Class Composition and Educational Outcomes Evidence from the Abolition of Denominational Schools

Ilka Gerhardts, University of Munich *

Uwe Sunde, University of Munich

Larissa Zierow, ifo Institute and University of Munich

March 2021

Abstract

Denominational schools are an important provider of education in many countries around the world. Due to their focus, these schools often operate with multigrade classes, in which more than one age cohort is taught in one classroom. Multigrade classes are a cost-effective way to provide education and play a crucial role in education policy in the context of demographic change. This paper presents estimates of the causal effect of attending denominational schools with multigrade classes on schooling and short-run labor market outcomes. The analysis combines administrative records of schools with comprehensive population census data, and exploits the abolition of denominational schools in the Saarland, a German state, in 1969, for identification of the effect. The findings document significantly detrimental effects on final grade attainment and labor market participation. Notably, the negative impact is most pronounced in the outcomes of girls.

JEL Classification: C21, I26, J13, J16, J21, J61, J62

Keywords: Education, multigrade schools, gender effects

*Corresponding author. Email: Ilka.Gerhardts@econ.lmu.de

1 Introduction

Many schools are operated on a basis of multigrade classes. Multigrade teaching represents a cost-effective way of providing children with education in the context of limited resources. In fact, in large parts of the world schools with multigrade classes, often run by different religious denominations, represent the typical way of teaching children. Around the globe, approximately one third of all classes across all countries, including some of the more developed countries, are multigrade classes (2005 UNESCO Agenda for Educational Planning).

Multigrade classes have recently become a principal adjustment device for enrollment fluctuations also in many parts of Europe where demographic aging puts increasing pressure on class sizes. Warnings have been raised regarding the potentially detrimental effects of teaching students of different ages and maturity within the same room. At the same time, teaching several cohorts in one classroom has been suggested to have advantageous pedagogical side effects by providing more intense interactions between students of different ages that foster student-based learning.

Historically, many schools were restricted to particular religious denominations, which led to a restriction of student numbers, and consequently multigrade teaching, as the result of religious segregation. With denominational affiliation losing importance, this led to the abolition of denominational schools in many parts of Europe. At the same time, this abolition lifted size restriction and led to the abolition of multigrade classes. Mixed empirical evidence regarding the effects of abolishing denominational schools with multigrade classes on subsequent outcomes continues to fuel heated debates regarding the appropriate school organization.

This paper investigates the impact of the abolition of denominational schools with predominantly multigrade teaching on the long-term returns to education. The identification strategy exploits the natural experiment of a large-scale reform that led to the abolition of denominational schools in the Saarland, a state in Germany, in 1969. Prior to the reform, more than 95% of primary and lower secondary schools were church-maintained. In scarcely populated regions, the strict tracking by religious denomination imposed severe restrictions on the allocation of students. As a consequence, schools were relatively small, implying that students of different ages and skills were taught within the same classroom, i.e. in multigrade classes. The abolition of denominational schools in 1969 led to the dissolution of hundreds of these rural multigrade schools within less than a year. The remaining schools obtained a single-grade structure, similar to the larger schools in more urban environments.

The identification approach exploits differential treatment exposure of students depending on how many students of the same birth cohort have the same denomination. In more rural municipalities, multigrade teaching in denominational schools was the norm prior to 1969, but not afterwards. By contrast, in more urban municipalities multigrade teaching in denominational schools was not necessary due to higher student numbers. To estimate the effects of the reform on schooling and labor market outcomes we use an enhanced differences-in-differences approach.

By exploring the heterogeneity of the effects across gender, the evidence also provides new insights into the roots of gender inequality. In particular, the large-scale natural experiment enables insights into the socialization mechanisms at school that might lead to gender differences in labor market participation and occupational choice later on in life.

The empirical analysis is based on a unique combination of administrative records and comprehensive population census data. The dataset has been collected and digitized specifically for this research project, which to our knowledge is the first to exploit the abolition of denominational schools as a natural experiment in this context. Using municipality codes and schools' denominations, we are able to link individual-level census data on virtually all of Saarland's households in 1970 and 1987 to a comprehensive schools' index that comprises more than 7,500 school-year observations on a municipality-denomination-level. The availability of a wide range of schooling covariates allows us to control for channels like class size, school size, school consolidation, gender composition, etc. that might confound the multigrade effects.

The empirical results suggest that the abolition of multigrade classes had positive effects on final grade attainment and labor market participation. While all students profited from the abolition of denominational schools in terms of the higher grade attainment and a greater likelihood to become a white-collar worker, the effect is notably stronger for girls. The abolition of denominational schools in municipalities where multigrade teaching was the norm before 1969 led to an increase in the number of girls who attained a higher educational degree and a decrease in the number of girls becoming housewives. The results therefore suggest an interplay of gender socialization and the mode of teaching in terms of multigrade classes on subsequent outcomes.

The question how denominational schools with multigrade classes affect students' outcomes touches upon several research strands related to class composition, educational infrastructure, peer and tracking studies. Our empirical approach contributes to the literature in several ways. First, the natural experiment of the sudden abolition of denominational schools allows for a credible identification of the causal impact of denominational schools with multigrade classes, whereas many existing studies suffer from insufficient randomization which renders identification problematic (mainly because of self-selection). Second, we present effects that are placed in a Western European society. Many studies on multigrade classes with credible identification (due to controlled randomization) have been conducted mainly in developing countries, at the cost of limited external validity for more developed countries. Moreover, recent studies on multigrade teaching with credible identification focus on short-term educational outcomes. Third, the high-quality dataset covering virtually the complete population of our region of study minimizes selection and response biases and affords statistical power whereas existing research mostly relies on evidence from small samples. Fourth, provided with large-scale evidence, we are able to link gender mechanisms at school not only to final grade attainment but also to labor market participation and occupational choice. Our analysis thereby extends earlier work that mainly focused on the gender specific effect of class composition on schooling outcomes. Overall, our results are in line with the findings of earlier studies that suggest rather negative effects of multigrade classes.

