfer to the German context.

sequent variations. The dependent variables variables are observed binary variables indicating the obtained school degree. The response probability, i.e. the probability that the dependent variable is equal to 1, is linear in the parameters. This also makes the residuals naturally heteroskedastic (Wooldridge, 2012, p. 249). An alternative approach using a logit or probit estimation is beyond the scope of this thesis.

One concern regarding identification is that pupils may have self-selected into the treatment and control group; i.e. that treatment status is endogenous to unobserved covariates such as ability or household characteristics. This is likely not an issue. School attendance in Germany is linked to one's place of residence. Thus, self-selecting into treatment or control group requires moving one's family either out of or into a treatment state, depending on preference-status for treatment. However, the final decisions to change grading practices in the first two years of elementary school were only made shortly before the start of the new school year, leaving little time for parents to move their children to a different state.[24] Furthermore, general interstate mobility is rather limited in Germany (Cygan-Rehm, 2018, cf.) and relocation is costly to parents. The expected benefits/cost of the policy reform are likely not clear enough to outweigh the cost-considerations of relocation. The households most likely to respond to the policy by moving to another state are households living close to the border of a neighboring state, which should be a negligible proportion.

### 4.3 Standard errors

Another concern regarding the empirical strategy relates to the appropriate standard errors. In difference-in-difference research designs exploiting variation across states and years, as proposed here, Bertrand et al. (2004) point to a possibly severe serial correlation problem. Difference-in-difference estimations often rely on long time series (here: 23 years), dependent variables that are often highly positively serially correlated, and limited variation in the treatment variable within states or over time. These three factors reinforce each other and lead to underestimated standard deviations (i.e. over-rejection of the null hypothesis) under normal or heteroskedasticity-robust standard errors. As a solution to this problem of a serially correlated error-term, Bertrand et al. (2004) propose to calculate cluster-robust standard errors that permit for heteroskedasticity and within-cluster error correlation, clustering on state rather than on state-year. However, this is only a viable solution if the number of clusters is sufficiently large (e.g. around 50 clusters) because the asymptotic approximation relevant for clustered data relies on a large number of clusters (Angrist and Pischke, 2008, p. 222).[25]

With too few clusters, OLS leads to overfitting, with the estimated residuals systematically too close to zero. Secondly, even with bias correction, the cluster-robust estimate of the variance matrix leads to overrejection (Cameron and Miller, 2015, p. 24). Following the above guidance

[24] The reforms were finalized in Lower Saxony on May 26th, 1977, in North Rhine-Westphalia on May 30th 1979, in Hesse on December 30th, 1980, and in Rhineland-Palatinate on July 21st 1988. See North Rhine-Westphalia (1979); Hesse (1980); Rhineland-Palatinate (1988); Lower Saxony (1977).

[25] Similarly, Mackinnon and Webb (2017) point out that the cluster-robust variance estimator is consistent under the three key assumptions that (1) the number of clusters goes to infinity, (2) the within-cluster error correlations are the same for all clusters, and (3) each cluster contains an equal number of observations (Mackinnon and Webb, 2017, p. 233).

in my empirical strategy would imply clustering by state.[26] However, there are only 6 states in my sample. Given the cluster size requirement, the estimate of clustered standard errors in this case would be biased and likely lead to overrejection. A solution is provided by Cameron et al. (2008), who show that Wild cluster bootstrap-t procedures provide asymptotic refinement and reduce the rejection rate to the nominal size of 5% for as few as 6 clusters.[27] The main specification is reported with standard errors clustered at the state level. The appropriate Wild t bootstrapped p-values are reported separately at the bottom of the tables.[28]

# 5 Results

## 5.1 Main results

This section discusses the baseline results as well as the models analyzing effect heterogeneity across gender and parental educational background. The results are presented in tables 4, 5, and 6. The tables are all similarly structured and contain four dependent variables: *Gymnasium*, *Realschule*, *Hauptschule*, and > *Hauptschule* (which groups *Gymnasium* and *Realschule*). Every individual in the sample graduated from one of these school types (except for 2% of individuals who have not obtained any degree). Therefore, a positive coefficient for one dependent variable has to appear as a negative coefficient for another (with a small margin of error due to dropouts). Importantly, it is mechanically impossible for all coefficients across the school types to be of the same sign. The coefficients are reported with standard errors clustered by state. As discussed above, clustered standard errors likely over-reject the null hypothesis of no effect. Therefore, bootstrapped p-values are reported at the bottom of the tables for the treatment variable and the relevant interaction term and the sum of the treatment coefficient and the interaction term, if applicable. I break with convention and do not indicate significance levels with stars in the table. Instead, bootstrapped p-values significant at or below the 5% level are reported in bold. I report the bootstrapped 95% confidence intervals for the treatment variable and for the sum of the treatment and interaction term, if applicable. I also report the mean of the control group for each respective school type to illustrate relative effect sizes.

Table 4 presents the baseline results of the estimation. Each outcome variable is estimated within a small model that only accounts for the treatment effect, state and cohort fixed effects, and a gender effect (*Basic* model, columns 1, 3, and 5) and a general model that furthermore controls for parental educational background (*ParentEduc*)[29] and whether or not an individual

[26] The common approach in the literature on economics of education in Germany is to cluster by state x year-of-birth cells (i.e. separate clusters for each birth cohort for each state, see Piopiunik (2014a); Pischke and Von Wachter (2008); Cygan-Rehm (2018)). This is explicitly not recommended by Bertrand et al. (2004).

[27] The Wild bootstrap is a kind of residual bootstrap that draws $\mathbf{X}_i'\hat{\beta} + \hat{e}_i$ with probability 0.5 and $\mathbf{X}_i'\hat{\beta} - \hat{e}_i$ otherwise. This preserves the relationship between the residual variances and the $\mathbf{X_i}$ in the original sample (Angrist and Pischke, 2008, p. 226). *Asymptotic refinement* refers to the quality that the sampling distribution obtained from the bootstrap is actually closer to the finite-sample distribution than the asymptotic approximation (Angrist and Pischke, 2008, p. 227). On the other hand, standard errors calculated from bootstrapped samples do not provide asymptotic refinement.

[28] The cluster bootstrap is implemented in Stata using the `boottest` command: `boottest 1.treatment=0, cluster(location) weight(webb) nograph seed(10101)`. See Roodman (2015).

[29] *ParentEduc* groups all observations with valid parental educational information (includes observations where "don't know" is a valid answer, that are indicated as "other", or "no degree") and is 1 for all observations where

has a second generation migrant background (*ParentMig*, *General* model, columns 2, 4, 6, and 7).[30] The point estimate of the treatment effect on students propensity to graduate from the *Gymnasium* in the basic model is -0.006, but is not statistically significantly different from zero. The bootstrapped 95% confidence interval indicates that later grading did not decrease the propensity to obtain a *Gymnasium* degree by more than 4.3 percentage points or increase it by more than 4.9 percentage points, relative to a control group mean of 27.9 percent. Including the background control and the migrant dummy decreases the point estimate slightly to -0.008, but this estimate is also not statistically significantly different from zero with a similar confidence interval. The point estimates for the treatment effect on the propensity to obtain a *Realschule* degree (including the subsequent *Fachhochschulreife*) are 0.023 and 0.024 for the basic and general model respectively. Neither of these estimates are statistically significantly different from zero either. Based on the bootstrapped confidence interval of the *General* model, this indicates a decrease no larger than 4.9 percentage points and an increase no larger than 6.7 percentage points, relative to a control group mean of 41 percent. Finally, for the *Hauptschule* degree, the point estimate of the basic model is -0.022. Including the additional background controls hardly changes the point estimate to -0.021. However, neither of these coefficients are significantly different from zero. Based on the bootstrapped confidence interval for the *General* model, I can rule out a decrease larger than 4.9 percentage points, and an increase larger than 0.3 percentage points, relative to a control group mean of 29.4 percent. The point estimate for the grouped *Gymnasium* and *Realschule* degrees in column 7 is with 0.015 also insignificant. Based on the bootstrapped confidence interval I can rule out a decrease larger than 1.9 percentage points and an increase larger than 6.1 percentage points, relative to a control group mean of 68.9 percent. Thus, on the aggregate level, I fail to reject the null hypothethis that later grading does not affect pupils' track choice.

The question remains for now whether this failure to reject a zero effect masks treatment heterogeneity across some subgroups. Two candidates for effect heterogeneity present themselves: gender and parental educational background. Another candidate for group heterogeneity analysis would be parental migration background. A separate analysis of the effect heterogeneity over this variable is beyond the scope of this thesis.

## 5.2 Heterogenous treatment effects by gender

A large literature in economics of education has been concerned with gender differences, both in terms of absolute achievement and of relative response to policy reforms. In the context of grading specifically, girls appear to be subject to lighter, more favourable grading by their teachers (see Bonesrønning, 2008; Falch and Naper, 2013; Rangvid, 2015). A general gender difference in educational attainment is also evident in the point estimate of the female dummy in Table 4: females are around 8.9 percentage points more likely to obtain an intermediate secondary degree and around 9.1 percentage points less likely to obtain only a basic secondary degree, compared to males (holding constant state and cohort fixed effects, migrant status and

at least one parent has a *Gymnasium* degree.

[30] Migrant background, in this respect, refers only to second generation migrants, as the dataset is restricted to individuals born in Germany. See section 3.

**Table 4** Results: Basic difference-in-differences

| | Gymnasium | | Realschule | | Hauptschule | | > Hauptschule |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| | Basic | General | Basic | General | Basic | General | General |
| Treatment | -0.006 | -0.008 | 0.023 | 0.024 | -0.022 | -0.021 | 0.015 |
| | (0.017) | (0.017) | (0.017) | (0.017) | (0.009) | (0.010) | (0.014) |
| Female | 0.011 | 0.008 | 0.089 | 0.089 | -0.093 | -0.091 | 0.098 |
| | (0.008) | (0.008) | (0.011) | (0.011) | (0.011) | (0.010) | (0.009) |
| ParentEduc | | 0.456 | | -0.186 | | -0.256 | 0.271 |
| | | (0.005) | | (0.030) | | (0.027) | (0.028) |
| ParentMig | | -0.050 | | -0.052 | | 0.089 | -0.102 |
| | | (0.015) | | (0.019) | | (0.016) | (0.018) |
| State FE | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x |
| P-val 'treatment' | 0.790 | 0.703 | 0.318 | 0.284 | 0.113 | 0.093 | 0.393 |
| P-val 'female' | | 0.291 | | **0.011** | | **0.011** | **0.010** |
| CI 'treatment' | [-.043, .049] | [-.042, .055] | [-.05, .07] | [-.049, .067] | [-.056, .005] | [-.049, .003] | [-.019, .061] |
| Control group mean | 0.279 | 0.279 | 0.410 | 0.410 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parentheses. Standard errors are clustered by state. 'Basic' model only contains treatment variable, gender dummy, and state and cohort fixed effects. 'General' model includes background covariates: 'ParentEduc' is a dummy variable that equals 1 if at least one parent obtained the Abitur, zero otherwise. 'ParentMig' is a dummy variable that equals 1 if one of the individuals' parents immigrated to Germany (but not the individual themselves), zero otherwise. "P-val 'treatment"' is the bootstrapped p-value for the coefficient on the treatment variable. 'CI 'treatment" indicates the bootstrapped 95% confidence interval for the coefficient on the treatment variable.

household educational background). Given this gender disparity, it is important to consider whether later grading in elementary school affects males and females differently.

Table 5 extends the baseline regression to account for effect heterogeneity across genders. The *General* model in columns 1, 4, and 7 is the *General* model from Table 4 for comparison. Columns 2, 5, and 8 (*Gender interacted*) extend the *General* model to include an interaction term between the female dummy and the treatment variable, to partial out the difference of the effects between genders. Columns 3, 6, and 9 furthermore add a full set of interaction terms between the female dummy and the control variables, including with state and cohort fixed effects. The results are then equivalent to results obtained with an alternative regression that splits the sample into males and females.

As is the case in the baseline model, I fail to reject the null hypothesis of no effect of later grading on *Gymnasium* (columns 1–3) for both males and females. Regarding the *Realschule* degree, including the interaction term between gender and treatment increases the point estimate on the treatment variable to 0.056. Including the full set of interactions further increases the point estimate to 0.061. These coefficients, however, are not significant at any conventional significance level. I thus cannot reject the null hypothesis that later grading affects males' propensity to obtain a *Realschule* degree. The coefficients on the interaction term between female and treatment are with $-0.061$ and $-0.066$ of the opposite sign than the treatment coefficients. These interaction terms are significant at the 5% level. The sum of the treatment coefficients and the interaction terms, which yields the treatment effect on females, is with $-0.005$ in both models virtually zero and insignificant. Based on the bootstrapped confidence intervals I can reject a decrease smaller than 7.9 and 7.0 percentage points, and an increase larger than 3.6 and 4.4 percentage points, for both models respectively.

The results for the *Hauptschule* degree paint a different picture. Including only the interaction term reduces the magnitude of the treatment coefficient to $-0.041$. Including the full set of interaction terms further increases the magnitude of the coefficient to $-0.067$. Both estimates are significant at the 5% level and indicate that later graded boys are, on average and all else equal, between 4.1 and 6.7 percentage points less likely than earlier graded boys to obtain a *Hauptschule* degree, relative to a control group mean of 29.4 percent. For females, I cannot reject the null hypothesis of no effect of later grading on their propensity to obtain a *Hauptschule* degree. While the interaction terms in model 8 and 9 are significant at the 5% and 1% level, the sum of the treatment coefficient and the interaction term is with $-0.005$ and 0.015 not significantly different from zero. I thus cannot reject the null hypothesis that later grading affects females' propensity to obtain a *Hauptschule* degree. Based on the bootstrapped confidence intervals for the fully interacted model I can rule out a decrease larger than 2.5 percentage points and an increase larger than 5.1 percentage points, relative to a control group mean of 29.4 percent.

Finally, column 10 analyses the extent to which later grading affects pupils' propensity to obtain a degree higher than the *Hauptschule* degree. For males, the results indicate that later grading improves their propensity to obtain a higher degree, on average and all else equal, by 6.0 percentage points, relative to a control group mean of 68.9 percent. This estimate is significant at the 5% level. For females, I fail to reject the null hypothesis of no effect. Even though the interaction term of $-0.081$ is significant at the 1% level, the sum of the treatment and

**Table 5** Results: Heterogenous effects by gender

| | Gymnasium | | | Realschule | | | Hauptschule | | | > Hauptschule |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | General | Gender interacted | Fully interacted | General | Gender interacted | Fully interacted | General | Gender interacted | Fully interacted | Fully interacted |
| Treatment | -0.008 | -0.025 | -0.001 | 0.024 | 0.056 | 0.061 | -0.021 | -0.041 | -0.067 | 0.060 |
| | (0.017) | (0.014) | (0.016) | (0.017) | (0.018) | (0.024) | (0.010) | (0.005) | (0.011) | (0.015) |
| Female | 0.008 | -0.002 | -0.027 | 0.089 | 0.109 | 0.137 | -0.091 | -0.102 | -0.110 | 0.110 |
| | (0.008) | (0.007) | (0.025) | (0.011) | (0.010) | (0.029) | (0.010) | (0.006) | (0.018) | (0.021) |
| Female x Treatment | | 0.031 | -0.015 | | -0.061 | -0.066 | | 0.036 | 0.082 | -0.081 |
| | | (0.022) | (0.028) | | (0.010) | (0.030) | | (0.012) | (0.020) | (0.023) |
| State FE | x | x | x | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x | x | x | x |
| Full interactions | | | x | | | x | | | x | x |
| P-val 'treatment' | 0.703 | 0.204 | 0.961 | 0.284 | 0.084 | 0.088 | 0.093 | **0.013** | **0.012** | **0.012** |
| P-val 'interaction' | | 0.164 | 0.618 | | **0.018** | **0.011** | | **0.016** | **0.009** | **0.008** |
| P-val 'treatment + interaction' | | 0.849 | 0.704 | | 0.831 | 0.859 | | 0.859 | 0.470 | 0.591 |
| P-val 'F-test' | | 0.813 | 0.580 | | 0.807 | 0.820 | | 0.775 | 0.370 | 0.372 |
| CI 'treatment' | [-.042, .055] | [-.058, .034] | [-.047, .064] | [-.049, .067] | [-.022, .108] | [-.024, .143] | [-.049, .003] | [-.051, -.022] | [-.107, -.046] | [.03, .112] |
| CI 'treatment + interaction' | | [-.043, .078] | [-.069, .062] | | [-.079, .036] | [-.07, .044] | | [-.045, .029] | [-.025, .051] | [-.066, .031] |
| Control group mean | 0.279 | 0.279 | 0.279 | 0.410 | 0.410 | 0.410 | 0.294 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parantheses. Standard errors are clustered by state. 'General' models are models with only treatment variable and covariates. 'Gender interacted' models add an interaction term between treatment and female dummy. 'Fully interacted' models add further interaction terms between female dummy and all covariates, including state and cohort fixed effects. The results are identical to running separate regressions for males and females. In the 'interacted' models, coefficient on 'Treatment' variable is the treatment effect on males. The treatment effect for females is the sum of the coefficient on the 'Treatment' variable and the interaction term 'Female x Treatment'. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'Female x Treatment', and 'Treatment + (Female x Treatment)'. "P-val 'F-test"' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. 'CI 'treatment" and 'CI 'treatment + interaction" indicate the bootstrapped 95% confidence intervals for the coefficients on the treatment variable and the sum of the coefficients on the treatment variable and the interaction term.

the interaction term coefficient of −0.021 is not significantly different from zero. Based on the bootstrapped confidence interval I can rule out a decrease larger than 6.6 percentage points and an increase larger than 3.1 percentage points, relative to a control group mean of 68.9 percent.

In summary, the results from the decomposition of the effect heterogeneity across genders show evidence that later grading affects males to a greater degree than females and that this effect is concentrated at the lower end of the tracks. Later graded males are, on average and all else equal, between 4.1 and 6.7 percentage points less likely to obtain a *Hauptschule* degree compared to earlier graded males, and around 6 percentage points more likely to obtain a degree higher than the *Hauptschule* degree. For females, I fail to reject a zero effect for all dependent variables.

### 5.3 Heterogenous treatment effects by background

Another concern is that treatment may affect pupils from backgrounds with lower education differently than pupils from educated backgrounds. Using data from the SOEP covering six decades, Dustmann (2004) finds a strong relationship between parental background and secondary school track choice, similar to the association indicated by the *ParentEduc* control in the *General* models of Table 4. A similar association is reported by Ermisch and Francesconi (2001) using data from the British Household Panel Study. *Educated background*, in the context of this analysis, refers to households where at least one of the parents obtained an *Abitur*. In the sample, this is true for roughly 16% of the individuals (see Table 1). Table 6 extends the basic model to account for the potential effect heterogeneity across pupils' household background. Columns 1, 4, and 7 again depict the *General* model from Table 4 for comparative purposes. Columns 2, 5, and 8 (*Background interacted*) extend the model to include an interaction term between the treatment variable and the background control *ParentEduc*. Columns 3, 6, and 9 (*Fully interacted*) further extend this model to include a full set of interaction terms between the background control and the other control variables (*Female* and *ParentMig* dummies), including the state and cohort fixed effects. As with the gender heterogeneity model previously discussed, the results are equivalent to two separate estimations of the effects of later grading on a sub sample of pupils from low-educated households and on a sub sample of pupils from educated households.

For the first set of regressions with the dependent variable *Gymnasium* degree, I fail to reject the null hypothesis of no effect both for pupils from low-educated backgrounds (coefficient on the treatment variable) and for pupils from educated households (sum of the coefficients on the treatment variable and the interaction term) for both the background interacted and the fully interacted specification. For pupils from low-educated households I can exclude a decrease larger than 3.5 percentage points and an increase larger than 5.3 percentage points, based on the bootstrapped confidence interval for the fully interacted model. For pupils from educated households, I can exclude a decrease larger than 11.3 percentage points and an increase larger than 9.8 percentage points, relative to a control group mean of 27.9 percent.