The remaining part of the paper is structured as follows. Section 2 gives an overview of the existing literature on class composition. Section 3 describes the institutional background. Section 4 presents the identification strategy, followed by a compact presentation of the data in Section 5. Section 6 presents the empirical results, discusses robustness with respect to sensitivity checks and shows the results of the subgroup analysis. Section 7 concludes.

2 Literature Review

Multigrade classes[1] produce multiple forms of *peer effects*. Peer effects are central aspects of education research. They have been modeled as inputs to the education production function ever since Coleman (1968) made them popular, among others by (Iversen and Bonesrø nning, 2015; Jones, 2013). There exists relatively less research on peer effects of class composition than, e.g., on class size (Jones, 2013), but the absolute number of class composition studies is still vast. Many of those have been criticized for low methodological quality, however, as detailed in Lindström and Lindahl (2011) or Mason and Burns (1996). In general, a variety of peer effects can arise in a system of multigrade classrooms which has been touched upon as follows.

Between-student spillovers may be positive if more knowledgeable, skilled or able classmates serve as natural role models (Duflo *et al.*, 2011; Hanushek *et al.*, 2003). Practical relevance of peer collaboration, however, is told to be rather limited (Hattie, 2002). There is also evidence that peer effects are rendered negative if age gaps arise due to grade repeating and redshirting which is often the case in developing countries (Lavy *et al.* (2012) as well as Jones (2013)).

Finally, peer effects among teachers in the sense of shared experiences have been mentioned in the multigrade context. The probability of beneficial spillovers prerequisites at least two teachers per school and is likely to increase in larger teaching staff which puts rural schools at a disadvantage (McEwan, 2008).

Besides peer effects, also effects of (no) adjustments of teacher training, curricula, materials and incentives need to be reconsidered upon collapsing the grade level structure. Traditional teacher colleges prepare single-grade teaching although multigrade teaching is strategically more demanding and stressful (Mason and Burns (1996) as well as Russell *et al.* (1998)). Therefore, it is likely that multigrade schools have negative effects on students if the pedagogical infrastructure

[1] Multigrade classes, as opposed to single-grade classes (Veenman, 1995), do not sort students by age and skill. Furthermore, they are created out of some necessity, not pedgogical purpose, as other types of combination classes are.

is not adapted to multigrade teaching.

Current research on multigrade classes is frequently located in developing countries. See Little (2001) or McEwan (2008) for overviews in Africa, Asia and Latin America respectively. While some randomized control studies conducted in these countries convince by providing internal validity, their external validity is rarely given.[2] First, there are several institutional deficiencies that make it difficult to compare the examined multigrade settings to each other. For example, in some cases the mixed grade levels are not even adjacent (Mulkeen and Higgings, 2009) which increases the heterogeneity in the classroom substantially.[3] Second, unsafe school ways complicate school attendance asymmetrically for girls which changes the classroom gender distribution (Mulkeen and Higgings, 2009). Third, grade attainment may not mean anything regarding knowledge and skills (Jones, 2013). Due to this range of peculiarities in developing countries estimation of the effects of multigrade classrooms is challenging even to (quasi-)experimental designs that are good practice in the sense of Angrist (2004).[4]

Even though the major part of research on multigrade classes studies multigrade settings in development countries multigrade classrooms are also prevalent in more developed countries. Contemporaneously, multigrade classes make up one third of all classes on earth, and even in countries like Finland, the Netherlands, India, Peru, Sri Lanka and Pakistan multigrade predominate single-grade classes (Mulkeen and Higgings, 2009).

Existing studies on multigrade classes that were (mostly) conducted in industrialized countries up to 1995 are summarized in a meta-analysis by Veenman (1995). He concludes there are no significant effects on cognitive and/or social-emotional outcomes after averaging over 43 combination class studies meeting his econometric criteria. Apart from being quite outdated today these criteria were already criticized by contemporary scholars Mason and Burns (1996). They point out that Veenman (1995) draws on studies that use non-random samples. They argue that multigrade classes have better teachers and students. By that the group composition in multigrade classrooms biases an actually negative effect of less effective teaching in this setting towards zero.[5]

[2]Not only randomized control studies deliver evidence for multigrade effects in developing countries. Jones (2013) relies on an IV strategy to circumvent selection issues. He presents strongly negative effects by African overage-for-grade peers thus being supportive of Lavy *et al.* (2012).

[3]Furthermore, teachers in these countries often undergo very different trainings and the rate of teacher absence is very high. Enrollment is not compulsory but rather an achievement in itself, at any age (Jones, 2013).

[4]Vivalt (2015) establishes the overall limited external validity of impact evaluation studies formally.

[5]Concretely, multigrade teaching is found to cover less curriculum, especially in higher grades. Russell *et al.* (1998) back up the hypothesis that multigrade teaching is increasingly detrimental beyond basic skill acquirement. Furthermore he finds numeracy skills to suffer more than literacy from a multigrade structure in elementary schooling. To the extent of bias due to peer ability Mason and Burns (1996)'s critic is mitigated by Cullen *et al.* (2006). They present evidence from US school choice lotteries claiming no significant influence on student attainment by higher peer quality associated with the preferred schools. Their quality indicator measures the difference between (single-grade) classmates' average test scores after winning or loosing the lottery. Insignificance applies uniformly to ability, gender and race strata. It is also robust to all intensities of lottery-induced peer improvement.

A rather recent study on combination classes is the one by Lindström and Lindahl (2011). They rely on survey data and compare non-random but observationally equivalent single-grade and mixed-age classes in Sweden. They report a negative impact as sizable as that observed for larger classes in the STAR experiment.[6] Another recent approach to estimate effects of multigrade classrooms is presented by Leuven and Ronning (2016). Looking at multigrade schools in Norway they highlight the idea of *perspective-dependent* peer instruments obtaining contrastive signs out of the same data. They find younger students to benefit from having older ones around while older students get worse results when younger ones are around.[7] Leuven and Ronning (2016) conclude seemingly inconsistent evidence to be rooted in researchers' unilateral approaches. Furthermore, they claim to reconcile the literature finding small but significantly positive peer effects conditional on an optimal allocation.[8] Subsequent investigations by Carrell *et al.* (2013), however, point out limitations of peer group interventions as proposed by Leuven and Ronning (2016) in the face of endogenous subgroup formation. They deliberately allocate weak and strong ability students enabling theoretically the largest possible spillovers. They do not foresee more able students to cut less able ones out of their circle leaving them with even worse academic attainments. Recent work by Checchi and De Paola (2018) estimate the effect of multigrade classes on the formation of student cognitive and non-cognitive skills exploiting institutional features of the Italian educational system establishing a minimum number of students per class. In a companion paper (Gerhardts *et al.*, 2021) we provide evidence on the causal effect of multi-grade teaching in primary schools on literacy skills by the end of primary school exploiting the variation i policies across the federal states in Germany.