Regarding the dependent variable *Realschule* degree, I also fail to reject the null hypothesis of no effect for both the educated and the low-educated group. The coefficient on the treatment

**Table 6** Results: Heterogenous effects by background

| | Gymnasium | | | Realschule | | | Hauptschule | | | > Hauptschule |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | General | Background interacted | Fully interacted | General | Background interacted | Fully interacted | General | Background interacted | Fully interacted | Fully interacted |
| Treatment | -0.008 | -0.009 | -0.005 | 0.024 | 0.040 | 0.036 | -0.021 | -0.037 | -0.036 | 0.031 |
| | (0.017) | (0.018) | (0.016) | (0.017) | (0.019) | (0.016) | (0.010) | (0.012) | (0.012) | (0.017) |
| ParentEduc | 0.456 | 0.454 | 0.378 | -0.186 | -0.153 | 0.007 | -0.256 | -0.286 | -0.367 | 0.385 |
| | (0.005) | (0.006) | (0.035) | (0.030) | (0.027) | (0.024) | (0.027) | (0.024) | (0.021) | (0.025) |
| ParentEduc x Treatment | | 0.006 | -0.024 | | -0.096 | -0.040 | | 0.089 | 0.063 | -0.064 |
| | | (0.014) | (0.030) | | (0.027) | (0.032) | | (0.023) | (0.015) | (0.021) |
| State FE | x | x | x | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x | x | x | x |
| Full interactions | | | x | | | x | | | x | x |
| P-val 'treatment' | 0.703 | 0.702 | 0.792 | 0.284 | 0.134 | 0.130 | 0.093 | **0.018** | **0.020** | 0.184 |
| P-val 'interaction' | | 0.731 | 0.561 | | **0.034** | 0.188 | | **0.025** | **0.015** | **0.014** |
| P-val 'treatment + interaction' | | 0.855 | 0.526 | | 0.050 | 0.915 | | **0.025** | **0.017** | **0.018** |
| P-val 'F-test' | | 0.856 | 0.453 | | 0.066 | 0.914 | | **0.015** | **0.009** | **0.004** |
| CI 'treatment' | [-.042, .055] | [-.044, .053] | [-.035, .051] | [-.049, .067] | [-.035, .091] | [-.033, .081] | [-.049, .003] | [-.069, -.008] | [-.068, -.006] | [-.011, .078] |
| CI 'treatment + interaction" | | [-.037, .101] | [-.113, .098] | | [-.149, 0] | [-.157, .078] | | [.011, .102] | [.013, .05] | [-.06, -.02] |
| Control group mean | 0.279 | 0.279 | 0.279 | 0.410 | 0.410 | 0.410 | 0.294 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parantheses. Standard errors are clustered by state. 'General' models are models with only treatment variable and covariates. 'Background interacted' models add an interaction term between treatment and the background dummy 'ParentEduc' which is 1 if either parent obtained an Abitur, 0 otherwise. 'Fully interacted' models add further interaction terms between the background dummy and all covariates, including state and cohort fixed effects. The results are identical to running separate regressions for educated and low-educated households. In the 'interacted' models, coefficient on 'Treatment' variable is the treatment effect on pupils from non-educated households (i.e. households where neither parent obtained an *Abitur*). The treatment effect for pupils from educated households is the sum of the coefficient on the 'Treatment' variable and the interaction term 'ParentEduc x Treatment'. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'ParentEduc x Treatment', and 'Treatment + (ParentEduc x Treatment)'. "P-val 'F-test"' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. 'CI 'treatment" and 'CI 'treatment + interaction" indicate the bootstrapped 95% confidence intervals for the coefficients on the treatment variable and the sum of the coefficients on the treatment variable and the interaction term.

variable is with 0.040 and 0.036 similar across the interacted and fully interacted models. The sum of the coefficients on the treatment variable and the interaction term, which yields the treatment effect on pupils from educated households, varies more. The estimate of −0.056 in column 5 is just not significant at the 5% level. The estimate of −0.004 in column 6, on the other hand, is insignificant.

For the dependent variable *Hauptschule* degree, including the interaction term increases the size of the point estimate on the treatment variable to −0.037. This point estimate is significant at the 5% level. Including the full set of interaction terms changes the point estimate only slightly to −0.036, which is also significant at the 5% level. Thus, I can reject the null hypothesis of no effect for pupils from low-educated households at the 5% level. This indicates that later graded pupils from low-educated households are, on average and all else equal, around 3.6 percentage points less likely than their earlier graded peers to obtain a *Hauptschule* degree. Regarding the effect of later grading on pupils from educated households, I observe the opposite effect. In the background interacted model in column 8, the coefficient on the interaction term is 0.089. Together with the coefficient on the treatment variable, this sums to an estimated effect on pupils from educated households of 0.052, significant at the 5% level. In the fully interacted model, this effect is estimated as 0.027, which is also significant at the 5% level. I can therefore reject the null hypothesis of no effect on pupils from educated households at the 5% level. The results indicate that later graded pupils from educated households are, on average and all else equal, between 2.7 and 5.2 percentage points more likely than their earlier graded peers to obtain a *Hauptschule* degree.

Finally, regarding the grouped dependent variable for all degrees higher than the *Hauptschule* degree, the point estimate of 0.031 is insignificant. For pupils from low-educated households, I therefore cannot reject the null hypothesis of a zero effect of later grading on the propensity to obtain a degree higher than the *Hauptschule*. For pupils from educated households, on the other hand, the results indicate that later grading decreases their propensity to obtain a degree higher than the *Hauptschule* degree, on average and all else equal, by 3.3 percentage points. This estimate is significant at the 5% level.

These results thus indicate that while there is no evidence of an effect of later grading on pupils propensity to obtain either a *Gymnasium* or a *Realschule* degree, later grading does seem to affect pupils propensity to pursue only a *Hauptschule* degree. For pupils from low-educated households, the evidence only supports a decreased propensity to obtain a *Hauptschule* degree but offers no evidence of a propensity to obtain a higher degree. For pupils from educated households, on the other hand, later grading decreases their propensity to obtain a degree higher than the *Hauptschule* degree by around 3.3 percentage points, relative to a control group mean of 68.9 percent.

In summary, the baseline results persistently fail to reject a zero effect on the dependent variable *Gymnasium* degree. The results for the dependent variable *Realschule* degree similarly fail to reject a zero average effect. I furthermore cannot reject a zero average effect on the propensity to obtain a *Hauptschule* degree. Later graded *males*, however, as well as later graded pupils from low-educated households exhibit a decreased propensity to obtain the basic degree. Later graded pupils from educated households are found to be more likely to obtain a *Hauptschule*

degree. For males, this translates into a statistically significantly increased propensity to obtain a degree higher than the *Hauptschule* degree. For pupils from educated households, on the other hand, the results indicate a decreased propensity to obtain such a higher degree.

# 6 Robustness

The econometric model deployed thus far has been subject to a number of modelling choices, discussed in section 3. In this section I outline a number of alternatives and discuss the sensitivity of the results to these alternative specifications. First, I discuss the sensitivity of the results to an alternative estimation that accounts for the weights provided by the SOEP. Second, I discuss the sensitivity of the results to alternative data imputations and sample constructions.

## 6.1 Weighted vs unweighted estimation

Section 3 briefly discussed why a weighted estimation might be sensible given the SOEP data and the weights provided. A researcher may consider weighing the estimation by a specific weighting factor if the probability of selection varies with the dependent variable even after conditioning on the explanatory variabels (Solon et al., 2015). Some population groups are oversampled in a number of subsamples contained in the SOEP sample used in this thesis. These groups are, for instance, single parent households, or low/high income households. Since I cannot control for these characteristics (the SOEP does not contain historical information on income, for instance), this might introduce bias into my estimates if the probability of pursuing a certain secondary school track varies along these characteristics. School track choice may be correlated with household characteristics if single parent households are less able to assist their children in their academic development. Income is also likely a strong determinant of educational attainment, both through the resources potentially allocated to a child and through the greater expectation that richer parents may have for their childrens' education. In order to assess the sensitivity of the main results to this alternative approach I re-estimate the results for Tables 4, 5, and 6. The full results and a more thorough discussion of the weighted estimation can be found in Tables B.1, B.2, and B.3 in Appendix B.

The results are qualitatively similar across the two approaches. The effect signs are robust to the alternative estimation, as are the magnitudes of the estimates. For the dependent variable *Gymnasium*, in both weighted and unweighted estimates, I fail to reject the null hypothesis of a zero effect for the general, as well as the gender- and background-heterogeneity models. For the *Realschule* model I also consistently fail to reject the null hypothesis of a zero average effect as well as for an effect for males or females. For pupils from low-educated households, the weighted estimation indicates a positive effect between 3.0 and 4.8 percentage points, significant at the 5% level. Regarding the dependent variable *Hauptschule* degree, the unweighted estimation fails to provide evidence of a non-zero average effect, while the weighted estimation yields a negative effect between 5.1 and 5.2 percentage points, significant at the 5% level. Furthermore, later graded *males* are observed to have a consistently and statistically significant reduced propensity to obtain a *Hauptschule* degree (between 6.2 and 7.7 percentage points in the weighted estima-

tion), as do treated pupils from low-educated households, regardless of the weighting decision (effects around 3.7 percentage points in the unweighted estimation compared to between 5.2 and 7.0 percentage points in the weighted estimation). Importantly, the point estimate for pupils from low-educated backgrounds in the weighted estimation is about twice as large as the point estimate in the unweighted estimation. For males, this translates into a 6.4 percentage points increased propensity to obtain a *higher* degree, statistically significantly different from zero at the 1% level in the weighted estimation, compared to a 6.0 percentage points increased propensity in the unweighted estimation, significant at the 5% level. For pupils from low-educated household, there is only suggestive evidence for an increased propensity to obtain a *higher* degree. The effect is only distinguishable from zero at the 5% level in the weighted estimation. For later graded pupils from educated households, the unweighted estimates indicated a 3.3 percentage points decreased likelihood to obtain a degree higher than the *Hauptschule* degree and a 2.7 percentage points increased likelihood to obtain a *Hauptschule* degree, but these effects disappear in the weighted estimation.

### 6.2 Alternative data imputation and sample constructions

Apart from the decision whether to weigh the estimation, a set of choices relating to the data and sample construction may skew the estimates in other ways. In this subsection I briefly discuss the robustness of the main estimates to six alternatives. An extended discussion is included in Appendix B. In general, the results are robust in terms of effect sign and magnitude. The largest variation occurs with respect to the treatment estimates for pupils from low-educated households. The results of the individual robustness tests are presented in Tables B.4, B.5, and B.6 in Appendix B. Table B.4 reports only the treatment estimates from the primary model and the robustness models, while Tables B.5 and B.6 also report the estimates for the respective interaction terms. The results in each table are reported in four Panels. Panel A reports the results for the dependent variable *Gymnasium*. Panel B reports the results for the dependent variable *Realschule*. Panel C reports the results for the dependent variable *Hauptschule*. Panel D reports the results for the dependent variable *higher than Hauptschule*. The model underlying the estimates in Table B.4 is the *General* model, including a full set of covariates. The models underlying the estimates in Tables B.5 and B.6 are the *Fully interacted* models, including a full set of covariates and interaction terms.

In a first alternative (Column 2) I drop all observations from Rhineland-Palatinate and all cohorts entering school after 1986 because of the large time gap between the policy implementations in Rhineland-Palatinate and the three other treatment states. This alternative does not significantly change the results. In a second robustness check (Column 3) I am concerned with measurement error around the school entry of individuals for whom the month of birth is missing. In the main analysis, these individuals were assumed to have entered school in the year they turned 7. The alternative would have been to assume that they entered school in the year they turned 6. As expected, the results under this alternative specification are almost identical to the main results. As a third robustness check (Column 4) I recode all observations with invalidly missing information on parental education as neither parent having obtained an *Abitur*

(i.e. *ParentEduc*= 0). The results are very similar in the general and the gender-heterogeneity models. Some differences surface in the background-heterogeneity model. The coefficient on the treatment variable in the *Hauptschule* model decreases in magnitude from -0.036 to -0.028. With a new bootstrapped p-value of 0.083, this new point estimate is no longer statistically significant. In a fourth robustness check (Column 5), I omit the first affected cohort to alleviate concerns around measurement error with the first affected cohort. Omitting these observations also decreases the magnitude of the point estimate of the treatment variable in the background-heterogeneity *Hauptschule* model from $-0.036$ to $-0.032$, which is statistically insignificant with a new bootstrapped p-value of 0.096. As a fifth robustness check (Column 6), I extend on the fourth robustness check and exclude all cohorts around the policy implementation (i.e. the cohorts entering school the year before the implementation and the year after). With this variation, the point estimate for the propensity of pupils from low-educated households to obtain a *Hauptschule* degree decreases from $-0.036$ to 0.034, which is no longer significant. A final concern (Column 7) relates to confounding reforms. The most likely confounding reform is the relaxation of the degree to which the schools' recommendation for a childs' secondary school track was binding, which Lower Saxony implemented in 1978, only one year after the late grading reform. I thus drop all observations for Lower Saxony from the sample. The results are robust across the models and panels with only small changes in effect size and no changes in the indicated significance levels.

In general, the estimates of the primary specifications presented in this thesis prove robust to the alternative specifications outlined above. I fail to reject zero effects on pupils propensity to pursue either a *Gymnasium* or a *Realschule* track. The analysis presented herein does offer suggestive evidence of a treatment effect on males concentrated at the lower end of the tracks. This effect is consistent across weighted and unweighted estimations and across the different sample variations. The estimated effect on pupils from low-educated backgrounds is less robust to the alternative estimations. On the one hand, for the sample variations, 3 out of 6 alternative point estimates for the *Hauptschule* model indicate insignificance (compared to significant results in the other three robustness estimations and the main estimation). On the other hand, all weighted estimates for the *Realschule*, *Hauptschule* and higher than *Hauptschule* models are statistically significant, compared to mostly insignificant results in the baseline estimation. The results pertaining to the effect on pupils from low-educated backgrounds should thus be interpreted cautiously. Regarding the results for pupils from educated backgrounds, there are no changes in the significance levels for any of the variations across the different samples. However, while the unweighted estimations yielded statistically significant results across the *Hauptschule* and higher than *Hauptschule* models, these effects disappear in the weighted estimation.

### 6.3 Advanced difference-in-differences diagnostics

The preceeding robustness checks have provided evidence of the stability of the estimates to alternative specifications. More advanced difference-in-difference diagnostics offer insights into the underlying properties of the estimators and thus allow to further assess the robustness of the results. The classical application of the difference-in-differences (DD) estimator has

been the two-period, two-group application that estimates the difference between the change in outcomes before and after a treatment between a treatment and a control group. Many applications differ from this simplified set-up in that across a longer time-horizon the treatment status varies across sub-groups. Some sub-groups tend to be treated earlier than others, while yet another sub-group may not be treated at all. The institutional set-up exploited in this thesis is an example of such a staggered adoption design. A recent literature has developed a set of advanced tools to analyse these properties with a special emphasis on estimators derived under time-heterogeneous treatment effects and staggered adoption designs. In this section I apply two of the most most relevant tools by calculating the weights attached to the average treatment effects in each group and period as proposed by de Chaisemartin and D'Haultfoeuille (2019), and by decomposing the difference-in-difference estimator into its 2x2 pairwise combinations as developed by Goodman-Bacon (2019).

de Chaisemartin and D'Haultfoeuille (2019) show that the estimate derived from two-way fixed effects difference-in-differences estimation under the common trends assumption is a weighted sum of the average treatment effect in each group and period. The control group in some of the individual comparisons of the outcome between consecutive time-periods across groups may be treated at both periods. If this is the case then the treatment effect at the second period gets differenced out by the difference-in-difference estimate, which can lead to a negative weight attached to that group-period estimate. Due to these negative weights, the coefficient from the linear regression may be negative while all the average treatment effects are in fact positive. Negative weights are of particular concern when "treatment effects differ between many vs few treated groups, or between groups treated for many vs few periods" (de Chaisemartin and D'Haultfoeuille, 2019, p. 9). The authors recommend to implement their alternative estimator, which is valid even if the treatment effect is heterogeneous over time or across groups, if many of the weights attached to the regression are negative. Thus, in order to assess whether in the context of my analysis the alternative estimator should be considered, I estimate the weights attached to the group-period clusters. I obtain 0 negative weights attached to a total of 55 average treatment effects on the treated (ATTs). This indicates that my results are robust to the potential limitation highlighted by de Chaisemartin and D'Haultfoeuille.[31]

Goodman-Bacon (2019) shows that the general estimator in a two-way fixed effects difference-in-differences estimation equals a weighted average of all possible two-group/two-period difference-in-differences estimators and supplies a decomposition that scatters the estimator against its weights across treatment and timing groups. This displays heterogeneity in the estimated components and clarifies which relationships and groups matter most. This helps, for example, to illustrate why coefficients change when some states are excluded. The constructed weights are proportional to group size and to the variance of the treatment dummy in each pair, meaning that units treated towards the middle of the panel are weighted highest.

[31]The results in de Chaisemartin and D'Haultfoeuille (2019) on two-way fixed effects regressions with controls apply to group x period level controls. Since my estimation relies on individual level data, the controls on gender, parental education, and parental migration background differ within the group x period cells. Therefore, the twowayfeweights-command replaces these control variables by their average value in each group x period cell. Stata code: `twowayfeweights hauptschule location entry_school treatment, type(feTR) controls(female hh_non_abitur mig)`.

I perform the Bacon Decomposition separately for each dependent variable and for separate samples of only males and only pupils from low-educated households (as these groups exhibited the strongest effects in my main analysis). The resulting plots are in Appendix D. The difference-in-differences estimators derived from the Bacon Decomposition differ slightly from the main estimates due to technical reasons related to the Stata command.[32] Because of the technical reasons, the precise weights attached to each 2x2 estimate estimated for my models are not informative. More informative are the magnitudes of the 2x2 estimates. Estimates scattered around the zero-line should thus be indicative of a failure to reject a zero effect. Estimates clustered to one side of the zero-line should be indicative of an effect with the respective sign.

The weights attached to the three types of groups (Earlier Group Treatment vs. Later Group Control (ET); Later Group Treatment vs. Earlier Group Control (LT); Treatment vs. Never Treated (TC)) are constant across the dependent variables and the models. The ET-2x2-estimates receive a weight of 0.160. This is to say that 16% of the DD-estimate from the Bacon Decomposition is derived from the ET-estimates. The LT-2x2-estimates receive a weight 0f 0.216. The TC-estimates receive a weight of 0.623. Thus, in the Bacon Decomposition, the majority of the DD-estimate is due to the Treatment vs. Never Treated comparison. The same likely holds for the individual level estimates in the main analysis, although this cannot be verified at this point.

Figure D.1 depicts the Bacon Decompositions for the *Basic Gymnasium* model (see Table 4, column 1), the *Gymnasium* model only for males, and the *Gymnasium* model only for pupils from low-educated backgrounds. The individual ET- and LT-estimate receive relatively small weights and a distributed on both sides of the 0-line. The TC-estimates receive more weight but are similarly split around the 0-line, with half the estimates below and half the estimates above the 0-line in all three models. This is indicative of the robust failure to reject a zero effect that the later grading reforms had on pupils propensity to obtain a *Gymnasium* degree. While the specific weights attached to the estimates are non-interpretable with respect to the primary results, the close clustering of the individual estimates indicates that the failure to reject a zero effect is not driven by individual outliers but is robust across the 2x2 pairs.

Figure D.2 depicts the analogous Bacon Decompositions for the *Basic Realschule* model, the *Realschule* model only for males, and the *Realschule* model only for pupils from low-educated backgrounds. As with the *Gymnasium* model above, the *Basic Realschule* decomposition is consistent with the zero-average effect. The individual ET- and LT-estimates are distributed around the 0-line, slightly negatively skewed. The TC-estimates indicate a slightly more positive treatment effect, with three of the four TC-estimates above the 0-line. For the *Realschule* model

[32] The decomposition developed by Goodman-Bacon (2019) and implemented in Stata by Goodman-Bacon et al. (2019) requires panel data. I thus have to collapse the individual observations of my sample to their means by cohort and state. The aggregate analogues of the individual level regressions yield the same coefficients when weighing by the number of observations in each cohort-state cell. Unfortunately, the decomposition-command does not allow for the inclusion of such weights. Thus, the resulting difference-in-difference estimates differ slightly from the main results. A second qualification relates to the actual implementation. The illustrative decomposition into "Early Group Treatment vs. Late Group Control", "Late Group Treatment vs. Early Group Control", and "Treatment vs. Never Treated" currently does not allow for control variables. Thus, the decompositions are performed on a set of regressions without controls, most closely similar to the "Basic" models in Table 4. Exemplary stata code: `bacondecomp hauptschule treatment, ddetail`.

for males, the Bacon Decomposition corroborates the positive skew of the treatment estimate. All of the TC-estimates are positive (between 0.04 an 0.15) and so are most of the individual ET- and LT-estimates. The decomposition for the model with only pupils from low-educated households paints a more mixed picture. The TC-estimates are still all positive but closer to zero, indicative of a less pronounced effect. The individual ET- and LT-estimates are roughly evenly distributed on either side of the 0-line.

Finally, Figure D.3 depicts the analogous decompositions for the *Hauptschule* models. Later graded males were found to be significantly less likely to obtain a *Hauptschule* degree, as were pupils from low-educated households. The Bacon Decompositions corroborate these findings. For the *Basic Hauptschule* model, all of the TC-estimates are negative, scattered roughly between -0.01 and -0.06. The individual ET- and LT-estimates, on the other hand, are roughly evenly distributed across the 0-line. For the *Hauptschule* model for males, the decomposition further corroborates the main results. All of the TC-estimates are negative, scattered between -0.04 and -0.11. The individual ET- and LT-estimates are almost all negative. The decomposition for the pupils from low-educated households, finally, also underscores the primary results. All of the TC-estimates are negative (between -0.03 and -0.07), while most of the individual ET- and LT-estimates are also negative. Compared to the model for the male sample, the effect is less pronounced, with more individual estimates close to or above zero.

# 7 Discussion

This section discusses the results presented above in relation to the small existing literature on the effects of grading reforms and considers potential mechanisms that work to explain the empirical findings. It concludes with a discussion of limitations of the research presented in this thesis and outlines avenues for further research.

The limited literature on comparable reforms has yielded inconclusive evidence of the effects of exposure to grading on pupils performance and educational attainment (see section 1). The naive expectation with respect to the treatment effect anticipated a positive effect on pupils from low-educated households, which the empirical results did not corroborate. Instead, the results presented in this thesis have offered suggestive evidence of a positive effect of later grading on males and a negative effect on pupils from educated households in the lower tracks. Males in particular are found to respond to later grading by slightly increasing their educational attainment by being marginally more likely to pursue a *Realschule* degree or higher as opposed to a basic *Hauptschule* degree. At the lower tracks, later grading appears to impact pupils from educated households less favourably than their peers from low-educated households. In this respect the findings are more in line with the findings reported by Klapp et al. (2014). Since this thesis has not explicitly analysed the effect of later grading on the propensity to complete high-school, a direct comparison with the primary finding of Sjögren (2010) is not possible. To the extent that completion and higher educational attainment are similar, the sign of the point estimates of the treatment effect estimated here does differ from the results of Sjögren (2010) and implies a different sign of the treatment effect on pupils from low-educated households. Given the suggestive nature of the evidence presented in this thesis, further work is required

to confidently establish the treatment effect of late grading reforms. I outline a number of promising research questions at the end of this section.

Another question altogether relates to the potential mechanism through which late grading may have affected track choice in the first place. I propose two mechanisms. One relates to the effect that later grading may have on pupils' engagement and performance. The other relates to the information available to parents which likely shapes their beliefs about their child's ability and future prospects. Both mechanisms are illustrated in Figure 3.

The first mechanism may work through pupils' performance. Some empirical research suggests that the motivational effect of grades is actually negative at the lower end and only positive at the upper end of the grade distribution (Betts and Grogger, 2003; You and Sharkey, 2009; Poorthuis et al., 2015). This is to say that high performing students tend to receive further motivation from good grades, while poorly performing students tend to become demotivated. Previous research indicates further that girls receive better grades in the classroom relative to their performance as evaluated by external standardized tests (Emanuelsson and Fischbein, 1986; Bonesrønning, 2008). Similarly, Rangvid (2015) finds that boys, pupils from low educated backgrounds, and migrants are systematically assessed lower by teacher scores than girls and pupils from educated backgrounds. Thus, as later graded males and pupils from low-educated backgrounds were exposed later to their on average likely worse grades, this likely affected their motivation, participation, and performance at the margin, which potentially translated into better performance and better secondary track recommendations.

Pupils' performance may have further been affected by a side effect of the late grading reforms – the crowding in of alternative forms of assessment and feedback. In the years and semesters before the pupils received number grades, they instead received written assessments of their relative strengths, weaknesses, potentials, and interests. To the extent that some pupils respond better to the type of feedback communicated through a written assessment compared to a simple number grade, the change in the exposure to written feedback induced by the postponed grading reforms may also have affected their educational performance. However, it is not clear that males or pupils from low-educated backgrounds respond differently to this form of feedback than do females or pupils from educated backgrounds. The evidence presented in this thesis does not amount to a formal analysis of this mechanism. Using supplementary SOEP data on reported measures of aptitude that may be correlated with an increase in motivation, ability, and performance, I fail to find evidence of this mechanism. The results are reported and discussed in Appendix E. However, given the preliminary nature of this analysis, little can be inferred from the insignificant results.

Another potential mechanism relates to the information available to parents to form their beliefs about their child's abilities and future prospects. It was briefly discussed above that performance alone does not determine a pupils' track choice. Rather, their parents' track desire also plays an important role. A large literature in development economics has found that parents' and youths' expectations of future earnings matter for enrollment decisions (Jensen, 2010; Attanasio and Kaufmann, 2014). An emerging literature in economics of education has furthermore explored the role of parents' beliefs about their childrens' abilities and their expectations of their childrens' performance on educational choices. For instance, in an experimental set-

ting in Malawi, Dizon-Ross (2019) finds that increasing parents information about their child's performance causes them to increase the school enrollment of their higher-performing children, decrease the enrollment of their lower-performing children, and choose educational inputs that are more closely matched to their children's academic level. The late grading reforms may thus have altered the quality of the information available to parents about their child's performance. Receiving grades later may have reduced the quality of the information available to parents, which would be expected to lead to a track desire less suited to the child's abilities. On the other hand, the supplementary written assessments of the child's strengths, weaknesses, and interests may have provided richer information to build beliefs about a child's abilities and future prospects, which would be expected to lead to a track desire more in line with the child's actual abilities. The net effect (if any) of this information mechanism remains unclear.

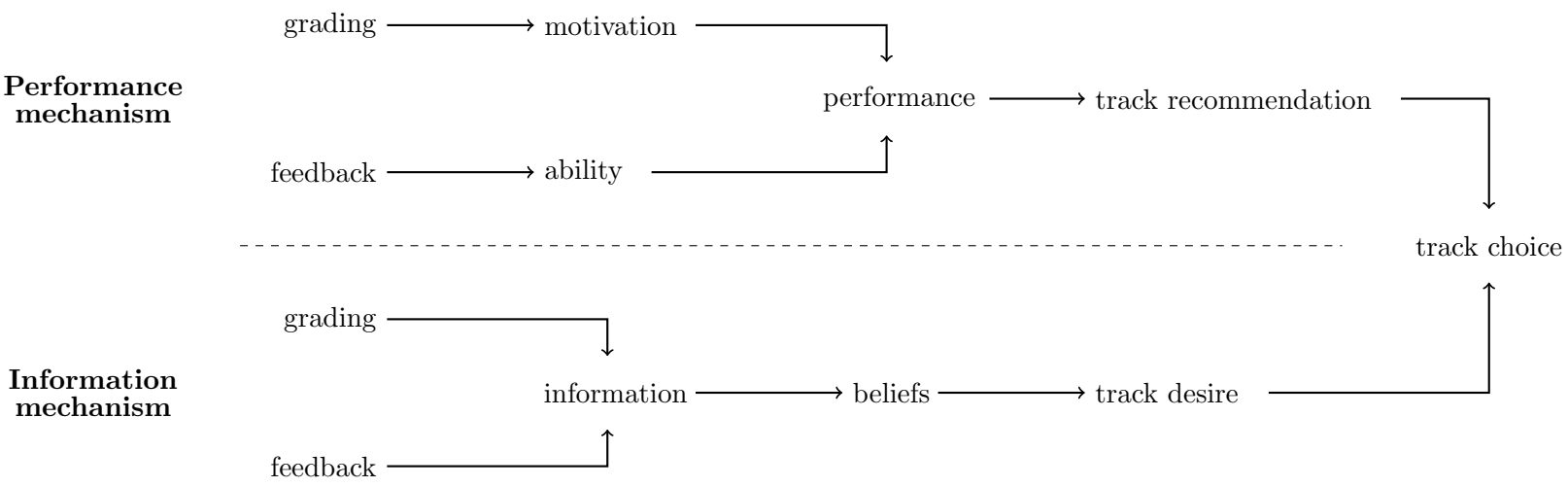
The suggestive evidence for an increased propensity of pupils from educated households to pursue a *Hauptschule* degree may offer a clue. In a German context, Piopiunik (2014a) find a significant association between increased parental education and parents' valuation of their children's education. Similarly, Schneider (2011) reports a high association between parents' socio-economic status and their child's probability of receiving a *Gymnasium* recommendation, even after controlling for performance. This suggests an upward bias for educated parents on their children's secondary school track. Assuming this upward bias, reducing the quality of information about a child's ability likely does not decrease their preferred secondary school track. It seems more likely that more accurate beliefs about a child's ability would decrease school track choice. This would point to the alternative form of feedback as an initiator of this mechanism.

An alternative explanation could be that the quality of the information available about a child's performance may change the additional resources that parents devote to their development. If a child performs badly in school and their parents are aware of this, they may invest more time and resources into homework, tutoring, or other beneficial activities. If parents become aware of the performance-gap only in Grade 3 instead of in Grade 2, the time left until the tracking decision may not suffice to catch up. Assuming that more educated parents are more likely to invest these resources if their child lags behind in their performance, this could explain the increased propensity of later graded pupils from educated households to pursue a *Hauptschule* degree. It would not, however, explain the observed gender gap.

## 7.1 Limitations and further research

The analysis presented in this thesis may be subject to a number of limitations. One limitation of the results may be due to the fact that I only observe the degree that an individual obtains, not the track they actually choose after Grade 4. This potentially introduces two sources of measurement error with opposite signs. Pupils may choose a higher track after elementary school and then change tracks during their secondary school years. As inter track mobility is usually limited to a downward mobility (Jürges and Schneider, 2007), this could lead to an under estimation of the track choice. On the other hand, pupils may pursue a consecutive degree after their first secondary school degree. The most common consecutive track after the

**Figure 3** Proposed mechanisms



*Realschule* is the *Fachhochschulreife*. I account for this consecutive track by coding individuals who I observe as having obtained a *Fachhochschulreife* as having pursued a *Realschule* degree first. Other consecutive track choices (such as pursuing an *Abitur* after a *Realschule* degree) are rather uncommon. To the extent that they are present, my estimates then over estimate the track choice for the higher degrees. An alternative would be to use a data set that directly observes the school track rather than the degree obtained but I am not aware of a data set that contains this information and sufficient background information.

Another concern with the research presented in this thesis may be that the heterogeneity analysis may increase the risk of identifying false positives due to multiple hypothesis testing. In order to alleviate concerns of p-hacking I have limited the analysis of the treatment effect heterogeneity to two subgroups commonly used in the economics of education literature and have refrained from conducting further heterogeneity analyses on more unusual subgroups.

One may also be concerned with the limited sample size of the SOEP, which leads to relatively few observations in each state-year cell. Given the representative sampling of the SOEP, the sample average of the respective degrees per state-year cell should at least approximate the population average, but it is not immediately clear that outliers do not skew the results. The limited sample size may furthermore decrease the precision of the estimates due to larger standard errors. This concern is evidenced by the relatively large confidence intervals discussed in section 5. An alternative, larger data set would be the German Micro Census, which annually samples 1% of German households. However, the size of the data set comes at a trade off: information on pupils' parental background is more limited in the Micro Census, as information on parents is only gathered while parents and child still live in the same household.[33] Nonetheless, it would be valuable to attempt to replicate the analysis of this thesis using data available from the Micro Census.

Another line of research, subject to data availability, should in more detail assess the effect of the late grading reforms on student performance directly. The results of this research would shed light on the respective mechanism that drives the effects. Researchers should also study the extent to which treatment effects differ by interactions of parental educational background and gender and the extent to which later grading affects pupils from migrant backgrounds. Given the trade offs indicated by the results on pupils from educated and low-educated backgrounds, it would furthermore be important to quantify the loss/increase in income due to alternative track choices. It could be that the potential earnings effect of a lower track choice for pupils from educated households may be mitigated by other mechanisms like their parents' professional connections. Conversely, any marginally improved educational outcomes for pupils from low-educated backgrounds may result in a substantial earnings effect. These cost considerations have to be left to future research.

Finally, it should be studied whether the effects found in this thesis also hold for reverse policy reforms. The Saarland enacted the late grading reform in 1994/95 only to move back to early grading in 1999/2000. Similarly, Hesse aborted its late grading policy in 1998/99. Further research should study the extent to which these opposite reforms resulted in the anticipated

[33] Furthermore, I am not eligible to access the Micro Census, as access to it is restricted to institutions in the Federal Republic of Germany.

reform effects: disadvantaging males and pupils from low-educated backgrounds to the advantage of pupils from educated households in the lower tracks.

An important question regarding the applicability of these results is the extent to which they may transfer into different contexts. The first point to note is Germany's uniquely early tracking decision. If late grading is to have an effect on pupils' track choice then that track choice should happen reasonably close to the change in grading practice, especially early in the school system. Postponing grading for a year from Grades 2 to 3 in a school system that tracks after Grade 9 likely has a lower effect than the suggestive effects found here. Another important aspect to consider is the extent to which grades are only symbolic. The legislature and experience with the German grading system suggests that teachers, pupils, and parents take the grades distributed even in the early years quite seriously. In other contexts, grades in early Grades may be more symbolic and a tool to gradually introduce pupils to the notion of performance assessments. If grades in early Grades perform a more symbolic function, postponing the distribution of such grades may not do much to affect pupils' motivation and performance (not to speak of the low informational value of such types of grades).

# 8 Conclusion

This thesis estimates the effect of a German reform that postponed the school Grade at which pupils first receive formal number grades on their secondary school track choice. I consistently fail to reject a zero average effect on the three main secondary school tracks. This is shown to mask treatment heterogeneity across genders and parental educational background. Later graded males are found to more often pursue a higher track (*Realschule* or higher) as opposed to the basic track (*Hauptschule*). Specifically, later graded males are found to be, on average and all else equal, around 6 percentage points more likely than earlier graded males to pursue a track higher than the *Hauptschule*. I consistently fail to reject a zero effect for females. With respect to pupils household background, on the one hand, later graded pupils from low-educated households are found to be less likely to pursue a *Hauptschule* track than their earlier graded peers, but this effect does not translate into a statistically significant effect on their propensity to pursue a higher degree and was less robust to alternative sample choices. Later graded pupils from educated households, on the other hand, are shown to be more likely to pursue a basic *Hauptschule* degree compared to their earlier graded peers. The results for males are shown to be robust to a range of sensible robustness checks and qualitatively similar to an alternative weighted estimation. The results on the interaction by parental educational background exhibit more variation across alternative sample constructions and weighted estimations. Two potential mechanisms are explored. However, a rigorous analysis of their relative contribution to the effects observed is left to future research.

Against the backdrop of the continuing cross-state variation in grading timing in Germany, the gender and background gradient in track choice, and the long-lasting consequences of pupils' secondary school degrees, any extent to which later grading improves pupils' educational attainment at the margin should be relevant to policy makers. The evidence presented here suggests that this introduces a trade off. While there is some evidence that later grading benefits males

unilaterally, across parental educational backgrounds, the it suggests that later grading disadvantages pupils from educated households without an off setting improvement on the part of pupils from low-educated households. Thus, insofar as later grading reduces inequality in educational mobility, this appears to stem from making pupils from educated backgrounds worse off. Furthermore, it is not clear whether the marginal changes are due to a performance effect or changes in the quality of information available to parents. Policy makers wishing to implement later grading should thus be clear in their objective function. They should also be prepared to weather the reasonable objections that educated parents may have to such policies.

# References


Angrist, J. D. and Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Attanasio, O. P. and Kaufmann, K. M. (2014). Education choices and returns to schooling: Mothers' and youths' subjective expectations and their role by gender. *Journal of Development Economics*, 109:203–216.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust difference-in-difference estimates? *The Quarterly Journal of Economics*, (February):1–27.

Betts, J. R. (2011). The economics of tracking in education. In *Handbook of the Economics of Education*, chapter 7, pages 341–382. North-Holland, 3rd edition.

Betts, J. R. and Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review*, 22(4):343–352.

Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research*, 60(3):245–264.

Borghans, B. L., Diris, R., Smits, W., and de Vries, J. (2020). Should we sort it out later? The effect of tracking age on long-run outcomes. *Economics of Education Review*, 75:101973.

Brunello, G. and Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, 22(52):781–861.

Brunello, G. and Giannini, M. (2004). Stratified or comprehensive? The economic efficiency of school design. *Scottish Journal of Political Economy*, 51(2):173–193.

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

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

Cygan-Rehm, K. (2018). Is additional schooling worthless? Revising the zero returns to compulsory schooling in Germany. *CESIFO Working Papers*.

de Chaisemartin, C. and D'Haultfoeuille, X. (2019). Two-way fixed effects estimators with heterogeneous treatment effects. *National Bureau of Economic Research Working Paper Series*, 25904:1–35.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21.

Deutscher Bildungsrat (1970). Structural plan for the education system [Strukturplan für das Bildungswesen]. Technical report.

Deutscher Philologenverband (2016). DPhV welcomes results of you gov poll [DPhV begrüßt Ergebnisse der YouGov-Meinungsumfrage]. Press release.

Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review*, 109(8):2728–2765.

Duflo, E. (2001). Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American Economic Review*, 91(4):795–813.

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

Emanuelsson, I. and Fischbein, S. (1986). Vive la difference? A study on sex and schooling. *Scandinavian Journal of Educational Research*, 30(2):71–84.

Ermisch, J. and Francesconi, M. (2001). Family matters: Impacts of family background on educational attainments. *Economica*, 68(270):137–156.

Falch, T. and Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, 36:12–25.

Federal Statistical Office of Germany (Destatis) (1970). Population: states, cutoff date [Bevölkerung: Bundesländer, Stichtag] [Dataset]. Technical report, Statistisches Bundesamt, Wiesbaden.

Federal Statistical Office of Germany (Destatis) (2018). Schools at a glance [Schulen auf einen Blick]. Technical report.

Federal Statistical Office of Germany (Destatis) (2019). Students at general schools by educational area and school type [Schüler/-innen an allgemeinbildenden Schulen nach Bildungsbereichen und Schularten] [Dataset].

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2019). The German Socio-Economic Panel (SOEP). In *Jahrbücher für Nationalokonomie und Statistik*, volume 239.

Goodman-Bacon, A. (2019). Difference-in-differences with variation in treatment timing. *National Bureau of Economic Research Working Paper Series*, 25018.

Goodman-Bacon, A., Goldring, T., and Nichols, A. (2019). Bacondecomp: Stata module to perform a Bacon decomposition of difference-in-differences estimation. Statistical Software Components, Boston College Department of Economics.

Grundschulverband (2018). Fact check elementary school [Faktencheck Grundschule]. Report.

Grundschulverband (2019). Position performance culture [Standpunkt Leistungskultur]. Position paper.

Hallinan, M. T. (1994). Tracking: from theory to practice. *Sociology of Education*, 67(2):79–84.

Hanushek, E. A. and Woessmann, L. (2005). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *National Bureau of Economic Research Working Paper Series*, 11124.