In view of the existing research on multigrade classes our study contributes to the literature in several ways: Our study focuses on the impact of the multigrade setting in German schools and uses a natural experiment – the sudden abolition of denominational schools – for identification of the causal effect of multigrade schools. By contrast, existing studies like those of Lindström and Lindahl (2011) and Leuven and Ronning (2016) suffer from insufficient randomization and rely on selection-on-observables methods which render causal identification problematic. Furthermore, we present effects of multigrade classes that are placed in a Western European society while those studies on multigrade classes with credible identification have been conducted mainly in developing countries. But, as described above, there are quite a few limitations of the in-

[6]In the STAR framework the presence of about six more students reduces test scores of classmates by 4 percentage points in the first year and 1 additional percentage point in subsequent years (Krueger, 1999).

[7]Concretely, they refer to Sims (2008) deriving negative impacts from measuring exposure to lower grade levels thus taking the perspective of the harmed older students. Along the same pattern Thomas (2012) is expected to find positive peer effects because he considers higher grade levels that are taught together with the treated younger students.

[8]Similarly Duflo *et al.* (2011) uncover contrastive spillover effects for high and low achievers in Indonesian (single-grade) schools. However, after taking into account lasting consequences of more adequate curricula (detailed in Glewwe *et al.* (2009)) and teachers' tendency to teach to the top of the class, Duflo *et al.* (2011) find tracking to be beneficial for all students. Yet another (single-grade) example where curriculum adjustments persistently outweigh peer effects is presented by Cortes and Goodman (2014) looking at US schools.

stitutional settings in these countries which diminishes the external validity of the findings for industrialized countries. Additionally, we possess a high-quality dataset covering virtually the complete population of our region of study. Thus, we do not have to deal with selection and response biases as much as studies relying on survey data (such as Lindström and Lindahl (2011)). Another advantage of being provided with large-scale evidence is that we are able to explore the effects of multigrade classrooms not only with respect to final grade attainment (as most existing research is confined to) but also to labor market participation and occupational choice. Extending the multigrade analysis to an interplay of medium-run outcomes (as pioneered in other contexts by Clark and Del Bono (2016) and Greenwood *et al.* (2016)) is new to the literature.

3 Institutional Background

This section describes the school reform in the region of our study, the framework of schooling laws, as well as potential confounders, using information from various sources.

Prior to the reform in 1969, almost all *Volksschulen* sorted students by denomination. This allocative restriction created multigrade classes in regions with a low population density. Figure 1 provides a first overview of the prevalence of multigrade classes in the Saarland prior to the reform.[9] With few exceptions denominational schools played a role only in the lowest educational track. For a concise description of ability tracking in German schools see Pischke and Wachter (2005).[10]

Schools providing primary or lower secondary education were uniformly labeled *Volksschule*, see Figure 2 in the appendix for a more details on the distribution of school types over time.

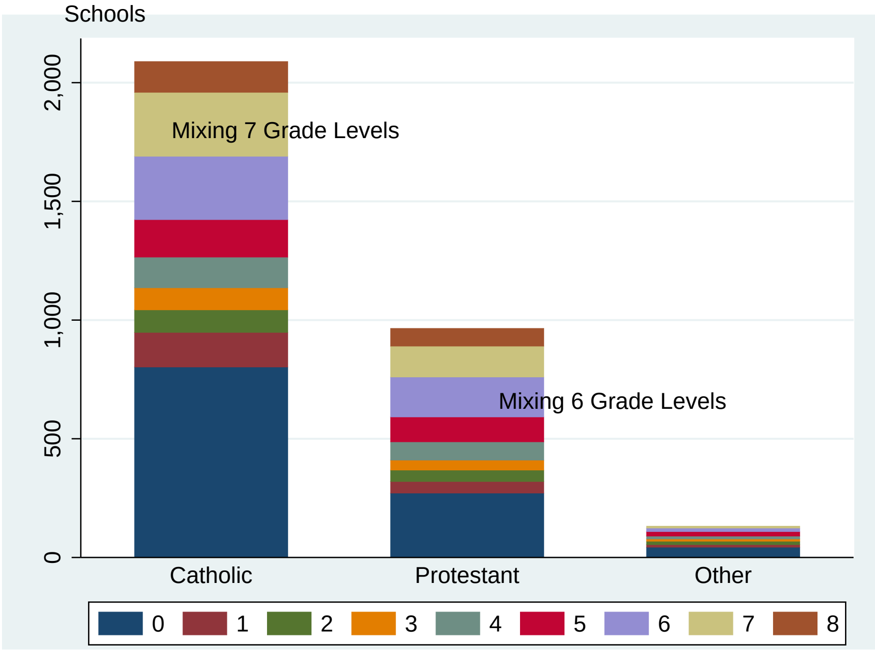
Prior to the abolition of denominational schools, the treatment exposure (the probability of being taught in a multigrade school) of students was dependent on how many students of the same birth cohort had the same denomination – due to the legal obligation to teach Catholics and Protestants separately.[11] In sum, 75% of schools in the Saarland resolved to a multigrade structure prior to the reform in 1969, all of which were schools in more rural regions. De-

[9]Rural *Volksschulen* create a multigrade setting not supported by pedagogical adjustments. First, the schools' records do not provide any evidence for adjustments. Moreover, albeit this is no rocket-science, there do exist alarming hints about amateurishly adapted teaching practices, available at `http://www.spiegel.de/spiegel/print/d-46265072.html` (01 May 2015). which highlights the comparability problem to mixed-age classes (Mulkeen and Higgings, 2009).