Helbig, M. and Nikolai, R. (2015a). *Helbig / Nikolao collection of important school policy in the German states from 1949 to 2010 [Helbig / Nikolai Sammlung wichtiger schulrechtlicher Regelungen in den deutschen Bundesländern von 1949 bis 2010]*. Julius Klinkhardt.

Helbig, M. and Nikolai, R. (2015b). *Uncomparable: Changing school systems in German states since 1949 [Die Unvergleichbaren: Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949]*. Julius Klinkhardt.

Hesse (1980). Regulation concerning report cards in grades 1 and 2 in elementary school [Verordnung über Zeugnisse in der Klasse 1 und 2 der Grundschule]. Legislation.

Huebener, M. and Marcus, J. (2017). Compressing instruction time into fewer years of schooling and the impact on student performance. *Economics of Education Review*, 58:1–14.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *Quarterly Journal of Economics*, 125(2):515–548.

Jürges, H. and Schneider, K. (2007). What can go wrong will go wrong: Birthday effects and early tracking in the German school system. *CESifo Working Paper, No. 2055*.

Jürges, H., Schneider, K., Senkbeil, M., and Carstensen, C. H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31(1):56–65.

Kahn-Lang, A. and Lang, K. (2019). The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business and Economic Statistics*, 2019:1–26.

Klapp, A., Cliffordson, C., and Gustafsson, J. E. (2014). The effect of being graded on later achievement: evidence from 13-year olds in Swedish compulsory school. *Educational Psychology*, 36(10):1771–1789.

Konferenz der Ministerpräsidenten (1955). Düsseldorf accord [Düsseldorfer Abkommen]. Agreement.

Konferenz der Ministerpräsidenten (1964). Hamburg accord [Hamburger Abkommen]. Agreement.

Kultusministerkonferenz (1970). Recommendations for work in elementary school [Empfehlungen zur Arbeit in der Grundschule]. Report.

Kultusministerkonferenz (2008). Advancement through education - the educational qualification initiative for Germany [Aufstieg durch Bildung - Die Qualifizierungsinitiative für Deutschland]. Report.

Kultusministerkonferenz (2019a). Basic structure of the education system in the Federal Republic of Germany. Report.

Kultusministerkonferenz (2019b). *The Education System in the Federal Republic of Germany 2016/2017*. Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany.

Lower Saxony (1977). Report card regulation in elementary school [Zeugnisbestimmung für die Grundschule]. Legislation.

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

Marcus, J. and Zambre, V. (2019). The effect of increasing education efficiency on university enrollment: Evidence from administrative data and an unusual schooling reform in Germany. *Journal of Human Resources*, 54(2):468–502.

Meyer, T., Thomsen, S. L., and Schneider, H. (2019). New evidence on the effects of the shortened school duration in the German states: An evaluation of post-secondary education decisions. *German Economic Review*, 20(4):201–253.

North Rhine-Westphalia (1979). Regulation for the educational program in elementary school [Verordnung über den Bildungsgang in der Grundschule]. Legislation.

Piopiunik, M. (2014a). Intergenerational transmission of education and mediating channels: Evidence from a compulsory schooling reform in Germany. *Scandinavian Journal of Economics*, 116(3):878–907.

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

Piopiunik, M., Schwerdt, G., and Woessmann, L. (2013). Central school exit exams and labor-market outcomes. *European Journal of Political Economy*, 31:93–108.

Pischke, J. S. (2007). The impact of length of the school year on student performance and earnings: Evidence from the German short school years. *Economic Journal*, 117(523):1216–1242.

Pischke, J. S. and Von Wachter, T. (2008). Zero returns to compulsory schooling in Germany: Evidence and interpretation. *Review of Economics and Statistics*, 90(3):592–598.

Pischner, R. (2007). Data documentation 22 cross sectional weights and scaling factors in the socio-economic panel [Data Documentation 22 Die Querschnittsgewichtung und die Hochrechnungsfaktoren des Sozio-oekonomischen Panels (SOEP) ab release 2007 (Welle W)].

Poorthuis, A. M., Juvonen, J., Thomaes, S., Denissen, J. J., de Castro, B. O., and van Aken, M. A. (2015). Do grades shape students' school engagement? The psychological consequences of report card grades at the beginning of secondary school. *Journal of Educational Psychology*, 107(3):842–854.

Rangvid, B. S. (2015). Systematic differences across evaluation schemes and educational choice. *Economics of Education Review*, 48:41–55.

Reinhold, S. and Jürges, H. (2010). Secondary school fees and the causal effect of schooling on health behavior. *Health Economics*, 19(8):994–1001.

Rhineland-Palatinate (1988). School regulation for public elementary schools [Schulordnung für die öffentlichen Grundschulen]. Legislation.

Riphahn, R. T. (2012). Effect of secondary school fees on educational attainment. *Scandinavian Journal of Economics*, 114(1):148–176.

Roodman, D. (2015). Boottest: Stata module to provide fast execution of the wild bootstrap with null imposed. Statistical Software Components, Boston College Department of Economics.

Rubin D. B (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Schlotter, M. (2011). The effect of preschool attendance on secondary school track choice: Evidence from siblings. *Ifo Working Paper No. 106*.

Schneider, T. (2011). The relevance of social origin and migration background for teacher assessments by example of the elementary school recommendation [Die Bedeutung der sozialen Herkunft und des Migrationshintergrundes für Lehrerurteile am Beispiel der Grundschulempfehlung]. *Zeitschrift fur Erziehungswissenschaft*, 14(3):371–396.

Siahaan, F., Lee, D. Y., and Kalist, D. E. (2014). Educational attainment of children of immigrants: Evidence from the national longitudinal survey of youth. *Economics of Education Review*, 38:1–8.

Sjögren, A. (2010). Graded children: Evidence of longrun consequences of school grades from a nationwide reform. *IFAU Working Paper*.

SOEP (2019). Socio-Economic Panel (SOEP), data for years 1984-2017, version 34 [Dataset].

Solon, G., Haider, S. J., and Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources*, 50(2):301–316.

Urabe, M. (2009). *Function and history of the German report card [Funktion und Geschichte des deutschen Schulzeugnisses]*. Julius Klinkhardt.

van de Werfhorst, H. G. (2018). Early tracking and socioeconomic inequality in academic achievement: Studying reforms in nine countries. *Research in Social Stratification and Mobility*, 58:22–32.

Van Elk, R., van der Steeg, M., and Webbink, D. (2011). Does the timing of tracking affect higher education completion? *Economics of Education Review*, 30(5):1009–1021.

Wegener, B. (1985). Does social prestige exist? [Gibt es Sozialprestige?]. *Zeitschrift für Soziologie*, 14(3):209–235.

Woessmann, L. (2009). International evidence on school tracking: A review. *CESifo DICE Report*.

Wooldridge, J. M. (2012). *Introductory Econometrics. A Modern Approach*. South-Western, 5th edition.

You, S. and Sharkey, J. (2009). Testing a developmental-ecological model of student engagement: a multilevel latent growth curve analysis. *Educational Psychology*, 29(6):659–684.

# A Supplementary tables and full results

**Table A.1** Summary statistics: Policy implementation by state and school year

| State | School year |
|---|---|
| Bremen | 1971/72 |
| Hamburg | 1979/80 |
| **Hesse** | 1981/82 |
| **Lower-Saxony** | 1977/78 |
| **North Rhine-Westphalia** | 1979/80 |
| **Rhineland Palatinate** | 1988/89 |
| Schleswig-Holstein | 1990/91 |
| **Bavaria** | – |
| **Baden-Wuerttemberg** | – |
| Saarland | – |

*Notes*: States in bold are included in the sample. Bremen and Hamburg are small city states. Schleswig-Holstein and Saarland are very small states. Small and city states are not included in the sample to mitigate the effect of outliers.

**Table A.2** Identification: State of school enrollment and last recorded location

| | school enrolment, Fed. State | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | BW | BY | BE | HE | NI | NRW | RP | SH | Total |
| Location (last recorded) | | | | | | | | | |
| Baden-Wuerttemberg (BW) | **147** | 1 | 2 | 0 | 0 | 2 | 0 | 0 | 152 |
| Bavaria (BY) | 6 | **140** | 0 | 1 | 0 | 4 | 0 | 0 | 151 |
| Berlin (BE) | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| Brandenburg (BB) | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Bremen (HB) | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Hamburg (HH) | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 3 |
| Hesse(HE) | 1 | 1 | 0 | **53** | 0 | 3 | 0 | 0 | 58 |
| Lower-Saxony (NI) | 0 | 0 | 1 | 0 | **71** | 2 | 0 | 1 | 75 |
| North Rhine-Westphalia (NRW) | 1 | 1 | 0 | 2 | 5 | **197** | 0 | 0 | 206 |
| Rhineland-Palatinate (RP) | 1 | 3 | 0 | 2 | 0 | 0 | **51** | 0 | 57 |
| Saarland (SL) | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 8 |
| Schleswig-Holstein (SH) | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 3 |
| Total | 156 | 146 | 3 | 58 | 79 | 213 | 60 | 2 | 717 |

*Notes*: The table reports the correspondence of individuals' last recorded location to the state in which they are observed to have enrolled in school. This should give a rough measure of the accuracy of using individuals' *first* recorded location as a proxy for the state of school attendance. Information on the state of school enrollment is only available for 717 individuals in the sample. The first column indicates the last state in which individuals are observed in the sample. Columns 2 – 10 indicate the state in which individuals are observed to have enrolled in school. Individuals for whom the state of school enrollment corresponds to the last recorded location are indicated in **bold**. This is true for 93% of the individuals for whom this information is available.

**Table A.3** Identification: State of first exit from school and last recorded location

| | first exit from school, Fed. State | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | BW | BY | BE | HB | HH | HE | NI | NRW | RP | SH | Total |
| Location (last recorded) | | | | | | | | | | | |
| Baden-Wuerttemberg (BW) | **462** | 8 | 1 | 0 | 0 | 2 | 3 | 9 | 3 | 0 | 488 |
| Bavaria (BY) | 11 | **425** | 1 | 0 | 0 | 7 | 4 | 9 | 0 | 0 | 457 |
| Berlin (BE) | 1 | 0 | 1 | 0 | 0 | 2 | 1 | 2 | 2 | 0 | 9 |
| Brandenburg (BB) | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Bremen (HB) | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 3 |
| Hamburg (HH) | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 1 | 5 |
| Hesse (HE) | 5 | 3 | 0 | 0 | 0 | **213** | 4 | 8 | 3 | 0 | 236 |
| Lower Saxony (NI) | 1 | 1 | 1 | 1 | 1 | 2 | **252** | 4 | 0 | 4 | 267 |
| North Rhine-Westphalia (NRW) | 2 | 5 | 1 | 0 | 0 | 3 | 7 | **701** | 6 | 0 | 725 |
| Rhineland-Palatinate (RP) | 6 | 4 | 0 | 0 | 0 | 4 | 0 | 3 | **138** | 0 | 155 |
| Saarland (SL) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 14 |
| Saxony (SN) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Schleswig-Holstein (SH) | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 1 | 0 | 1 | 8 |
| Thuringia (TH) | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Total | 489 | 447 | 5 | 2 | 2 | 233 | 279 | 741 | 166 | 6 | 2370 |

*Notes*: The table reports the correspondence of individuals' last recorded location to the state in which they are observed to have first graduated from school. This should give a rough measure of the accuracy of using individuals' *first* recorded location as a proxy for the state of school attendance. Information on the state of school graduation is only available for 2370 individuals in the sample. The first column indicates the last state in which individuals are observed in the sample. Columns 2 – 12 indicate the state in which individuals are observed to have graduated from school. Individuals for whom the state of school graduation corresponds to the last recorded location are indicated in **bold**. This is true for 92% of the individuals for whom this information is available.

**Table A.4** Results: Heterogenous effects by gender (full results)

| | Gymnasium | | | Realschule | | | Hauptschule | | | > Hauptschule |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | General | Gender interacted | Fully iteracted | General | Gender interacted | Fully iteracted | General | Gender interacted | Fully iteracted | Fully iteracted |
| Treatment | -0.008 | -0.025 | -0.001 | 0.024 | 0.056 | 0.061 | -0.021 | -0.041 | -0.067 | 0.060 |
| | (0.017) | (0.014) | (0.016) | (0.017) | (0.018) | (0.024) | (0.010) | (0.005) | (0.011) | (0.015) |
| Female | 0.008 | -0.002 | -0.027 | 0.089 | 0.109 | 0.137 | -0.091 | -0.102 | -0.110 | 0.110 |
| | (0.008) | (0.007) | (0.025) | (0.011) | (0.010) | (0.029) | (0.010) | (0.006) | (0.018) | (0.021) |
| Female x Treatment | | 0.031 | -0.015 | | -0.061 | -0.066 | | 0.036 | 0.082 | -0.081 |
| | | (0.022) | (0.028) | | (0.010) | (0.030) | | (0.012) | (0.020) | (0.023) |
| ParentEduc | 0.456 | 0.456 | 0.446 | -0.186 | -0.185 | -0.127 | -0.256 | -0.256 | -0.305 | 0.319 |
| | (0.005) | (0.005) | (0.012) | (0.030) | (0.030) | (0.031) | (0.027) | (0.027) | (0.027) | (0.029) |
| ParentMig | -0.050 | -0.050 | -0.067 | -0.052 | -0.053 | -0.036 | 0.089 | 0.089 | 0.084 | -0.104 |
| | (0.015) | (0.016) | (0.023) | (0.019) | (0.019) | (0.019) | (0.016) | (0.016) | (0.026) | (0.033) |
| Female x ParentEduc | | | 0.019 | | | -0.108 | | | 0.090 | -0.089 |
| | | | (0.018) | | | (0.017) | | | (0.007) | (0.007) |
| Female x ParentMig | | | 0.033 | | | -0.030 | | | 0.009 | 0.003 |
| | | | (0.027) | | | (0.017) | | | (0.023) | (0.031) |
| State FE | x | x | x | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x | x | x | x |
| Full interactions | | | x | | | x | | | x | x |
| P-val 'treatment' | 0.703 | 0.204 | 0.961 | 0.284 | 0.084 | 0.088 | 0.093 | **0.013** | **0.012** | **0.012** |
| P-val 'interaction' | | 0.164 | 0.618 | | **0.018** | **0.011** | | **0.016** | **0.009** | **0.008** |
| P-val 'treatment + interaction' | | 0.849 | 0.704 | | 0.831 | 0.859 | | 0.859 | 0.470 | 0.591 |
| P-val 'F-test' | | 0.813 | 0.580 | | 0.807 | 0.820 | | 0.775 | 0.370 | 0.372 |
| CI 'treatment' | [-.042, .055] | [-.058, .034] | [-.047, .064] | [-.049, .067] | [-.022, .108] | [-.024, .143] | [-.049, .003] | [-.051, -.022] | [-.107, -.046] | [.03, .112] |
| CI 'treatment + interaction' | | [-.043, .078] | [-.069, .062] | | [-.079, .036] | [-.07, .044] | | [-.045, .029] | [-.025, .051] | [-.066, .031] |
| Control group mean | 0.279 | 0.279 | 0.279 | 0.410 | 0.410 | 0.410 | 0.294 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parantheses. Standard errors are clustered by state. 'General' models are models with only treatment variable and covariates. 'Gender interacted' models add an interaction term between treatment and female dummy. 'Fully interacted' models add further interaction terms between female dummy and all covariates, including state and cohort fixed effects. The results are equivalent to running separate regressions for males and females. In the 'interacted' models, the coefficient on the 'treatment' variable is the treatment effect on males. The treatment effect for females is the sum of the coefficient on the 'Treatment' variable and the interaction term 'Female x Treatment'. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'Female x Treatment', and 'Treatment + (Female x Treatment)'. "P-val 'F-test"' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. CI 'treatment' and CI 'treatment + interaction' report the bootstrapped 95% confidence intervals for the coefficient on the treatment variable and for the sum of the coefficients on the treatment variable and the interaction term.

**Table A.5** Results: Heterogenous effects by background (full results)

| | Gymnasium | | | Realschule | | | Hauptschule | | | > Hauptschule |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | General | Background interacted | Fully iteracted | General | Background interacted | Fully iteracted | General | Background interacted | Fully iteracted | Fully iteracted |
| Treatment | -0.008 | -0.009 | -0.005 | 0.024 | 0.040 | 0.036 | -0.021 | -0.037 | -0.036 | 0.031 |
| | (0.017) | (0.018) | (0.016) | (0.017) | (0.019) | (0.016) | (0.010) | (0.012) | (0.012) | (0.017) |
| ParentEduc | 0.456 | 0.454 | 0.378 | -0.186 | -0.153 | 0.007 | -0.256 | -0.286 | -0.367 | 0.385 |
| | (0.005) | (0.006) | (0.035) | (0.030) | (0.027) | (0.024) | (0.027) | (0.024) | (0.021) | (0.025) |
| Treatment x ParentEduc | | 0.006 | -0.024 | | -0.096 | -0.040 | | 0.089 | 0.063 | -0.064 |
| | | (0.014) | (0.030) | | (0.027) | (0.032) | | (0.023) | (0.015) | (0.021) |
| Female | 0.008 | 0.008 | 0.005 | 0.089 | 0.090 | 0.107 | -0.091 | -0.091 | -0.106 | 0.112 |
| | (0.008) | (0.008) | (0.007) | (0.011) | (0.011) | (0.012) | (0.010) | (0.010) | (0.010) | (0.009) |
| ParentMig | -0.050 | -0.050 | -0.048 | -0.052 | -0.053 | -0.065 | 0.089 | 0.090 | 0.099 | -0.113 |
| | (0.015) | (0.015) | (0.020) | (0.019) | (0.018) | (0.022) | (0.016) | (0.015) | (0.018) | (0.020) |
| ParentEduc x Female | | | 0.019 | | | -0.109 | | | 0.091 | -0.090 |
| | | | (0.017) | | | (0.014) | | | (0.005) | (0.005) |
| ParentEduc x ParentMig | | | -0.012 | | | 0.088 | | | -0.071 | 0.076 |
| | | | (0.036) | | | (0.022) | | | (0.024) | (0.030) |
| State FE | x | x | x | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x | x | x | x |
| Full interactions | | | x | | | x | | | x | x |
| P-val 'treatment' | 0.703 | 0.702 | 0.792 | 0.284 | 0.134 | 0.130 | 0.093 | **0.018** | **0.020** | 0.184 |
| P-val 'interaction' | | 0.731 | 0.561 | | **0.034** | 0.188 | | **0.025** | **0.015** | **0.014** |
| P-val 'treatment + interaction' | | 0.855 | 0.526 | | 0.050 | 0.915 | | **0.025** | **0.017** | **0.018** |
| P-val 'F-test' | | 0.856 | 0.453 | | 0.066 | 0.914 | | **0.015** | **0.009** | **0.004** |
| CI 'treatment' | [-.042, .055] | [-.044, .053] | [-.035, .051] | [-.049, .067] | [-.035, .091] | [-.033, .081] | [-.049, .003] | [-.069, -.008] | [-.068, -.006] | [-.011, .078] |
| CI 'treatment + interaction' | | [-.037, .101] | [-.113, .098] | | [-.149, 0] | [-.157, .078] | | [.011, .102] | [.013, .05] | [-.06, -.02] |
| Control group mean | 0.279 | 0.279 | 0.279 | 0.410 | 0.410 | 0.410 | 0.294 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parantheses. Standard errors are clustered by state. 'General' models are models with only treatment variable and covariates. 'Background interacted' models add an interaction term between treatment and the background dummy 'ParentEduc' which is 1 if either parent obtained an Abitur, 0 otherwise. 'Fully interacted' models add further interaction terms between the background dummy and all covariates, including state and cohort fixed effects. The results are equivalent to running separate regressions for educated and low-educated households. In the 'interacted' models, coefficient on 'Treatment' variable is the treatment effect on pupils from non-educated households (i.e. households where neither parent obtained an *Abitur*). The treatment effect for pupils from educated households is the sum of the coefficient on the 'Treatment' variable and the interaction term 'ParentEduc x Treatment'. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'ParentEduc x Treatment', and 'Treatment + (ParentEduc x Treatment)'. "P-val 'F-test"' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. CI 'treatment' and CI 'treatment + interaction' report the bootstrapped 95% confidence intervals for the coefficient on the treatment variable and for the sum of the coefficients on the treatment variable and the interaction term.