[10]Multigrade classes in remote regions pool children of very different abilities. Do the observed spillovers of our study provide guidance for inclusion of handicapped children as well? This depends on the multigrade school employing a full inclusion policy. Iversen and Bonesrø nning (2015) explore spillovers in Norwegian elementary schools where special education happens to be integrated within ordinary classrooms. They find that spillovers interact with the level of special education provided. In Germany the *Volksschule* and special schools are kept apart. After reforming lower secondary education the separation persists (Figure 2). Thus the insights by Iversen and Bonesrø nning (2015) formalize the lack-of-comparability argument forwarded in Veenman (1995) by which he excludes studies on gifted as well as handicapped children from his synthesis.

[11]Verfassung des Saarlandes (1947) Art. 27 (Amtsbl. des Saarlandes, Nr. 41) Vom 05.11.1969, *available at* `http://www.verfassungen.de/de/saar/saarland47-index.htm` (23 May 2015).

Figure 1: Mixed Grade Levels by Denomination



Notes: This figure shows the prevalence of multigrade teaching prior to the reform in 1969 by denomination. The category 'Other' mainly consists of non-denominational schools. Each color represents the amount of grade levels that were taught together. Red, for instance, shows the number of schools that were teaching 5 grade levels simultaneously.

Source: Schools' Index 1964-1986. Own calculations.

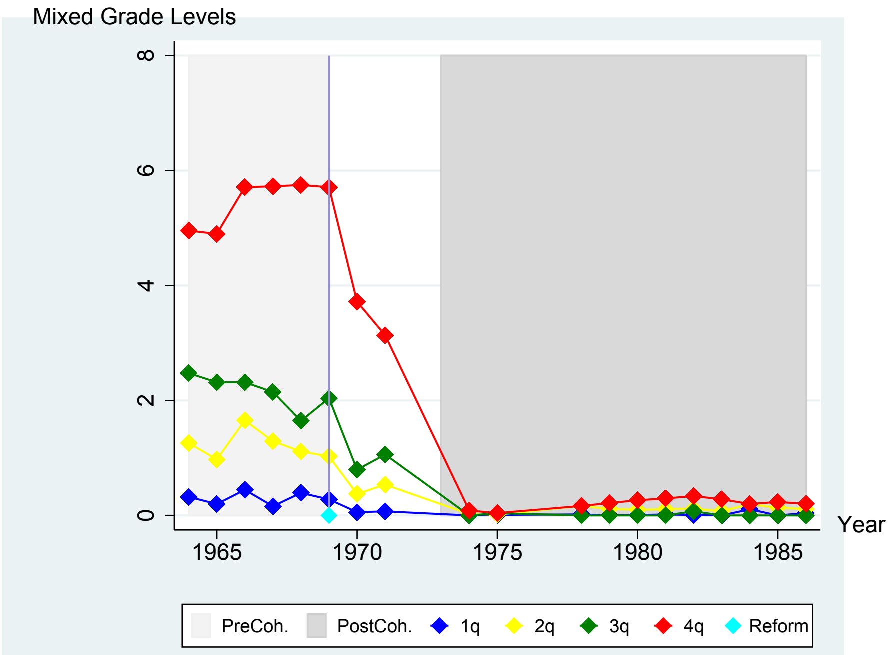
nominational schools in more urban regions, by contrast, were characterized by a single-grade structure.

The reform of 1969 had a direct impact on schools offering basic education. Inducing a change in students' distribution across school types it also indirectly affected higher education though. When denominational schools were legally abolished in various states all over Germany, this raised hot debates and interventions on behalf of the church and parents likewise[12] but in the Saarland the reform was carried out neatly.

Due to the reform the number of multigrade schools decreased by two thirds in less than a year and from 1974 onwards the share of multigrade schools was negligible. Thus, the reform changed the learning environment for children in more rural regions where multigrade schools predominated prior to the reform in 1969 substantially. Tiny schools were wrapped up into normal-size ones reducing the number of village schools by more than 50% while diminishing the frequency of more urban schools only moderately. In consequence, from 1974 onwards the prevalence of multigrade teaching was close to zero in both treated and control regions, see Figure 2 for the development of multigrade teaching in Catholic schools over time and Figure 3

[12] `http://www.spiegel.de/spiegel/print/d-46369565.html` (01 May 2015).

Figure 2: Mixed Grade Levels by Treatment Probability over Time: Catholic Students



Notes: This figure shows, for the case of Catholic students, the prevalence of multigrade teaching (diyplaying the number of mixed grade levels) over time by treatment probability (in quartiles). The treatment probability depends on the number of schools in a municipality-denomination-cohort-cell that were offering multigrade teaching prior to the reform.

Source: Schools' Index 1964-1986. Own calculations.

for the case of Protestant schools.[13]

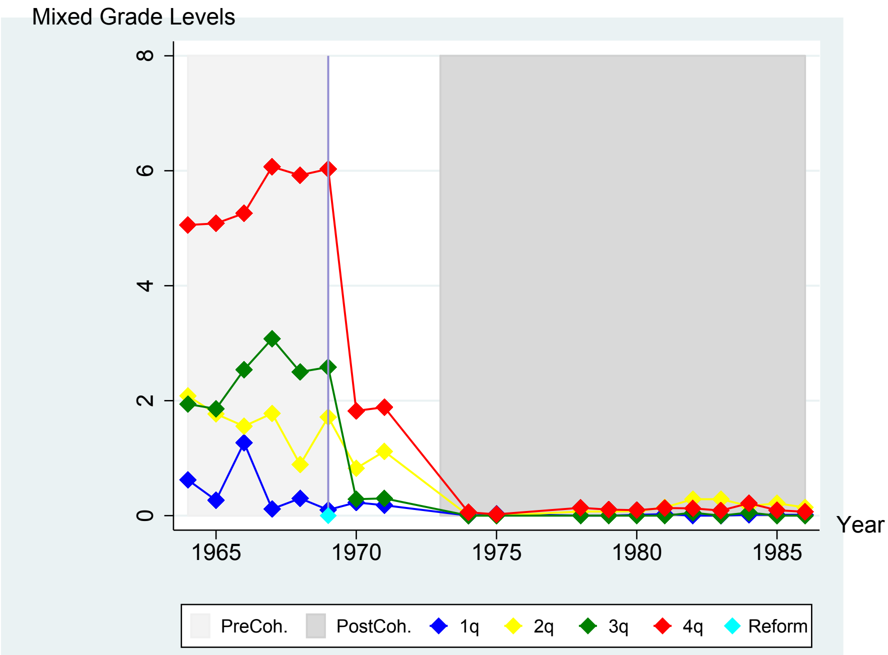
The abolition of denominational schools left some villages without an own school altogether and required their children to become commuters. Having to commute anyway changed relative commuting costs to higher education schools that might previously have been prohibitive. Attending a restructured *Volksschule* or even opting for a higher education school, either way rural students were taught in much more homogeneous classes.

All key features of schools are summarized in Tables 1 and 2, partitioning the universe of *Volksschulen* into four groups, namely treated and control schools, each before and after 1969 (and separately for Catholic and Protestant schools).[14] As the tables show, by construction the reform reshaped the educational infrastructure in multiple ways and also implied more students and more teachers per school in absolute terms (EENEE, 2015). For example in the case of Protestants living in treated municipalities where multigrade teaching was the norm prior to the reform, average school size increased from 89 students per school to 227 students per school and from 2.8 teachers per school to 11.4 teachers per school (see Table 2). At first sight surprisingly, average class size shrank because the inflow of remote area children into more urban school

[13]Tables A.1, 8 and 9 in the appendix compare the number of mixed grade levels in treated and control regions prior and after 1969 separately for Catholic, Protestant and (the few) non-denominational schools.

[14]The key features of non-denominational schools are shown in Table 10 in the appendix.

Figure 3: Mixed Grade Levels by Treatment Probability over Time: Protestant Students



Notes: This figure shows, for the case of Protestant students, the prevalence of multigrade teaching (diy-playing the number of mixed grade levels) over time by treatment probability (in quartiles). The treatment probability depends on the number of schools in a municipality-denomination-cohort-cell that were offering multigrade teaching prior to the reform.

Source: Schools' Index 1964-1986. Own calculations.

Table 1: School Characteristics by Treated and Control Status of Catholic Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Class Size	37.509	34.42	-3.089	(-9.987)	23.24	21.701	-1.539	(-8.478)
Pupils/Teacher	36.354	34.621	-1.733	(-5.678)	20.076	21.002	.926	(4.471)
Pupils/School	369.435	109.818	-259.617	(-47.896)	284.636	127.83	-156.806	(-28.302)
Girls' Share	.527	.49	-.037	(-6.04)	.48	.49	.01	(4.738)
Female Teachers' Share	.459	.427	-.033	(-3.965)	.526	.529	.004	(.471)
Teachers/School	10.125	3.056	-7.069	(-48.393)	14.292	6.155	-8.137	(-29.917)
Observations	1216	1021			2667	872		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Catholic students and the schools they attended are considered.

Source: Schools' Index 1964-1986. Own calculations.

districts was mitigated by a demographic decline in enrollment. It drastically reduced overall class size from 39 (1964) to 19 (1986) students on average, but the relative change was identical for treated and control regions. However commuting students coming from remote areas might have encountered higher quality peers from more urban municipalities (Leuven and Ronning, 2016).

Table 2: School Characteristics by Treated and Control Status of Protestant Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Class Size	32.409	31.009	-1.4	(-3.193)	23.045	22.344	-.7	(-3.913)
Pupils/Teacher	31.465	31.404	-.061	(-.105)	20.336	20.215	-.122	(-.599)
Pupils/School	270.118	88.962	-181.156	(-27.511)	252.835	226.88	-25.955	(-4.335)
Girls' Share	.51	.494	-.016	(-1.933)	.483	.483	0	(-.06)
Female Teachers' Share	.517	.463	-.054	(-4.265)	.526	.528	.002	(.288)
Teachers/School	8.607	2.772	-5.835	(-27.904)	12.599	11.414	-1.185	(-3.986)
Observations	374	448			2607	932		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Protestant students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

For the comparison between treated municipalities (where multigrade teaching was the norm prior to 1969) and control municipalities (where single-grade teaching was the norm prior to 1969) to make sense a common trend between those regions is essential. The 1960s are called the decade of educational expansion and changes over time are indeed tremendous. We exploit that the reform eradicates multigrade classes which creates an asymmetry between otherwise parallel worlds. The following important education laws in the Saarland are all implemented well before the reform is rolled out in 1969 and they maintain a common denominator for treated and control municipalities – those with and those without a history of multigrade schools – over time.

To begin with the *Compulsory School Entry Age* fixes enrollment into primary school to age six with minor exceptions referring to each June's 30th as cut-off date.[15] Next *Compulsory Schooling Duration* requires that students stay in school for at least nine years and passing the ninth grade is rewarded with a lower secondary degree. It turns out that roughly 4:1 students finish a ninth grade already before the law inures in 1965 (Pischke and Wachter, 2005). However its implementation requires two short school years that actually compress schooling duration in 1966/67. Then, *No Tuition Fees* guarantee basic education to be free of charge, independent of the school being state- or church-maintained.[16] It limits the influence of parents' financial constraints and prevents a selection by the fee itself. Finally, *Limited School Choice* of the parents is achieved by allocating students over schools based on catchment areas.[17] To choose a certain *Volksschule* by its reputation would require the household to move into that school's catchment area. Rothstein (2006) investigates parental preferences over school choice and es-

[15] §2 Satz 1 Gesetz Nr. 826 Schulpflichtgesetz *available at* `http://sl.juris.de/cgi-bin/landesrecht.py?d=http://sl.juris.de/sl/gesamt/SchulPflG_SL.htm\#SchulPflG_SL_rahmen` (12 June 2015).
[16] §1 Satz 1 Gesetz Nr. 662 Schulgeldfreiheit *available at* `http://sl.juris.de/cgi-bin/landesrecht.py?d=http://sl.juris.de/sl/gesamt/SchulGFrhG_SL.htm` (12 June 2015).
[17] §29 Satz 2 Schulordnungsgesetz vom 5. Mai 1965.

tablishes that peer groups matter even more than schools' effectiveness. This underlines the importance of student allocation by catchment areas because it mitigates parental choice effects which interfere with the core mechanism of multigrade classes. Jointly these laws provide accuracy in comparing schooling circumstances. This is an advantage compared to class composition studies of developing countries.