# B Robustness (estimates and elaboration)

## Weighted estimation

For each state, some observations have a zero-weight. This is mostly true for new partial samples, because in the first waves, respondents tend to exhibit worse response behavior (Pischner, 2007, p. 2). Zero-weight observations are attributed to observations from 5 of the 17 sample waves contained in the baseline sample. 59 observations are zero-weight for the *D 1994/95 migration sample* (out of 882 sample observations), 76 observations are zero-weight for the *G 2002 High-income sample* (out of 555 sample observations), 453 observations are zero weight for the *I 2009 Innovation sample* (complete sample), 222 observations are zero weight out of the *L2 2010 Family Type sample* (Low income, single parent, multi-child) sample (out of 1612 sample observations), and 5 observations are zero weight for the *L3 2011 Family Type sample* (Single parent, multi-child) sample. Observations with zero weight are excluded from both the weighted and unweighted estimation. This Appendix reports the full results of the weighted estimation.

## Discussion of the weighted estimation results

In the basic model, the sign of the coefficient on the treatment variable in the *Gymnasium* model changes sign but is similarly indistinguishable from zero. In the weighted estimation, the bootstrapped confidence interval excludes a decrease of the propensity to pursue a *Gymnasium* larger than 5 percentage points, compared to 4.3 percentage points in the unweighted estimation. Similarly, for the upper bound, the confidence interval excludes an increase larger than 6 percentage points in the weighted estimation compared to 4.9 percentage points in the unweighted estimation. The coefficient on the treatment variable in the *Realschule* model is 0.025 in the weighted specification compared with 0.024 in the unweighted specification. The bootstrapped p-values on these coefficients are 0.235 in the weighted specification compared with 0.284 in the unweighted specification. The confidence interval in the weighted estimation excludes a decrease larger than 0.4 percentage points and an increase larger than 7.5 percentage points, compared to the unweighted estimation, which excludes a decrease larger than 4.9 percentage points and an increase larger than 6.7 percentage points. Thus, while both results are indistinguishable from zero, the weighted estimation is skewed towards a positive effect. The coefficient on the treatment variable in the *Hauptschule* model is -0.052 in the weighted specification compared with -0.021 in the unweighted specification. The coefficient from the weighted model is statistically significantly different from zero at the 5% threshold. The coefficient from the original model was not statistically significantly different from zero. The coefficient on the treatment variable in the higher than *Hauptschule* model is 0.043 in the weighted estimation, compared to 0.015 in the unweighted estimation, and is significant at the 5% level.

Moving to the effect heterogeneity across genders, the results are again qualitatively similar across the weighted vs unweighted estimation. The coefficient on the treatment variable for males in the fully interacted *Gymnasium* model is 0.027 in the weighted estimation, compared with $-0.001$ in the unweighted estimation. The bootstrapped confidence interval for the weighted estimates excludes decreases larger than 10.2 percentage points and increases larger

than 10.8 percentage points and is as such much less precise than the weighted estimation, which excludes decreases larger than 4.7 percentage points and increases larger than 6.4 percentage points. The coefficient on the treatment variable in the fully interacted *Realschule* model is with 0.037 in the weighted estimation smaller than the unweighted estimate of 0.061. Both are statistically insignificantly different from zero and relatively imprecise. The bootstrapped confidence interval from the weighted estimation only excludes decreases larger than 3.3 percentage points and increases larger than 13.6 percentage points. The coefficient on the treatment variable in the *Hauptschule* model is -0.077 in the weighted estimation compared with -0.067 in the unweighted estimation. The weighted coefficient is significant at the 1% threshold. The unweighted coefficient was similarly significant at the 5% threshold. Thus, across the weighted and unweighted estimations, there is consistent evidence that later graded males are less likely to obtain a *Haupschule* degree, with a decreased propensity between 6.7 and 7.7 percentage points. The coefficient on the treatment variable in the generic higher than *Hauptschule* model is with 0.064 in the weighted estimation slightly larger than the unweighted estimate of 0.060. The weighted estimate is significant at the 1% threshold, compared to the unweighted estimate which is significant at the 5% threshold. Thus, across both estimation approaches there is consistent evidence that later graded male pupils are between 6 and 6.4 percentage points more likely to obtain a degree higher than a *Haupschule* degree.

Finally, moving on to effect heterogeneity across educational backgrounds, a similar picture emerges. In the *Gymnasium* model, the coefficient on the treatment effect in the weighted estimation is with 0.030 similarly indistinguishable from zero as the estimate under the unweighted estimation, -0.005. However, the bootstrapped confidence interval from the weighted estimation only excludes a decrease larger than 0.9 percentage points and an increase larger than 7 percentage points, thus indicating a slightly positively skewed estimate. The coefficient on the treatment effect in the *Realschule* model is at 0.030 in the weighted estimation very similar to the unweighted estimate of 0.036, relative to a control group mean of 41 percent. However, the weighted estimate is significant at the 5% threshold, whereas the unweighted estimate is not. The treatment coefficient in the *Hauptschule* model differs somewhat: the weighted estimation yields a considerably larger estimate of -0.069 compared to -0.036 in the unweighted estimation. Both estimates are significant at the 5% threshold. The results differ for pupils from educated households. Where the unweighted estimates indicated a significant increase in their propensity to pursue a *Hauptschule* degree, which was significant at the 5% level, the weighted estimates do not indicate such an effect. In fact, based on the bootstrapped confidence intervals, the estimation excludes a decrease larger than 4.8 percentage points and an increase larger than 4.9 percentage points, relative to a control group mean of 29.4 percent. The treatment-coefficient in the higher than *Hauptschule* model is with 0.060 in the weighted estimation also considerably larger than the unweighted estimate of 0.031. The weighted coefficient is significant at the 5% threshold, while the coefficient in the unweighted estimation is not significant at any conventional threshold of significance. For pupils from educated households, the results again differ. The unweighted estimation yielded a decreased propensity to obtain a degree higher than the *Hauptschule* degree significant at the 5% level. The unweighted estimation fails to detect this effect. In fact, the bootstrapped confidence intervals only excludes decreases larger than 5.5

percentage points and increases larger than 4.3 percentage points, compared to a control group mean of 68.9 percent.

Thus, the estimates are broadly consistent across the weighted and unweighted estimations, failing to reject a zero effect for the basic, as well as gender- and background-heterogeneity models for the dependent variable *Gymnasium*. The results for the *Realschule* model consistently fail to reject a zero average effect. The evidence regarding the effect across parental educational background is more mixed, and indicates a positive sign for pupils from low-educated households. As regards the dependent variable *Hauptschule* degree, the unweighted estimation fails to provide evidence of a non-zero average effect while the weighted estimation yields a negative effect significant at the 5% threshold. Later graded *males* are observed to have a consistently and statistically significant reduced propensity to obtain a *Hauptschule* degree, as do later graded pupils from low-educated households. For males, this translates into an increased propensity to obtain a *higher* degree, statistically significantly different from zero at the 5% threshold across weighted and unweighted estimations. For pupils from low-educated household, there is only suggestive evidence for an increased propensity to obtain a *higher* degree, as the effect is only distinguishable from zero at the 5% threshold in the weighted estimation. Later graded pupils from educated households are only less likely to obtain a higher degree than the *Hauptschule* degree, compared to their earlier graded peers, based on the unweighted estimation. In the weighted estimation, this effect disappears.

**Table B.1** Results: Basic difference-in-differences (weighted regression)

| | Gymnasium | | Realschule | | Hauptschule | | > Hauptschule |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| | Basic | General | Basic | General | Basic | General | General |
| Treatment | 0.014 | 0.018 | 0.028 | 0.025 | -0.051 | -0.052 | 0.043 |
| | (0.021) | (0.018) | (0.018) | (0.016) | (0.007) | (0.010) | (0.013) |
| Female | 0.013 | 0.011 | 0.076 | 0.076 | -0.082 | -0.081 | 0.088 |
| | (0.015) | (0.016) | (0.011) | (0.011) | (0.010) | (0.011) | (0.010) |
| ParentEduc | | 0.438 | | -0.159 | | -0.263 | 0.279 |
| | | (0.016) | | (0.042) | | (0.026) | (0.028) |
| ParentMig | | -0.054 | | -0.044 | | 0.081 | -0.098 |
| | | (0.017) | | (0.029) | | (0.021) | (0.016) |
| State FE | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x |
| P-val 'treatment' | 0.556 | 0.362 | 0.219 | 0.235 | **0.018** | **0.01**5 | **0.013** |
| CI 'treatment' | [-.05, .06] | [-.034, .078] | [-.006, .084] | [-.004, .075] | [-.089, -.041] | [-.089, -.032] | [.019, .082] |
| Control group mean | 0.279 | 0.279 | 0.410 | 0.410 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parentheses. Standard errors are clustered by state. 'Basic' model only contains treatment variable, gender dummy, and state and cohort fixed effects. 'General' model includes background covariates: 'ParentEduc' is a dummy variable that equals 1 if at least one parent obtained the Abitur, zero otherwise. 'ParentMig' is a dummy variable that equals 1 if one of the individuals' parents immigrated to Germany (but not the individual themselves), zero otherwise. "P-val 'treatment"' is the bootstrapped p-value for the coefficient on the treatment variable. CI 'treatment' reports the bootstrapped 95% confidence interval for the coefficient on the treatment variable.

**Table B.2** Results: Heterogenous effects by gender (full results, weighted regression)

| | Gymnasium | | | Realschule | | | Hauptschule | | | > Hauptschule |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | General | Gender interacted | Fully iteracted | General | Gender interacted | Fully iteracted | General | Gender interacted | Fully iteracted | Fully iteracted |
| Treatment | 0.018 | -0.002 | 0.027 | 0.025 | 0.050 | 0.037 | -0.052 | -0.062 | -0.077 | 0.064 |
| | (0.018) | (0.023) | (0.042) | (0.016) | (0.018) | (0.033) | (0.010) | (0.010) | (0.016) | (0.019) |
| Female | 0.011 | -0.002 | -0.014 | 0.076 | 0.093 | 0.121 | -0.081 | -0.088 | -0.100 | 0.106 |
| | (0.016) | (0.022) | (0.029) | (0.011) | (0.010) | (0.037) | (0.011) | (0.017) | (0.036) | (0.035) |
| Female x Treatment | | 0.043 | -0.018 | | -0.052 | -0.029 | | 0.020 | 0.058 | -0.047 |
| | | (0.035) | (0.063) | | (0.011) | (0.042) | | (0.022) | (0.028) | (0.037) |
| ParentEduc | 0.438 | 0.438 | 0.448 | -0.159 | -0.159 | -0.136 | -0.263 | -0.264 | -0.295 | 0.312 |
| | (0.016) | (0.016) | (0.018) | (0.042) | (0.042) | (0.030) | (0.026) | (0.026) | (0.021) | (0.023) |
| ParentMig | -0.054 | -0.053 | -0.065 | -0.044 | -0.045 | -0.037 | 0.081 | 0.082 | 0.076 | -0.103 |
| | (0.017) | (0.017) | (0.011) | (0.029) | (0.029) | (0.032) | (0.021) | (0.021) | (0.032) | (0.030) |
| Female x ParentEduc | | | -0.021 | | | -0.046 | | | 0.066 | -0.067 |
| | | | (0.035) | | | (0.036) | | | (0.018) | (0.015) |
| Female x ParentMig | | | 0.028 | | | -0.019 | | | 0.013 | 0.009 |
| | | | (0.029) | | | (0.024) | | | (0.030) | (0.038) |
| State FE | x | x | x | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x | x | x | x |
| Full interactions | | | x | | | x | | | x | x |
| P-val 'treatment' | 0.362 | 0.912 | 0.597 | 0.235 | **0.013** | 0.301 | **0.015** | **0.012** | **0.008** | **0.009** |
| P-val 'interaction' | | 0.206 | 0.819 | | 0.028 | 0.508 | | 0.615 | 0.268 | 0.411 |
| P-val 'treatment + interaction' | | 0.164 | 0.833 | | 0.920 | 0.704 | | 0.050 | 0.394 | 0.593 |
| P-val 'F-test' | | 0.192 | 0.789 | | 0.915 | 0.666 | | 0.083 | 0.371 | 0.558 |
| CI 'treatment' | [-.034, .078] | [-.076, .061] | [-.102, .108] | [-.004, .075] | [.016, .102] | [-.033, .136] | [-.089, -.032] | [-.095, -.035] | [-.136, -.049] | [.018, .122] |
| CI 'treatment + interaction' | | [-.016, .125] | [-.06, .094] | | [-.045, .047] | [-.049, .047] | | [-.119, 0] | [-.096, .028] | [-.037, .103] |
| Control group mean | 0.279 | 0.279 | 0.279 | 0.410 | 0.410 | 0.410 | 0.294 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parantheses. Standard errors are clustered by state. 'General' models are models with only treatment variable and covariates. 'Gender interacted' models add an interaction term between treatment and female dummy. 'Fully interacted' models add further interaction terms between female dummy and all covariates, including state and cohort fixed effects. The results are equivalent to two separate regression for the subgroups. In the 'interacted' models, coefficient on 'Treatment' variable is the treatment effect on males. The treatment effect for females is the sum of the coefficient on the 'Treatment' variable and the interaction term 'Female x Treatment'. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'Female x Treatment', and 'Treatment + (Female x Treatment)'. "P-val 'F-test"' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. CI 'treatment' and CI 'treatment + interaction' report the bootstrapped 95% confidence intervals for the coefficient on the treatment variable and for the sum of the coefficients on the treatment variable and the interaction term.

**Table B.3** Results: Heterogenous effects by background (full results, weighted regression)

| | Gymnasium | | | Realschule | | | Hauptschule | | | > Hauptschule |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | General | Background interacted | Fully iteracted | General | Background interacted | Fully iteracted | General | Background interacted | Fully iteracted | Fully iteracted |
| Treatment | 0.018 | 0.014 | 0.030 | 0.025 | 0.048 | 0.030 | -0.052 | -0.070 | -0.069 | 0.060 |
| | (0.018) | (0.019) | (0.013) | (0.016) | (0.019) | (0.015) | (0.010) | (0.010) | (0.010) | (0.013) |
| ParentEduc | 0.438 | 0.428 | 0.234 | -0.159 | -0.112 | 0.225 | -0.263 | -0.301 | -0.432 | 0.459 |
| | (0.016) | (0.023) | (0.042) | (0.042) | (0.041) | (0.044) | (0.026) | (0.019) | (0.026) | (0.036) |
| Treatment x ParentEduc | | 0.027 | -0.081 | | -0.138 | 0.008 | | 0.108 | 0.077 | -0.074 |
| | | (0.029) | (0.061) | | (0.041) | (0.068) | | (0.017) | (0.014) | (0.021) |
| Female | 0.011 | 0.011 | 0.014 | 0.076 | 0.076 | 0.084 | -0.081 | -0.081 | -0.092 | 0.098 |
| | (0.016) | (0.016) | (0.017) | (0.011) | (0.010) | (0.010) | (0.011) | (0.011) | (0.011) | (0.011) |
| ParentMig | -0.054 | -0.053 | -0.045 | -0.044 | -0.045 | -0.065 | 0.081 | 0.082 | 0.095 | -0.111 |
| | (0.017) | (0.016) | (0.017) | (0.029) | (0.028) | (0.033) | (0.021) | (0.021) | (0.025) | (0.019) |
| ParentEduc x Female | | | -0.010 | | | -0.063 | | | 0.073 | -0.073 |
| | | | (0.036) | | | (0.032) | | | (0.014) | (0.012) |
| ParentEduc x ParentMig | | | -0.041 | | | 0.122 | | | -0.086 | 0.081 |
| | | | (0.017) | | | (0.023) | | | (0.024) | (0.026) |
| State FE | x | x | x | x | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x | x | x | x | x |
| Full interactions | | | x | | | x | | | x | x |
| P-val 'treatment' | 0.362 | 0.523 | 0.113 | 0.235 | **0.016** | **0.027** | **0.015** | **0.015** | **0.016** | **0.012** |
| P-val 'interaction' | | 0.417 | 0.339 | | **0.042** | 0.885 | | **0.011** | **0.011** | **0.009** |
| P-val 'treatment + interaction' | | 0.130 | 0.567 | | 0.050 | 0.652 | | 0.262 | 0.632 | 0.494 |
| P-val | | 0.215 | 0.489 | | **0.035** | 0.593 | | 0.080 | 0.600 | 0.419 |
| CI 'treatment' | [-.034, .078] | [-.042, .067] | [-.009, .07] | [-.004, .075] | [.007, .102] | [.002, .073] | [-.089, -.032] | [-.104, -.051] | [-.103, -.05] | [.034, .093] |
| CI 'treatment + interaction' | | [-.023, .188] | [-.218, .226] | | [-.214, 0] | [-.181, .194] | | [-.021, .078] | [-.048, .049] | [-.055, .043] |
| Control group mean | 0.279 | 0.279 | 0.279 | 0.410 | 0.410 | 0.410 | 0.294 | 0.294 | 0.294 | 0.689 |
| *N* | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 | 12898 |

*Notes:* Standard errors in parantheses. Standard errors are clustered by state. 'General' models are models with only treatment variable and covariates. 'Background interacted' models add an interaction term between treatment and the background dummy 'ParentEduc' which is 1 if either parent obtained an Abitur, 0 otherwise. models add further interaction terms between female dummy and all covariates, including state and cohort fixed effects. The results are equivalent to two separate regression for the subgroups. In the 'interacted' models, coefficient on 'Treatment' variable is the treatment effect on pupils from non-educated households (i.e. households where neither parent obtained and abitur). The treatment effect for pupils from educated households is the sum of the coefficient on the 'Treatment' variable and the interaction term 'ParentEduc x Treatment'. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'ParentEduc x Treatment', and 'Treatment + (ParentEduc x Treatment)'. "P-val 'F-test"' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. CI 'treatment' and CI 'treatment + interaction' report the bootstrapped 95% confidence intervals for the coefficient on the treatment variable and for the sum of the coefficients on the treatment variable and the interaction term.

### Missing information on parents' education

Section 3 discusses the construction of the background-control variable *ParentEduc*. It is noted that information on parental education is missing for some observations. This missing information can be categorized into three categories: completely missing information, non-problematically missing information, and potentially problematically missing information. Completely missing is information on 518 observations on parental education. When information on reasons for missing information is provided, three reasons are distinguished. If a person refuses to provide the answer or does not know the answer, it is coded as "no answer / don't know". This is true for 440 observations on paternal education and for 404 observations on maternal education. A second reason is that information may be missing when a question is not asked because it is not relevant for a specific person. This is coded as "does not apply". This is true for 212 observations on paternal education and for 191 observations on maternal education. Lastly with the extension of the SOEP in recent years, entirely new samples have been added to the core. In these samples, sometimes questions are left out completely, e.g. to shorten the questionnaire or because the focus of the sample is different as in some of the related studies. In such a case, the variable are coded as "Not included in this version of the questionnaire" for an entire subsample. This is true for 959 observations on paternal education and 939 observations on maternal education. Missing information on parental education may introduce bias in the estimators if information is missing non-randomly (Wooldridge, 2012, p. 324). Missing information on parental education is likely not a problem if the question was not part of the respective survey to begin with. This is true for paternal education for 959 observations. For maternal education, this is true for 939 observations.

This Appendix reports treatment and interaction term coefficients across a range of robustness checks, using alternative data imputations and sample constructions.