We analyze a period of more than two decades of schooling conditions. Our setup is robust to symmetric shocks. When screening the most influential historical events that could have had asymmetric impacts on treated and control municipalities, a primary concern relates to fluctuations in economic activity centered in urban regions. The coal and steel crises depressed the Saarland even more than the rest of Germany (Lichtblau, 2009). They caused dramatic peaks in unemployment and overshadowed positive shocks such as the construction of the Ford plant or the infrastructure improvement by the Saar Canal. Geographic controls measuring the distance to former major smelting works, direct access to the river, etc. are one possible solution to control for these changes. It is worth mentioning that despite of these shocks the Saarland was politically nearly perfectly stable (ibid). Only the very last year of our study's time horizon is subject to a different government, therefore we expect its influence to be limited. The advantage of exploring inner-state differences becomes obvious here. By construction, many complicating aspects like tax schedules causing potential problems in Abramitzky and Lavy (2011), etc. are taken care of from the start.

4 Empirical Model

The key empirical question refers to the comparison of the performance of students in a multigrade environment to a single-grade environment, which is less heterogeneous in terms of birth cohorts. We tackle this question estimating a triple differences (DDD) model that exploits exogenous variation in the probability to be a multigrade student over time, region and age group. Let Y_{1imdcy} represent individual i's outcome in municipality m with denomination d, belonging to cohort c and age group y if she attended a multigrade school and Y_{0imdcy} otherwise.

A student is defined as treated if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. If in one municipality there was one Protestant school teaching at least two grade levels jointly in all pre-reform years, then a Protestant student will be labeled as living in a multigrade municipality. This is still true if in the same municipality there exist Catholic schools which might be single-grade schools. This definition underlies the balancing tables 1, 2 and 4. It ensures that within a treatment-municipality-denomination-cohort cell the probability to attend a multigrade school was 100%.[18] Yet, this definition might be overly retrictive as it dismisses multigrade exposure

[18]We estimate an intentention-to-treat effect. Apart from the standard assumptions for multiple differences

whenever the probability was not 100%. In other words, citing the example from above, even if only in one year prior to the reform the Protestant school obtained a single-grade structure the Protestant student will be labeled as non-treated. Therefore, building on the binary defnition we employ two alternative continuous treatment indicators in our regressions.[19] Consider a municipality with two Protestant schools, school A with 90 and school B with 10 students. The school-based indicator corresponds to the share of multigrade schools, the student-based indicator to the share of multigrade students of the respective municipality-denomination-cohort cell. Table 3 shows which indicator behaves more conservative, in the common computational scenarios.

Table 3: Treatment Status by Alternative Multigrade Indicators

	Multigrade Indicator		
Multigrade School?	Binary	School-based	student-based
Case I			
Both A, B	1	1	1
Case II			
School A	0	0.5	0.9
Case III			
School B	0	0.5	0.1
Case IV			
Neither A nor B	0	0	0

Note: Fictitious example considering a municipality with two Protestant schools, school A with 90 and school B with 10 students. The continuous school-based indicator corresponds to the share of multigrade schools, the continuous student-based indicator to the share of multigrade students of the respective municipality-denomination-cohort cell.

The binary indicator underlying our balancing tests is very conservative in assigning treatment status. Thus, it is most likely to reveal significant differences that potentially create non-common trends. Nevertheless, as any binary indicator, it disregards that treatment probability is gradual. Therefore it should be modeled as a continuous variable, just as we do in our preferred specifications discussed in this paper. As Table 3 shows the school-based indicator computes the probability to attend a multigrade school based on the number of schools per municipality-denomination-cohort cell (MDC). The student-based indicator models the probability to attend a school within a MDC cell to be proportional to the school's size, as a proxy for its capacity to take in students. Note however that the latter need not be a better indicator per se. Smaller multigrade schools were often much more extreme in collapsing grade levels

analyis our setup requires two non-technical assumptions. First, pre-reform denomination of student and school coincide and second, the likelihood for treated and control students to start their own household follows a common trend while they are under-age. Conditional on these assumptions the probability to be treated assigned by the binary multigrade indicator is 100%.

[19]The binary treatment indicator is used in a robustness check. The results do not provide additional insights and are available upon request.

than larger schools had to be. This motivates to condition on treatment intensity, something we are still working on. Of course treatment probability and treatment intensity are two different things. This is just one example to point out that apart from school size there exist multiple factors influencing the possible multigrade experience of a student. From this perspective, the school-based indicator is just a neutral and thus very useful benchmark.

We estimate the *reform effect* in a regression with $Multigrade_{md} \in [0,1]$, a continuous variable measuring the likelihood of being taught in a multigrade class, the binary variable $c \in \{Pre, Post(Reform)\}$ and the binary variable $y \in \{Young, Old\}$, and a triple interaction, reflecting the DDD estimator. *Post* equals one for observations of the 1987 Census and zero for 1970. *Young* equals one for people aged fifteen to twenty in either census year and is zero for people aged 32 to 37 years.

$$\begin{aligned} Y_{imdcy} = {} & \beta_0 + \beta_1 Multigrade_{md} + \beta_2 Post_c + \beta_3 Young_y \\ & + \beta_{12} Multigrade_{md} Post_c + \beta_{13} Multigrade_{md} Young_y + \beta_{23} Post_c Young_r \\ & + \beta \underbrace{Multigrade_{md} Post_c Young_y}_{D_{mdcy}} + \psi_m + \epsilon_{imdcy} \end{aligned} \tag{1}$$

To account for time-invariant confounders at the municipality level, we include municipality fixed effects ψ_m. To allow for correlation of errors within municipality we cluster on the municipality level (335 clusters).

Identification is thus based on the contrasts across municipalities with a different coverage of multigrade schools prior to the reform, age groups, and time. We estimate the DDD baseline reform effect including just the main effects *Multigrade, Post, Young* and their interaction terms.

We proceed by estimating the multigrade effect in more extensive specifications that include additional individual controls from population census data. These include *Age, Age Square, Young at School Entry, Female, Catholic* and *German. Young at School Entry* relates birth month and school entry cutoff date to indicate if a student is relatively young within her cohort. Combining this with administrative data from school records allows us to include additional controls. These comprise municipality-denomination-cohort level regressors *Class Size, School Size* (defined as the number of students) *Girls' Share* and *Female Teachers' Share.* We furthermore account for *Potential Commuting Costs* which we define as the average distance to the nearest *Realschule* or *Gymnasium* net of the distance to the nearest *Volksschule.*

The identifying assumption of our DDD strategy is that multigrade exposure is as good as randomly assigned conditional on observables and unobservable-but-fixed confounders. Adding a control group of elder people nets out region-specific changes that are not rooted in schooling conditions themselves. An example would be a boost in multigrade municipalities' neighborhood quality induced by state-level interventions to counteract drift to the cities (characterized by single-grade schools). The setup still requires unobservable asymmetries in teaching effectiveness and ability differences between multigrade municipalities' and single-grade municipalities' students to be time-constant, because – with only two periods in which region-specific outcomes are measured – trends are not identified, a drawback detailed in Stephens and Yang (2014). Moreover we rely on the aforementioned student allocation via catchment areas to ensure that students do not choose their school, and thus their multigrade exposure. To sum up, for multi-dimensional differencing to be applicable group composition needs to be spatially stable as well as groups should follow a common trend over time. Furthermore we assume *zero conditional mean, additive separability* and a *constant, weakly monotone causal effect* β.

5 Data

This section describes the data. Via municipality codes we combine two censuses and one schools' statistics, all of which are comprehensive, high-quality administrative datasets.[20]

Outcomes[21]

We construct schooling and labor market outcomes using individual-level census data from 1970 for the baseline and from 1987 for the follow-up cohorts. The data is available via remote execution at the German Federal Statistical Office. To evaluate final grade attainment we consider two separate dummies, namely (1) attainment of *Mittlere Reife or Fach-/ Abitur* (i.e. at least an intermediate secondary degree) and (2) attainment of *Fach-/ Abitur* (i.e. at least a high-school degree). Looking at grade attainment instead of years of schooling reflects longer schooling net of grade repetition and also identifies dropouts (EENEE, 2015). There are no test scores in the data. If there were, however their predictive power might have been limited anyway by grading on a reference curve, especially in a multigrade class, because relative grading depends on the presence of more advanced peers (Leuven and Ronning, 2016). Importantly, peer effects may trigger social competences not captured by test scores but perhaps reflected in post-schooling attainment. We therefore also use labor market outcomes to assess lasting or reemerging effects of schooling similar to Chetty *et al.* (2014b). In order to analyze labor market participation we use binary indicators on unemployment and labor market participation. Given labor market entry we distinguish further between blue- and white-collar occupations to capture the socio-economic status of the occupation. Note that wages are not reported in the

[20] Volkszaehlungsgesetz 1970 vom 14. April 1969 (BGBl. I S. 292); Volkszaehlungsgesetz 1987 vom 8. November 1985 (BGBl. I S. 2078).

[21] Nearly all our outcomes are binary. Accordingly, the OLS regressions represent linear probability models (LPMs) which means that causality draws on the CIA, predictions may violate the [0,1] range and the error term is heteroskedastic (Angrist and Pischke, 2008).

Census 1987.[22] Table 4 shows descriptive evidence on the differences between treated and control individuals with respect to their schooling and labour market outcomes. It shows that treated individuals prior to the reform were less likely to hold at least a *Realschule* degree (RS degree) than control individuals. Furthermore, they were more likely to have a blue-collar job and less likely to have a white-collar job. According to the descriptive statistics, these differences were less pronounced after the reform. In fact, after the reform treated individuals are more likely to hold at least a *Realschule* degree than control individuals.

Table 4: Descriptive Statistics: Treatment, Outcomes and Controls

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Treatment Indicators								
MDC MG School Share	.259	1	.741	(397.086)	.028	.122	.094	(59.844)
MDC MG Pupil Share	.088	1	.912	(821.797)	.005	.064	.06	(53.115)
Outcomes								
At least RS Degree	.094	.08	-.014	(-5.298)	.371	.392	.021	(3.751)
At least A-levels	.009	.007	-.002	(-1.834)	.067	.069	.002	(.614)
Employed	.651	.653	.001	(.328)	.688	.707	.019	(3.674)
Non-Participant LM	.071	.07	-.001	(-.349)	.045	.032	-.013	(-5.694)
Blue-Collar Job	.514	.548	.034	(7.485)	.525	.538	.013	(2.313)
White-Collar Job	.407	.364	-.043	(-9.674)	.428	.428	0	(-.019)
Controls								
15-17 Year-olds	.417	.43	.013	(2.919)	.218	.227	.009	(1.966)
1 VS in MDC cell	.297	.902	.604	(156.353)	.325	.82	.495	(97.561)
Mun: max.5000 inh.	.233	.882	.649	(178.032)	.307	.893	.586	(121.127)
Female	.498	.488	-.011	(-2.376)	.449	.435	-.014	(-2.442)
Age	17.846	17.794	-.052	(-3.58)	18.566	18.518	-.048	(-3.276)
Young Within Cohort	.396	.402	.007	(1.489)	.372	.38	.008	(1.513)
Catholic	.804	.692	-.112	(-30.104)	.804	.682	-.123	(-26.083)
Protestant	.187	.292	.106	(28.829)	.17	.277	.107	(23.929)
German	.967	.979	.012	(7.898)	.952	.968	.016	(7.055)
Single	.895	.893	-.002	(-.75)	.944	.951	.007	(2.654)
Household Size	4.376	4.65	.274	(15.244)	3.742	4.039	.297	(19.378)
MDC Class Size	37.037	34.447	-2.59	(-77.175)	23.215	22.337	-.878	(-52.595)
MDC Pupils	380.272	133.094	-247.178	(-252.53)	296.094	170.926	-125.168	(-112.362)
MDC Girls Share	.531	.494	-.037	(-71.901)	.482	.486	.005	(22.21)
MDC Fem.Teachers Share	.477	.405	-.072	(-73.767)	.531	.514	-.016	(-11.977)
Commuter to VS	.045	.173	.128	(55.238)	.03	.339	.308	(96.759)
Commuting to VS (km)	.132	.521	.389	(48.743)	.054	.996	.942	(55.032)
Commuting to RS (km)	3.045	6.412	3.368	(82.812)	1.909	3.915	2.006	(53.953)
Commuting to Gym (km)	2.604	6.383	3.779	(95.454)	2.672	5.12	2.448	(50.949)
Commuter	.566	.664	.098	(21.521)	.649	.71	.062	(11.168)
Observations	54465	15694			30245	10456		