### Discussion of the robustness estimates

The timing of the policy implementations across the four German states (see Table 2) shows that Lower-Saxony, North Rhine-Westphalia, and Hesse all implemented postponed grading within a four-year time-span (between 1977/78 and 1981/82). Rhineland-Palatinate is an outlier in that it only postponed grading in 1988/89 – seven years after Hesse. The large time gap between these policy changes may skew the results if those years capture other state-specific or time-variant variation. I thus drop Rhineland-Palatinate from the sample and furthermore drop all observations that enter school later than 1986. This reduces the number of available observations to 9,624 The new treatment and relevant group-interaction coefficients are reported in Column 2 (*R1*) in tables B.4, B.5, and B.6. Dropping Rhineland-Palatinate increases the point estimate on the treatment variable in the *Gymnasium* model from -0.008 to -0.014. This new coefficient is still statistically insignificant. The coefficients in the other models (*Realschule*, *Hauptschule*, and > *Hauptschule*) remain mostly unchanged (Column 2 in table B.4). The same is true for the coefficients in the gender-interacted model (results in table B.5). The sample changes do not change the results in the background-interacted model (see table B.6).

Some measurement error may have been introduced by the decision how to handle missing

information on month of birth to impute the entry cohort. In the baseline specification, observations with missing month of birth were coded as having entered school the year they turn 7 years old (in effect assuming a greater propensity to postpone enrollment in line with the tendency exhibited in official statistics). The alternative would have been to assume an entry point in the year the invididuals turn 6. The sensitivity of the results to this modelling choice is reported in Column 3 (*R2*). The results using this alternative specification are almost identical to the results from the main analysis.

Parental education plays a large role in the empirical analysis presented in this thesis. While information on parental education is available for a large majority of observations, information is incomplete for some 1,700 observations. In the baseline estimation, observations with non-usable missing information were dropped from the estimation. This may introduce bias into the estimators, if information is missing non-randomly. Since the control variable of choice is whether or not either of an individuals' parents obtained an *Abitur*, specifically accurate data on parental education may not be needed. Rather, missing information on parental education may indeed be a good proxy for neither parent having obtained an *Abitur*, as individuals propensity to know their parents' educational background is likely increasing in their parents schooling. Thus, an alternative would have been to code missing information on parental education as "non-Abitur". The sensitivity of the results to this alternative is reported in Column 4 (*R3*). The results are very similar to the baseline results both in the basic and in the gender-dissected models. Some minor differences surface in the background-heterogeneity model, as would be expected. Specifically, the coefficient on the treatment dummy for the dependent variable *Hauptschule* decreases from -0.036 to -0.028. The bootstrapped p-value for new alternative estimate indicates insignificance with a value of 0.083.

Another dimension that may have introduced measurement error into the estimation is the choice of first affected cohort, as this is subject to some measurement error. In the baseline estimation, the first treated cohort was assumed to be the cohort entering school *the year before* the policy took effect, i.e. the cohort entering Grade 2 in the year of the policy change. It is a reasonable assumption, as most grading regimes prior to the policy change did not grade in Grade 1 but even before the reform only started assigning number grades with Grade 2. Any measurement error this modeling decision may have introduced should be relatively small, given the long sample size. To investigate the sensitivity of the results to this measurement error, I drop the first affected cohort from the sample. The results are reported in Column 5 (*R4*) of tables B.4, B.5, and B.6. This adjustment leaves most estimates virtually unchanged. As in case *R3*, the greatest difference is in the point-estimate of the treatment dummy in the background-heterogeneity model. Here, with the dependent variable *Hauptschule* degree, the point estimate decreases in absolute terms from -0.036 to -0.032. The bootstrapped p-value for the new estimate is 0.096. Thus, the estimate is not significantly different from zero at any threshold of significance.

Another concern is general measurement error around the imputed entry cohorts and first treated cohorts. I drop the marginal cohorts (i.e. cohorts just treated and just not-treated) from the sample. The sensitivity of the results to this alternative are reported in Column 6 (*R5*). With this variation, the results remain mostly unchanged, both in terms of magnitude

and significance. Only the coefficient on the treatment effect on pupils from low-educated households decreases on their propensity to obtain a *Hauptschule* degree decreases from -0.036 to -0.034, which is no longer significant at the 5% level.

A final concern may be that other reforms are confounding the treatment estimate. Other reforms around the transition from elementary to secondary school are discussed in section 2. Most of these reforms take place before the intervention window considered here, or mostly apply to states excluded from the sample. The most likely confounding reform is the relaxation of the degree to which the schools' recommendation for a child's secondary school track was binding, which Lower Saxony implemented in 1978, only one year after the postponed grading reform. In order to assess the extent to which any effects stemming from this reform confound the estimates of the postponed grading reforms, I drop Lower Saxony from the sample and re-estimate my primary and heterogeneity models. The results are reported in Column 7 (*R6*) of tables B.4, B.5, and B.6. For the basic model, the results are virtually unchanged. Importantly, the alternative sample indicates no changes on the extensive margin of statistical significance. In the gender-heterogeneity model the alternative specification changes none of the results, relative to the baseline estimation.

In sum, the estimates of the primary specifications presented in this thesis prove robust to the alternative specifications outlined above. I robustly fail to reject a zero effect on pupils propensity to pursue either a *Gymnasium* or a *Realschule* track. The analysis presented herein does offer suggestive evidence of a treatment effect on males and pupils from educated backgrounds concentrated at the lower end of the tracks. The evidence of an effect on pupils from low-educated backgrounds is less robust to the alternatives, with 3 of the 6 alternative specifications changing the significance level of the treatment estimate in the background *Hauptschule* model.

**Table B.4** Robustness: Main results

| | (1) OLS | (2) R1 | (3) R2 | (4) R3 | (5) R4 | (6) R5 | (7) R6 |
|---|---|---|---|---|---|---|---|
| ***Panel A: Gymnasium*** | | | | | | | |
| Treatment | -0.008 | -0.014 | -0.007 | -0.013 | -0.009 | -0.012 | -0.010 |
| | (0.017) | (0.015) | (0.017) | (0.016) | (0.018) | (0.014) | (0.018) |
| P-val 'treatment' | 0.703 | 0.618 | 0.721 | 0.576 | 0.727 | 0.516 | 0.687 |
| ***Panel B: Realschule*** | | | | | | | |
| Treatment | 0.024 | 0.025 | 0.023 | 0.025 | 0.022 | 0.022 | 0.015 |
| | (0.017) | (0.013) | (0.016) | (0.020) | (0.018) | (0.018) | (0.017) |
| P-val 'treatment' | 0.284 | 0.187 | 0.288 | 0.347 | 0.375 | 0.352 | 0.528 |
| ***Panel C: Hauptschule*** | | | | | | | |
| Treatment | -0.021 | -0.019 | -0.021 | -0.016 | -0.017 | -0.016 | -0.016 |
| | (0.010) | (0.011) | (0.009) | (0.011) | (0.010) | (0.012) | (0.007) |
| P-val 'treatment' | 0.093 | 0.191 | 0.075 | 0.338 | 0.185 | 0.265 | 0.066 |
| ***Panel D: > Hauptschule*** | | | | | | | |
| Treatment | 0.015 | 0.011 | 0.016 | 0.012 | 0.012 | 0.010 | 0.005 |
| | (0.014) | (0.019) | (0.012) | (0.014) | (0.014) | (0.016) | (0.009) |
| P-val 'treatment' | 0.393 | 0.617 | 0.237 | 0.498 | 0.448 | 0.606 | 0.672 |
| $N$ | 12898 | 9624 | 12898 | 14559 | 12525 | 11759 | 11219 |

*Notes:* Standard errors in parantheses. Model as in section 5. Only reporting coefficients on treatment dummy. Robustness models: R1: without Rhineland-Palatinate, 1971–1986. R2: alternative cohort imputation for missing month-of-birth. R3: alternative parental education imputation. R4: first affected cohort excluded. R5: marginal cohorts (relative cohorts $\in\{-1,0,1\}$) excluded. R6: Lower Saxony excluded.

**Table B.5** Robustness: Gender

| | (1) OLS | (2) R1 | (3) R2 | (4) R3 | (5) R4 | (6) R5 | (7) R6 |
|---|---|---|---|---|---|---|---|
| ***Panel A: Gymnasium*** | | | | | | | |
| Treatment | -0.001 | -0.007 | 0.001 | -0.004 | 0.002 | 0.000 | 0.012 |
| | (0.016) | (0.018) | (0.016) | (0.016) | (0.012) | (0.015) | (0.015) |
| Female x Treatment | -0.015 | -0.016 | -0.016 | -0.019 | -0.021 | -0.023 | -0.043 |
| | (0.028) | (0.034) | (0.029) | (0.025) | (0.027) | (0.029) | (0.010) |
| P-val 'treatment' | 0.961 | 0.699 | 0.976 | 0.855 | 0.906 | 0.986 | 0.503 |
| P-val 'treatment + interaction' | 0.704 | 0.601 | 0.697 | 0.570 | 0.678 | 0.552 | 0.469 |
| ***Panel B: Realschule*** | | | | | | | |
| Treatment | 0.061 | 0.072 | 0.059 | 0.054 | 0.059 | 0.056 | 0.040 |
| | (0.024) | (0.025) | (0.024) | (0.026) | (0.023) | (0.024) | (0.016) |
| Female x Treatment | -0.066 | -0.083 | -0.064 | -0.050 | -0.066 | -0.059 | -0.042 |
| | (0.030) | (0.033) | (0.034) | (0.033) | (0.030) | (0.029) | (0.019) |
| P-val 'treatment' | 0.088 | 0.064 | 0.094 | 0.138 | 0.098 | 0.102 | 0.195 |
| P-val 'treatment + interaction' | 0.859 | 0.544 | 0.884 | 0.891 | 0.811 | 0.928 | 0.936 |
| ***Panel C: Hauptschule*** | | | | | | | |
| Treatment | -0.067 | -0.076 | -0.065 | -0.056 | -0.064 | -0.057 | -0.061 |
| | (0.011) | (0.010) | (0.011) | (0.012) | (0.012) | (0.012) | (0.007) |
| Female x Treatment | 0.082 | 0.103 | 0.079 | 0.072 | 0.085 | 0.073 | 0.082 |
| | (0.020) | (0.013) | (0.024) | (0.019) | (0.022) | (0.024) | (0.021) |
| P-val 'treatment' | **0.012** | **0.018** | **0.012** | **0.011** | **0.012** | **0.023** | **0.025** |
| P-val 'treatment + interaction' | 0.470 | 0.178 | 0.523 | 0.464 | 0.337 | 0.573 | 0.373 |
| ***Panel D: > Hauptschule*** | | | | | | | |
| Treatment | 0.060 | 0.065 | 0.060 | 0.050 | 0.061 | 0.056 | 0.052 |
| | (0.015) | (0.016) | (0.014) | (0.015) | (0.017) | (0.017) | (0.011) |
| Female x Treatment | -0.081 | -0.099 | -0.079 | -0.069 | -0.087 | -0.082 | -0.085 |
| | (0.023) | (0.018) | (0.026) | (0.019) | (0.024) | (0.026) | (0.023) |
| P-val 'treatment' | **0.012** | **0.022** | **0.012** | **0.022** | **0.029** | **0.021** | **0.027** |
| P-val 'treatment + interaction' | 0.591 | 0.377 | 0.502 | 0.472 | 0.423 | 0.572 | 0.224 |
| *N* | 12898 | 9624 | 12898 | 14559 | 12525 | 11759 | 11219 |

*Notes:* Standard errors in parantheses. Model as in section 5. Only reporting coefficients on treatment dummy and interaction term. Robustness models: R1: without Rhineland-Palatinate, 1971–1986. R2: alternative cohort imputation for missing month-of-birth. R3: alternative parental education imputation. R4: first affected cohort excluded. R5: marginal cohorts (relative cohorts ∈{-1,0,1}) excluded. R6: Lower Saxony excluded.

**Table B.6** Robustness: Background

| | (1) OLS | (2) R1 | (3) R2 | (4) R3 | (5) R4 | (6) R5 | (7) R6 |
|---|---|---|---|---|---|---|---|
| ***Panel A: Gymnasium*** | | | | | | | |
| Treatment | -0.005 | -0.013 | -0.004 | -0.011 | -0.003 | -0.009 | -0.007 |
| | (0.015) | (0.017) | (0.015) | (0.014) | (0.016) | (0.017) | (0.017) |
| ParentEduc x Treatment | -0.024 | -0.013 | -0.029 | -0.018 | -0.044 | -0.025 | -0.021 |
| | (0.038) | (0.045) | (0.038) | (0.038) | (0.040) | (0.043) | (0.041) |
| P-val 'treatment' | 0.792 | 0.623 | 0.843 | 0.601 | 0.879 | 0.567 | 0.763 |
| P-val 'treatment + interaction' | 0.526 | 0.221 | 0.550 | 0.525 | 0.432 | 0.693 | 0.591 |
| ***Panel B: Realschule*** | | | | | | | |
| Treatment | 0.036 | 0.035 | 0.035 | 0.035 | 0.030 | 0.037 | 0.026 |
| | (0.017) | (0.020) | (0.017) | (0.016) | (0.018) | (0.020) | (0.019) |
| ParentEduc x Treatment | -0.040 | -0.038 | -0.042 | -0.040 | -0.018 | -0.054 | -0.030 |
| | (0.044) | (0.052) | (0.044) | (0.044) | (0.046) | (0.050) | (0.048) |
| P-val 'treatment' | 0.130 | 0.112 | 0.120 | 0.201 | 0.246 | 0.166 | 0.306 |
| P-val 'treatment + interaction' | 0.915 | 0.768 | 0.872 | 0.914 | 0.825 | 0.790 | 0.941 |
| ***Panel C: Hauptschule*** | | | | | | | |
| Treatment | -0.036 | -0.030 | -0.036 | -0.028 | -0.032 | -0.034 | -0.030 |
| | (0.015) | (0.017) | (0.015) | (0.014) | (0.016) | (0.017) | (0.017) |
| ParentEduc x Treatment | 0.063 | 0.052 | 0.068 | 0.055 | 0.061 | 0.073 | 0.056 |
| | (0.039) | (0.046) | (0.039) | (0.039) | (0.040) | (0.044) | (0.042) |
| P-val 'treatment' | **0.020** | **0.044** | **0.028** | 0.083 | 0.096 | 0.130 | **0.042** |
| P-val 'treatment + interaction' | **0.017** | **0.024** | **0.017** | **0.017** | **0.023** | **0.023** | **0.033** |
| ***Panel D: > Hauptschule*** | | | | | | | |
| Treatment | 0.031 | 0.022 | 0.032 | 0.024 | 0.027 | 0.029 | 0.019 |
| | (0.015) | (0.018) | (0.015) | (0.014) | (0.016) | (0.017) | (0.017) |
| ParentEduc x Treatment | -0.064 | -0.051 | -0.070 | -0.058 | -0.063 | -0.079 | -0.051 |
| | (0.040) | (0.047) | (0.040) | (0.039) | (0.041) | (0.044) | (0.043) |
| P-val 'treatment' | 0.184 | 0.483 | 0.088 | 0.339 | 0.204 | 0.222 | 0.246 |
| P-val 'treatment + interaction' | **0.018** | **0.019** | **0.017** | **0.018** | **0.020** | **0.020** | **0.033** |
| *N* | 12898 | 9624 | 12898 | 14559 | 12525 | 11759 | 11219 |

*Notes:* Standard errors in parantheses. Model as in section 5. Only reporting coefficients on treatment dummy and interaction term. Robustness models: R1: without Rhineland-Palatinate, 1971–1986. R2: alternative cohort imputation for missing month-of-birth. R3: alternative parental education imputation. R4: first affected cohort excluded. R5: marginal cohorts (relative cohorts ∈{-1,0,1}) excluded. R6: Lower Saxony excluded.

# C Event study (estimates and graphs)

## Unweighted coefficients

**Table C.1** Event study estimates: *Gymnasium*

| $t$ | Coefficient | SD | CI Lower | CI Upper |
|---|---|---|---|---|
| -6 | -0.006 | 0.033 | -0.105 | 0.105 |
| -5 | 0.009 | 0.030 | -0.119 | 0.107 |
| -4 | 0.027 | 0.018 | -0.074 | 0.097 |
| -3 | 0.018 | 0.009 | -0.019 | 0.085 |
| -2 | 0.012 | 0.022 | -0.059 | 0.087 |
| -1 | 0.000 | . | . | . |
| 0 | 0.017 | 0.027 | -0.074 | 0.121 |
| 1 | -0.005 | 0.040 | -0.321 | 0.366 |
| 2 | 0.034 | 0.037 | -0.052 | 0.184 |
| 3 | 0.033 | 0.027 | -0.063 | 0.123 |
| 4 | 0.031 | 0.026 | -0.042 | 0.107 |
| 5 | -0.017 | 0.011 | -0.034 | 0.017 |
| 6 | -0.013 | 0.033 | -0.127 | 0.121 |
| 7 | -0.062 | 0.011 | -0.097 | -0.011 |
| 8 | -0.019 | 0.030 | -0.146 | 0.110 |
| 9 | -0.010 | 0.029 | -0.114 | 0.105 |
| 10 | 0.034 | 0.043 | -0.081 | 0.220 |
| 11 | -0.046 | 0.014 | -0.140 | 0.018 |
| 12 | -0.011 | 0.015 | -0.056 | 0.062 |
| $N$ | 12898 | | | |

*Notes*: The table reports the coefficients on the leading and lagging treatment dummies from an estimation of equation 4.3 for the dependent variable *Gymnasium* degree. $t$ reports the event times relative to first treatment. Relative event times before $-6$ are binned into the leading variable $t = -6$. Relative event times after 12 are binned into the lagging variable $t = 12$. $t = -1$ is omitted from the estimation. Standard-deviations and confidence intervals are boostrapped using the Wild t bootstrapp developed by Roodman (2015).

**Table C.2** Event study estimates: *Realschule*

| $t$ | Coefficient | SD | CI Lower | CI Upper |
|---|---|---|---|---|
| -6 | 0.006 | 0.040 | -0.137 | 0.090 |
| -5 | 0.009 | 0.038 | -0.171 | 0.082 |
| -4 | -0.063 | 0.035 | -0.253 | 0.009 |
| -3 | -0.001 | 0.013 | -0.029 | 0.044 |
| -2 | 0.003 | 0.030 | -0.095 | 0.078 |
| -1 | 0.000 | . | . | . |
| 0 | 0.020 | 0.017 | -0.059 | 0.056 |
| 1 | 0.020 | 0.044 | -0.328 | 0.228 |
| 2 | -0.017 | 0.037 | -0.188 | 0.083 |
| 3 | -0.020 | 0.027 | -0.123 | 0.046 |
| 4 | -0.029 | 0.035 | -0.173 | 0.083 |
| 5 | 0.043 | 0.025 | -0.039 | 0.088 |
| 6 | 0.003 | 0.041 | -0.199 | 0.158 |
| 7 | 0.040 | 0.045 | -0.241 | 0.151 |
| 8 | 0.038 | 0.040 | -0.204 | 0.147 |
| 9 | 0.023 | 0.033 | -0.134 | 0.123 |
| 10 | 0.102 | 0.053 | -0.272 | 0.291 |
| 11 | 0.108 | 0.032 | -0.074 | 0.156 |
| 12 | 0.042 | 0.027 | -0.125 | 0.117 |
| $N$ | 12898 | | | |

*Notes*: The table reports the coefficients on the leading and lagging treatment dummies from an estimation of equation 4.3 for the dependent variable *Realschule* degree. $t$ reports the event times relative to first treatment. Relative event times before $-6$ are binned into the leading variable $t = -6$. Relative event times after 12 are binned into the lagging variable $t = 12$. $t = -1$ is omitted from the estimation. Standard-deviations and confidence intervals are boostrapped using the Wild t bootstrapp developed by Roodman (2015).

**Table C.3** Event study estimates: *Hauptschule*

| $t$ | Coefficient | SD | CI Lower | CI Upper |
|---|---|---|---|---|
| -6 | 0.007 | 0.031 | -0.096 | 0.098 |
| -5 | -0.005 | 0.033 | -0.082 | 0.200 |
| -4 | 0.034 | 0.032 | -0.042 | 0.190 |
| -3 | -0.027 | 0.010 | -0.051 | 0.000 |
| -2 | -0.021 | 0.038 | -0.100 | 0.072 |
| -1 | 0.000 | . | . | . |
| 0 | -0.049 | 0.016 | -0.097 | -0.017 |
| 1 | -0.011 | 0.015 | -0.046 | 0.086 |
| 2 | -0.013 | 0.029 | -0.104 | 0.102 |
| 3 | -0.004 | 0.025 | -0.080 | 0.106 |
| 4 | -0.014 | 0.022 | -0.061 | 0.045 |
| 5 | -0.026 | 0.014 | -0.074 | 0.015 |
| 6 | -0.012 | 0.023 | -0.099 | 0.080 |
| 7 | 0.033 | 0.035 | -0.060 | 0.218 |
| 8 | -0.023 | 0.034 | -0.121 | 0.153 |
| 9 | -0.036 | 0.039 | -0.137 | 0.140 |
| 10 | -0.119 | 0.030 | -0.232 | 0.045 |
| 11 | -0.058 | 0.026 | -0.098 | 0.092 |
| 12 | -0.030 | 0.017 | -0.090 | 0.080 |
| $N$ | 12898 | | | |

*Notes*: The table reports the coefficients on the leading and lagging treatment dummies from an estimation of equation 4.3 for the dependent variable *Hauptschule* degree. $t$ reports the event times relative to first treatment. Relative event times before $-6$ are binned into the leading variable $t = -6$. Relative event times after 12 are binned into the lagging variable $t = 12$. $t = -1$ is omitted from the estimation. Standard-deviations and confidence intervals are boostrapped using the Wild t bootstrapp developed by Roodman (2015).

### Weighted coefficients

**Table C.4** Event study estimates: *Gymnasium* (weighted)

| $t$ | Coefficient | SD | CI Lower | CI Upper |
|---|---|---|---|---|
| -6 | -0.052 | 0.057 | -0.239 | 0.077 |
| -5 | -0.023 | 0.052 | -0.267 | 0.220 |
| -4 | 0.025 | 0.072 | -0.329 | 0.219 |
| -3 | -0.077 | 0.026 | -0.272 | 0.180 |
| -2 | -0.016 | 0.039 | -0.199 | 0.059 |
| -1 | 0.000 | . | . | . |
| 0 | -0.018 | 0.051 | -0.161 | 0.163 |
| 1 | -0.006 | 0.050 | -0.371 | 0.419 |
| 2 | 0.065 | 0.058 | -0.136 | 0.297 |
| 3 | -0.029 | 0.043 | -0.250 | 0.078 |
| 4 | 0.005 | 0.022 | -0.105 | 0.049 |
| 5 | -0.018 | 0.044 | -0.208 | 0.084 |
| 6 | 0.110 | 0.040 | -0.059 | 0.229 |
| 7 | -0.127 | 0.045 | -0.369 | -0.015 |
| 8 | -0.058 | 0.043 | -0.222 | 0.060 |
| 9 | -0.097 | 0.054 | -0.390 | 0.085 |
| 10 | 0.031 | 0.069 | -0.323 | 0.285 |
| 11 | -0.074 | 0.080 | -0.424 | 0.211 |
| 12 | -0.015 | 0.029 | -0.099 | 0.101 |
| $N$ | 12898 | | | |

*Notes*: The table reports the coefficients on the leading and lagging treatment dummies from a weighted estimation of equation 4.3 for the dependent variable *Gymnasium* degree. $t$ reports the event times relative to first treatment. Relative event times before $-6$ are binned into the leading variable $t = -6$. Relative event times after 12 are binned into the lagging variable $t = 12$. $t = -1$ is omitted from the estimation. Standard-deviations and confidence intervals are boostrapped using the Wild t bootstrapp developed by Roodman (2015).

**Table C.5** Event study estimates: *Realschule* (weighted)

| $t$ | Coefficient | SD | CI Lower | CI Upper |
|---|---|---|---|---|
| -6 | 0.036 | 0.024 | -0.039 | 0.126 |
| -5 | 0.028 | 0.043 | -0.190 | 0.158 |
| -4 | -0.052 | 0.077 | -0.384 | 0.290 |
| -3 | 0.072 | 0.029 | -0.207 | 0.243 |
| -2 | 0.045 | 0.054 | -0.057 | 0.260 |
| -1 | 0.000 | . | . | . |
| 0 | 0.037 | 0.036 | -0.106 | 0.113 |
| 1 | 0.041 | 0.060 | -0.406 | 0.399 |
| 2 | -0.002 | 0.038 | -0.096 | 0.079 |
| 3 | 0.056 | 0.049 | -0.069 | 0.175 |
| 4 | 0.029 | 0.020 | 0.000 | 0.101 |
| 5 | 0.068 | 0.037 | -0.015 | 0.187 |
| 6 | -0.031 | 0.029 | -0.080 | 0.081 |
| 7 | 0.047 | 0.056 | -0.177 | 0.216 |
| 8 | 0.095 | 0.040 | -0.012 | 0.367 |
| 9 | 0.090 | 0.012 | 0.074 | 0.125 |
| 10 | 0.142 | 0.071 | -0.012 | 0.519 |
| 11 | 0.108 | 0.076 | -0.199 | 0.377 |
| 12 | 0.081 | 0.031 | -0.019 | 0.156 |
| $N$ | 12898 | | | |

*Notes*: The table reports the coefficients on the leading and lagging treatment dummies from a weighted estimation of equation 4.3 for the dependent variable *Realschule* degree. $t$ reports the event times relative to first treatment. Relative event times before $-6$ are binned into the leading variable $t = -6$. Relative event times after 12 are binned into the lagging variable $t = 12$. $t = -1$ is omitted from the estimation. Standard-deviations and confidence intervals are boostrapped using the Wild t bootstrapp developed by Roodman (2015).

**Table C.6** Event study estimates: *Hauptschule* (weighted)

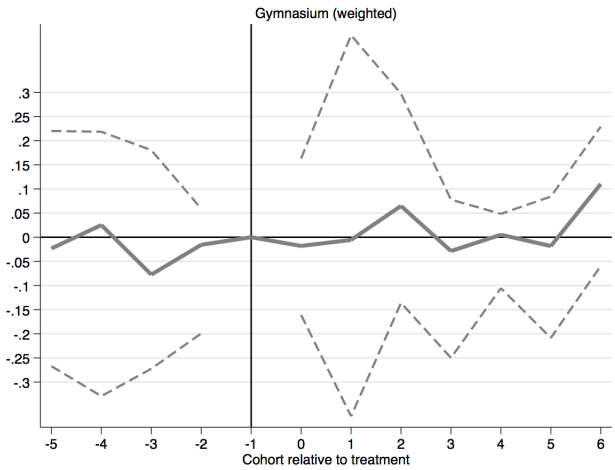
| $t$ | Coefficient | SD | CI Lower | CI Upper |
|---|---|---|---|---|
| -6 | 0.017 | 0.031 | -0.050 | 0.124 |
| -5 | 0.008 | 0.013 | -0.031 | 0.041 |
| -4 | 0.030 | 0.027 | -0.068 | 0.171 |
| -3 | -0.016 | 0.026 | -0.209 | 0.119 |
| -2 | -0.026 | 0.042 | -0.143 | 0.087 |
| -1 | 0.000 | . | . | . |
| 0 | -0.037 | 0.027 | -0.150 | 0.027 |
| 1 | -0.040 | 0.017 | -0.102 | 0.021 |
| 2 | -0.064 | 0.031 | -0.188 | 0.049 |
| 3 | -0.016 | 0.024 | -0.100 | 0.038 |
| 4 | -0.056 | 0.025 | -0.168 | 0.053 |
| 5 | -0.057 | 0.028 | -0.112 | 0.062 |
| 6 | -0.113 | 0.015 | -0.183 | -0.076 |
| 7 | 0.080 | 0.073 | -0.222 | 0.381 |
| 8 | -0.022 | 0.026 | -0.103 | 0.031 |
| 9 | -0.038 | 0.036 | -0.139 | 0.126 |
| 10 | -0.156 | 0.020 | -0.215 | -0.109 |
| 11 | -0.042 | 0.027 | -0.121 | 0.104 |
| 12 | -0.065 | 0.023 | -0.118 | 0.024 |
| $N$ | 12898 | | | |

*Notes*: The table reports the coefficients on the leading and lagging treatment dummies from a weighted estimation of equation 4.3 for the dependent variable *Hauptschule* degree. $t$ reports the event times relative to first treatment. Relative event times before $-6$ are binned into the leading variable $t = -6$. Relative event times after 12 are binned into the lagging variable $t = 12$. $t = -1$ is omitted from the estimation. Standard-deviations and confidence intervals are boostrapped using the Wild t bootstrapp developed by Roodman (2015).

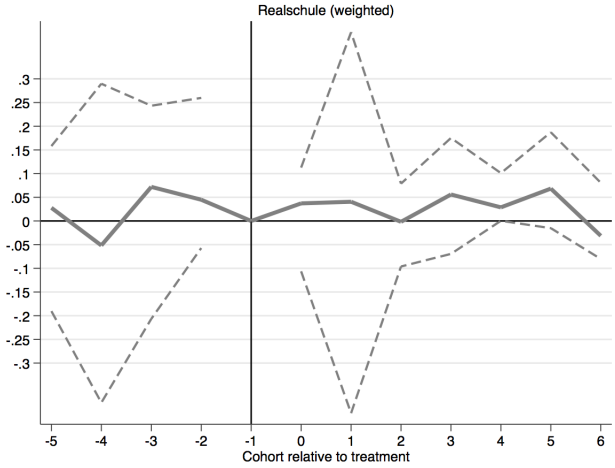
## Event study graphs – Weighted estimation

**Figure C.1** Event study graphs by degree (weighted)

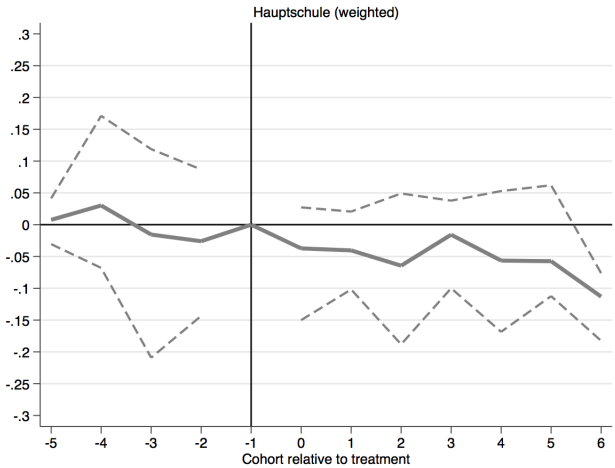
**(a)** Gymnasium (general university qualification)




**(b)** Realschule (intermediate degree)




**(c)** Hauptschule (basic degree)

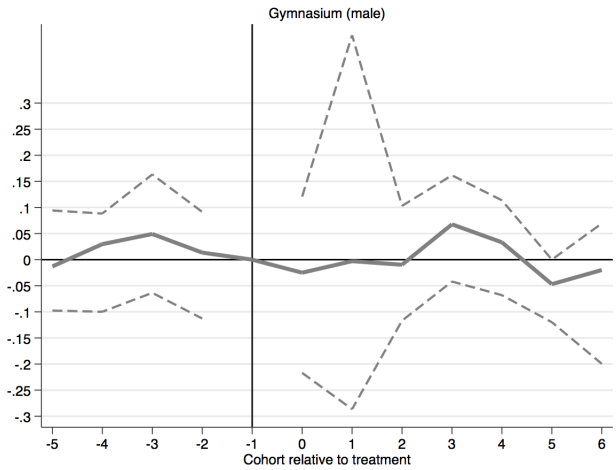



*Note:* Event studies are plotted with bootstrapped 95% confidence intervals.

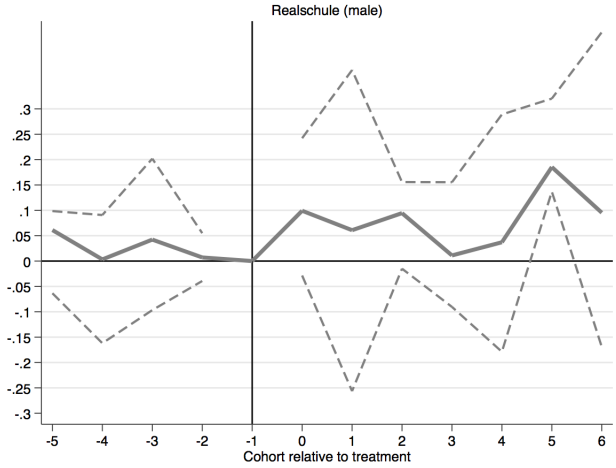
## Event study graphs – Male sample

**Figure C.2** Event study graphs by degree (male sample)

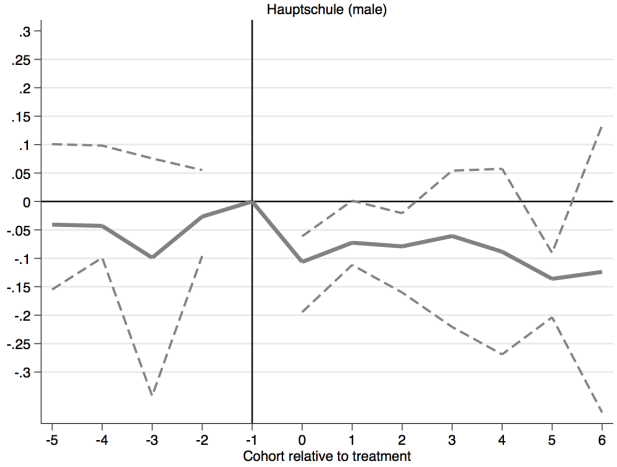
**(a)** Gymnasium (general university qualification)




**(b)** Realschule (intermediate degree)




**(c)** Hauptschule (basic degree)

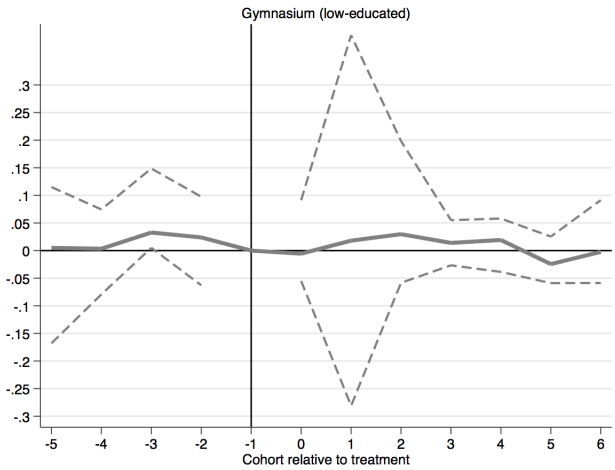



*Note:* Event studies are plotted with bootstrapped 95% confidence intervals.

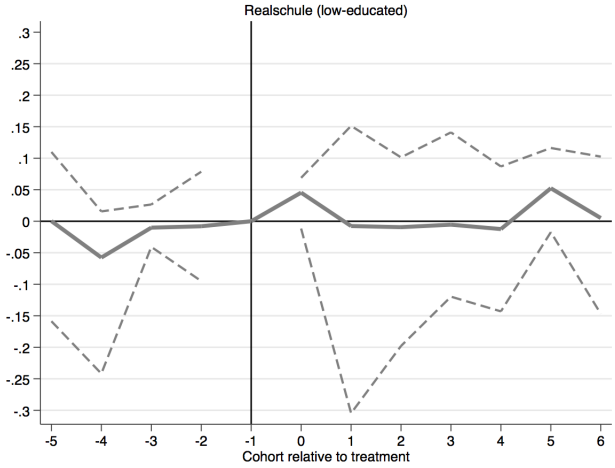
## Event study graphs – Low-educated sample

**Figure C.3** Event study graphs by degree (low-educated sample)

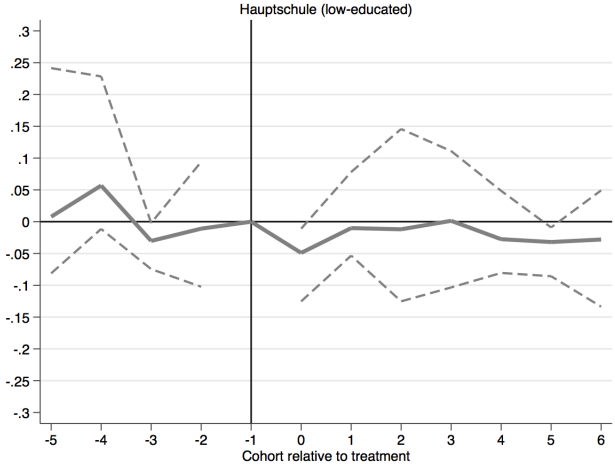
**(a)** Gymnasium (general university qualification)




**(b)** Realschule (intermediate degree)




**(c)** Hauptschule (basic degree)




*Note:* Event studies are plotted with bootstrapped 95% confidence intervals.

# D Decomposition

**Figure D.1** Bacon Decomposition of the DD estimate: *Gymnasium*

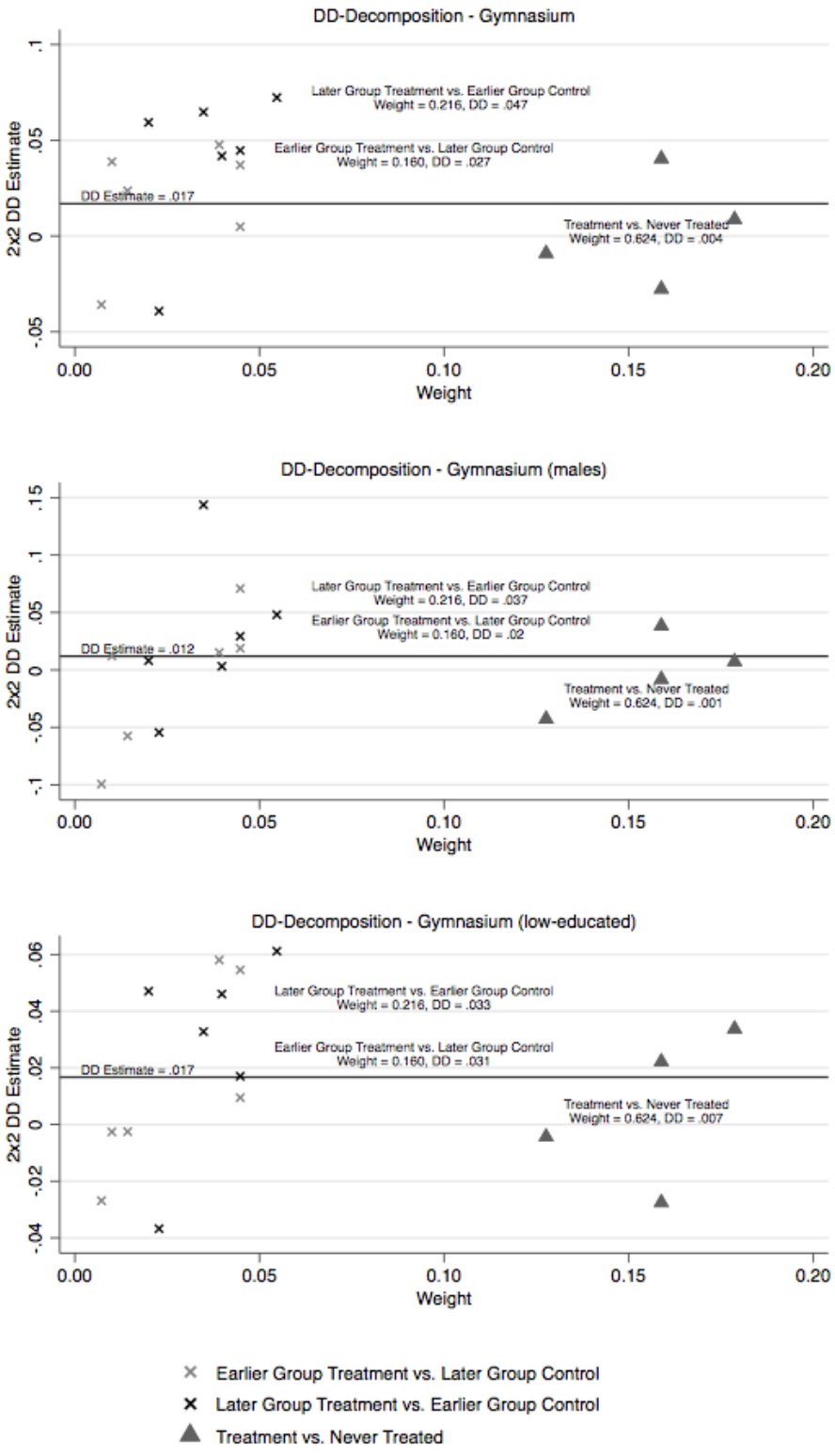


**Figure D.2** Bacon Decomposition of the DD estimate: *Realschule*

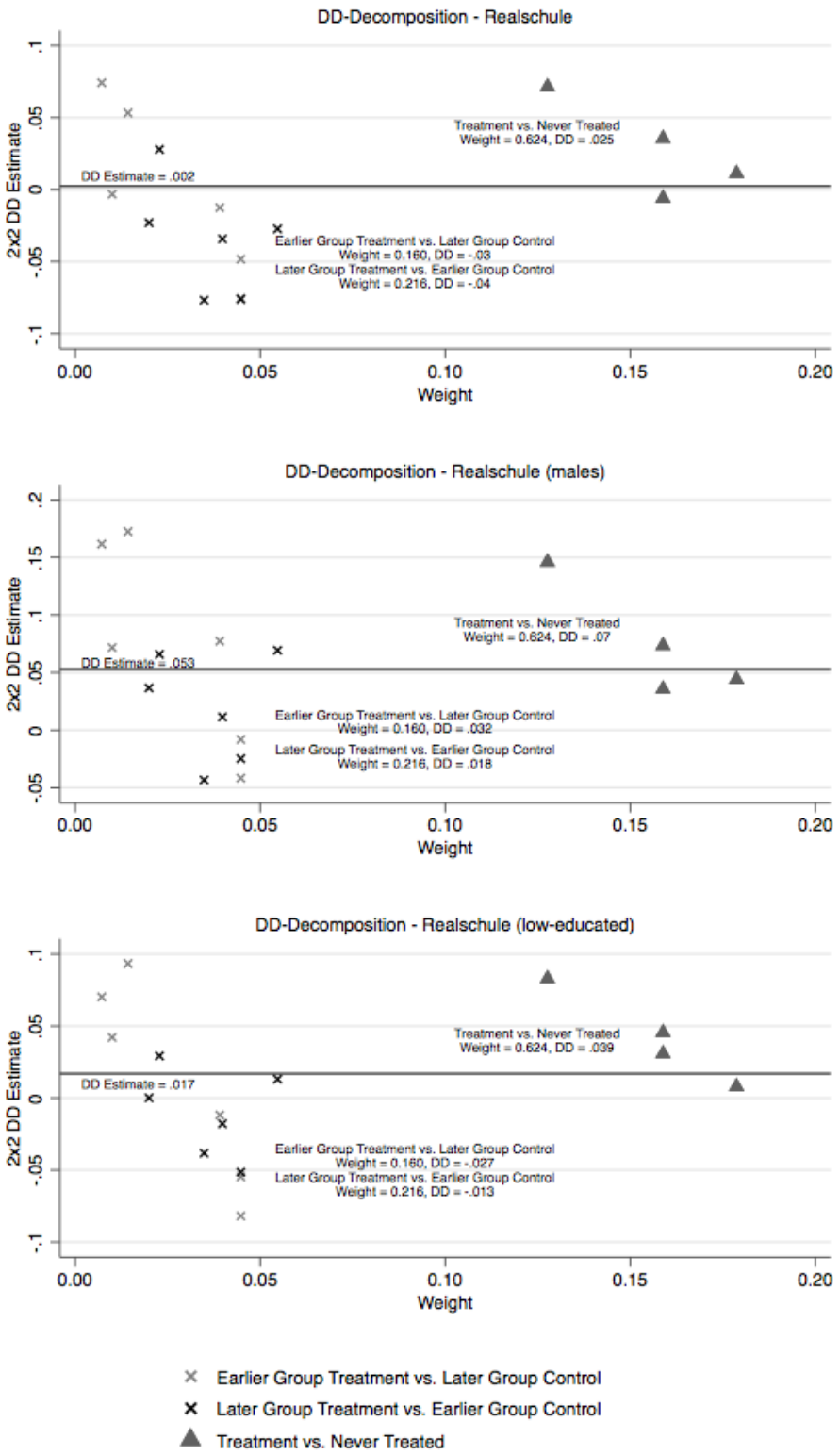
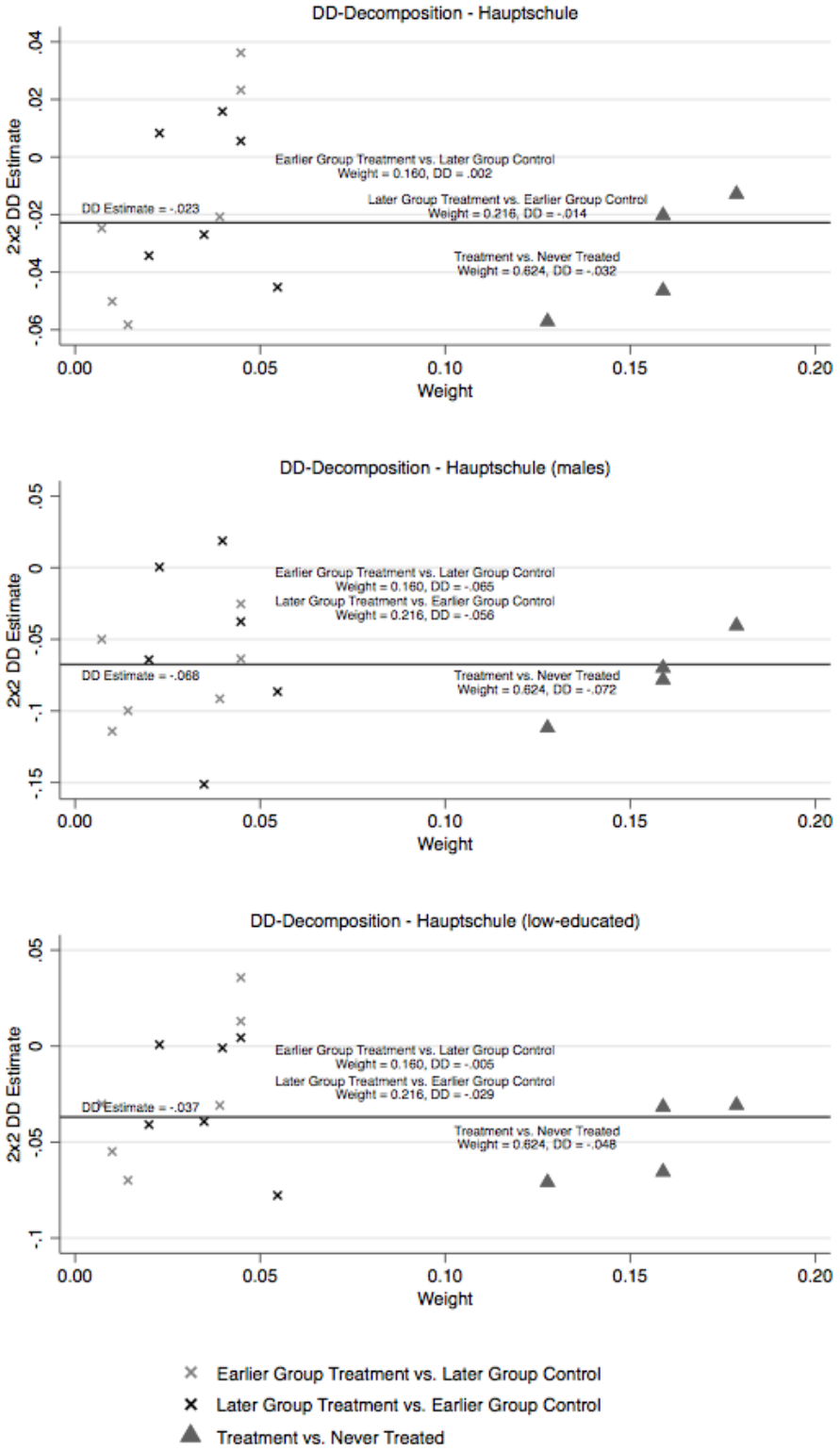


**Figure D.3** Bacon Decomposition of the DD estimate: *Hauptschule*



# E Additional estimations

## Mechanism

Section 7 briefly introduced two proposed mechanisms through which late grading may affect pupils' track choice. In this appendix I perform some approximate analyses to search for evidence of the two aspects of the performance mechanism: motivation and ability building. The results presented here offer only a first step in a further analyses of the exact mechanism through which late grading affects pupils' track choice. Using the data available from the SOEP I find no evidence for either of the proposed mechanisms.

The mechanisms proposed in section 7 trace the observed change in the outcome variable "track recommendation" to two ultimate causes: the effect due to exposure to grading, and the effect due to alternative forms of feedback. I propose that these two ultimate causes change either internally pupils' motivation or change externally pupils' ability building. Since motivation and ability building are not observable characteristics, I propose to approximate their relative importance by analysing the effect of treatment on likely correlated characteristics. Specifically, I propose to estimate the treatment effect on a set of attitudes on personality traits surveyed as part of the SOEP.

It is important to note that the survey data available poses a limitation with respect to the extent to which these mechanisms can be tested. Unlike the track-choice and degree behavior, which is relatively closely linked in time to the early years of elementary education, the measures of internal motivation and external ability building proposed here are collected long after the intervention. Therefore, the extent to which any variation in outcomes can be reasonably traced back to the policy itself is questionable. In light of these limitations, the evidence for the channels of transmission discussed in this section has to be considered suggestive.

The SOEP periodically surveys participants on a range of attitudes, values, and personality traits. In order to gauge the relative importance of either of the proposed mechanisms, I need to find characteristics and attitudes that are plausibly correlated to these ultimate causes.

With respect to the external ability mechanism I propose that the degree to which individuals perceive of themselves as the following are plausibly correlated with external sources of ability:

- The degree to which an individual perceives of themselves as working diligently
- The degree to which an individual perceives of themselves as working efficiently
- The degree to which an individual assesses their ability to handle stress well
- The degree to which an individual voices a propensity to doubt their ability when encountering an obstacle
- The degree to which an individual perceives of themselves as patient

With respect to the internal motivation mechanism I propose that the degree to which individuals perceive of themselves as the following should be plausibly correlated with their internal motivation:

- The degree to which they perceive their life as determined by their actions
- The degree to which they perceive their life as determined by themselves
- The degree to which they voice willingness to reciprocate favors

- The degree to which they voice willingness to go out of their way to reciprocate favors
- The degree to which they report having a vivid phantasy
- The degree to which they report as being interested and willing to learn

Table E.3 offers a brief description of the above variables, their scales, and the number of steps on the scale.

Tables E.1 and E.2 report the coefficients estimated by the following model:

$$y_{isc} = \beta_1 \cdot D_{sc} + \gamma_s + \lambda_c + \beta_2 \cdot female_i + \delta \mathbf{X}_{isc} + \varepsilon_{isc} \tag{E.1}$$

where $y_{isc}$ denotes the dependent variable, $D_{sc}$ denotes the treatment and equals 1 if state $s$ graded students from cohort $c$ from 3rd grade, $\gamma_s$ denotes a set of state fixed effects, $\lambda_c$ denotes a set of cohort fixed effects, $female_i$ is a female dummy that equals 1 if an individual is female and 0 otherwise, and $\mathbf{X}$ is a set of further control variables such as households' educational background, parental migration background, age, and interaction terms between the gender dummy and the background controls (except for age).

The proposed identification is virtually identical to the identification strategy of the main analysis. Using this identifying approach, I find no evidence for either the internal "motivation" mechanism nor of the external "ability" mechanism. The coefficient on the treatment variable estimates the effect of treatment on males. The boostrapped p-values reported at the bottom of tables E.1 and E.2 indicate insignificance for all estimated treatment coefficients. The sum of the treatment variable and the interaction term indicates the estimated treatment effect for females. The bootstrapped p-values indicate insignificance for all dependent variables except for the ability to handle stress (model (3), table E.1).

**Table E.1** Mechanism: Ability

| | (1) work_diligent | (2) work_efficient | (3) handle_stress | (4) prop_selfdoubt1 | (5) prop_selfdoubt2 | (6) patience |
|---|---|---|---|---|---|---|
| Treatment | -0.087 | -0.024 | 0.017 | -0.037 | 0.123 | 0.081 |
| | (0.053) | (0.032) | (0.046) | (0.056) | (0.077) | (0.182) |
| Female | -0.034 | -0.050 | -0.401 | 0.291 | 0.588 | -0.393 |
| | (0.051) | (0.052) | (0.052) | (0.057) | (0.056) | (0.073) |
| Female x Treatment | 0.081 | -0.034 | 0.013 | -0.059 | 0.041 | -0.158 |
| | (0.045) | (0.042) | (0.044) | (0.076) | (0.064) | (0.322) |
| State FE | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x |
| Age | x | x | x | x | x | x |
| Full interactions | x | x | x | x | x | x |
| P-val 'treatment' | 0.274 | 0.484 | 0.808 | 0.574 | 0.328 | 0.698 |
| P-val 'interaction' | 0.223 | 0.805 | 0.777 | 0.434 | 0.521 | 0.582 |
| P-val 'treatment + interaction' | 0.905 | 0.352 | **0.021** | 0.548 | 0.124 | 0.728 |
| P-val 'F-test' | 0.894 | 0.223 | **0.030** | 0.343 | 0.097 | 0.704 |
| *N* | 9572 | 9569 | 9568 | 2303 | 7555 | 4455 |

*Notes*: Standard errors in parentheses. Standard errors are clustered by state. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'Female x Treatment', and 'Treatment + (Female x Treatment)'. P-val 'F-test' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. Age is included as a control variable because responses to the surveyed questions may vary with age and responses are collected over various years. Thus, any age effect is not captured by the cohort FE.

**Table E.2** Mechanism: Engagement

| | (1) life_action_determined | (2) life_self_determined | (3) recipr | (4) recipr_effort | (5) vivid_phantasy | (6) interested |
|---|---|---|---|---|---|---|
| Treatment | 0.101 | -0.040 | -0.060 | -0.102 | 0.025 | -0.054 |
| | (0.047) | (0.028) | (0.049) | (0.071) | (0.044) | (0.024) |
| Female | 0.133 | -0.000 | 0.046 | -0.189 | 0.028 | -0.050 |
| | (0.053) | (0.087) | (0.056) | (0.029) | (0.052) | (0.038) |
| Female x Treatment | -0.006 | 0.057 | 0.031 | 0.059 | 0.032 | -0.076 |
| | (0.047) | (0.099) | (0.067) | (0.106) | (0.064) | (0.056) |
| State FE | x | x | x | x | x | x |
| Cohort FE | x | x | x | x | x | x |
| Age | x | x | x | x | x | x |
| Full interactions | x | x | x | x | x | x |
| P-val 'treatment' | 0.137 | 0.217 | 0.397 | 0.329 | 0.642 | 0.087 |
| P-val 'interaction' | 0.902 | 0.677 | 0.663 | 0.620 | 0.649 | 0.306 |
| P-val 'treatment + interaction' | 0.359 | 0.797 | 0.507 | 0.454 | 0.429 | 0.074 |
| P-val 'F-test' | 0.276 | 0.861 | 0.447 | 0.381 | 0.354 | 0.070 |
| *N* | 2059 | 2060 | 7562 | 7562 | 9566 | 8500 |

*Notes*: Standard errors in parentheses. Standard errors are clustered by state. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'Female x Treatment', and 'Treatment + (Female x Treatment)'. P-val 'F-test' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero. Age is included as a control variable because responses to the surveyed questions may vary with age and responses are collected over various years. Thus, any age effect is not captured by the cohort FE.

**Table E.3** Mechanism: Variables

| Variable | Description | Scale | Steps |
|---|---|---|---|
| *handle_stress* | can handle stress well | does not apply – applies | 7 |
| *prop_selfdoubt1* | when in difficulty, tend to doubt myself | disagree completely – agree completely | 4 |
| *prop_selfdoubt2* | when in difficulty, tend to doubt myself | does not apply – applies | 7 |
| *patience* | personal patience | very impatient – very patient | 10 |
| *recipr* | willingness to reciprocate favors | does not apply – applies | 7 |
| *recipr_effort* | particular effort to reciprocate | does not apply – applies | 7 |
| *work_diligent* | work diligently | does not apply – applies | 7 |
| *work_efficient* | complete tasks effectively, efficiently | does not apply – aplies | 7 |
| *life_action_determined* | life determined by my actions | agree completely – disagree | 4 |
| *life_self_determined* | life determined by me | agree completely – disagree | 4 |
| *vivid_phantasy* | have vivid phantasy | does not apply – applies | 7 |
| *interested* | interested, want to learn | does not apply – applies | 7 |

*Notes:* The table presents the dependent variables for the estimation of the 'ability' and 'engagement' mechanisms in Tables E.1 and E.2. Column 1 reports the name of the variable. Column 2 reports a abbreviated version of the question that respondents were asked. Column 3 reports the edges of the scale that applies to the question. Column 4 reports the steps on that scale, including the edges. To illustrate: row one reports the information for a variable *handle_stress* (Column 1), which asked respondents about their ability to handle stress (Column 2). Answers applied on a scale from 'does not apply' to 'applies' (Column 3). Respondents could answer on a scale from 1 'does not apply' to 7 'applies' (Column 4).

### Intergenerational mobility

A recurring topic in German educational debates is intergenerational educational mobility, i.e. children's ability to achieve greater educational attainment than their parents. Not least is it one of the proclaimed goals of current educational policy (Kultusministerkonferenz, 2008). Thus, table E.4 analyzes the propensity for pupils to obtain a higher degree than their parents. I drop all observations for whom information on both parents' education is imcomplete, missing, or unusable. I furthermore drop all observations for which either parent obtained an *Abitur*, as the dependent variable is necessarily coded as 0 for these observations. This leaves 10,147 observations for the baseline estimation (column (1), table E.4). The point estimate on the treatment dummy is 0.017, which is insignificant at any conventional threshold of significance. The point-estimate of the treatment effect does not differ across model specifications (columns (2) and (3) in table E.4, p-values= 0.106 and 0.227, respectively). The failure to reject a zero effect is furthermore persistent also for the subgroup of pupils from *Hauptschule* backgrounds (column (4), p-value= 0.256).

**Table E.4** Intergenerational mobility

| | (1) General | (2) Gender interacted | (3) Fully interacted | (4) Low-educated households |
|---|---|---|---|---|
| Treatment | 0.017 | 0.017 | 0.016 | 0.040 |
| | (0.017) | (0.010) | (0.012) | (0.024) |
| Female | 0.076 | 0.076 | 0.078 | 0.122 |
| | (0.009) | (0.012) | (0.014) | (0.012) |
| Female x Treatment | | 0.000 | 0.001 | -0.039 |
| | | (0.027) | (0.024) | (0.016) |
| ParentMig | 0.133 | 0.133 | 0.147 | -0.009 |
| | (0.012) | (0.012) | (0.023) | (0.020) |
| Female x ParentMig | | | -0.027 | -0.031 |
| | | | (0.024) | (0.024) |
| State FE | x | x | x | x |
| Cohort FE | x | x | x | x |
| Trend | | | x | x |
| P-val 'treatment' | 0.435 | 0.106 | 0.227 | 0.256 |
| P-val 'interaction' | | 0.995 | 0.951 | **0.030** |
| P-val 'treatment + interaction' | | 0.603 | 0.576 | 0.980 |
| P-val 'F-test' | | 0.569 | 0.531 | 0.977 |
| *N* | 10147 | 10147 | 10147 | 6601 |

*Notes*: Standard errors in parentheses. Standard errors are clustered by state. P-values 'treatment', 'interaction', and 'treatment + interaction' at the bottom of the table are bootstrapped p-values for the coefficients on 'Treatment', 'Female x Treatment', and 'Treatment + (Female x Treatment)'. P-val 'F-test' is the p-value on a test that the sum of the coefficient on the treatment variable and the interaction term is zero.