Notes: In this table, we differentiate between control and treated students (between 15 and 20 years old) pre and post to the reform in 1969. A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. MDC = municipality-denomination-cohort, MG = multigrade, VS = Volksschule, RS = Realschule, Gym = Gymnasium, LM = labor market.
Source: Integrated dataset of Census 1970 and 1987 and Schools' Index 1964-1986. Own calculations.

Treatment Indicator

We determine each individual's likelihood for having been a multigrade student – considering

[22] For a follow-up version of this paper, we consider to assign a standard income range based on each observation's meticulously reported profession (ISCO 88) for income mobility analysis in the sense of Chetty *et al.* (2014a).

each individual's municipality and denomination – computing two alternative continuous treatment indicators as explained in Section 4. The school-based indicator corresponds to the share of multigrade schools, the student-based indicator to the share of multigrade students of the respective municipality-denomination-cohort cell (MDC).[23] Table 4 shows that on average 26% of those students defined as control by the *binary* indicator are assigned a positive treatment probability by the *school-based* indicator. In contrast, 8.8% of those students defined as control by the *binary* indicator are assigned a positive treatment probability by the *student-based* indicator.

Controls

Using data from Saarland's Statistical Office, we obtain records on all primary and lower secondary schools from 1964 to 1986. Key figures like the numbers of male and female students and teachers, the number of classes, school's type, denomination and address are given for each school on an annual basis yielding more than 7500 school-year observations.[24] The school's address enables us to average over schooling conditions of schools of a given denomination in a given municipality in a given year. We then group the years into pre and post reform and match them to individuals in the baseline and follow up cohorts respectively via the municipality code while also considering an individual's denomination.[25] Importantly, for 80% of all schools (attended by roughly 50% of all students) a unique mapping between a student of a given denomination and the school of her denomination is possible (i.e. there is no need to match the student to an average of school characteristics of two or more schools of her denomination).

By help of the schools' records we compute pre- and post-reform municipality-denomination-cohort (MDC) averages of class size, student-teacher ratio, school size (in terms of number of students), girls' share and female teachers' share. Table 4 compares the main schooling characteristics between schools in treated and control municipalities. Importantly, class size, the principal rivaling input when estimating the effect of multigrade schools, is a bit lower in treated regions (on average, there were 2.6 students less per class). Since a smaller class size has presumably beneficial effects on students' achievement, this fact will rather lead to underestimating the effects of the abolition of multigrade classes when not controlling for class size.

The census data provide us with a set of individual-level controls all displayed in Table 4, most of which are commonly used and self-explanatory. The differences between treated and control individuals are in line with expectations: Treated individuals are more likely to live in

[23]See Table 3 for gaining an intuition of the different behavior of both indicators.

[24]We exclude special schools. Records for the years 1971/72 are missing completely. For 1966 one fifth of the data is missing but without region-specific missing patterns.

[25]In order to calculate average post-reform schooling conditions, we take schools' records from 1973-1986 into account. The cohorts of interest analyzed out of the 1987 Census are at most 20 years old in 1987 implying they entered primary school earliest in 1973.

municipalities with less than 5,000 inhabitants (88% vs. 23%), and are more likely to have only one Volksschule (VS) in their municipality-denomination-cohort cell (MDC), namely by 90% vs. 30%. Moreover, treated individuals are less likely to be Catholic (70% vs. 80%).

Here we briefly discuss those controls with non-standard implications. In our setting, some standard controls like household size and marital status are potentially bad control because the reform likely affects marriage and/or fertility behavior (Lundborg *et al.*, 2012). The bad control case is even more pronounced for potential commuting costs. Students forced to commute are facing different effort costs than those attending school in direct vicinity. So continuing school at all is decided on altered premises. Simultaneously the implicit 'vicinity bonus' of lower secondary schools over higher education schools disappears in rural regions. Commuting anyway, ability-based school choice seems more natural than it has been with a *Volksschule* at walking distance and higher education schools at multiple kilometers' distance. Therefore we control for the distance to the nearest *Realschule* and/or *Gymnasium*. Importantly, however, we only include household size, marital status and commuting costs in an extended version of our regressions because we cannot rule out they are bad controls.

Sample Restrictions

Census data virtually cover all Saarlanders in each of the two survey years providing us with an unrestricted sample exceeding two million observations. We drop individuals younger than fifteen years because that is the minimum age for the outcomes we observe. Furthermore it is crucial to drop individuals between 21 and 32 years for two reasons.

First, before turning 21, people are still underage[26] such that their mobility is low. This matters because census data provide the municipality code of current residence and of school attendance. Fortunately, the residence-of-household definition ties children to their parents' address until they begin their own household.

Nevertheless, concerned with individuals moving reform-induced away from more rural regions (characterized by a higher likelihood of offering multigrade teaching) to urban regions we impose that underage restriction. It leaves us with a sample of main interest consisting of five consecutive birth cohorts with individuals who are between fifteen and twenty years old in either census. All of them attend primary and lower secondary school either strictly before or strictly after the reform takes place.

Second, although there is no panel structure at the individual level, observations of the 1970 Census reappear in the survey of 1987. Individuals between 32-37 years olds in 1987 have been past schooling age already in 1970 and are therefore untreated in either census. By construction their mobility cannot change reform-induced, so it is safe to include them as a control group.

[26]Legal definition as of 1970. For a subset of outcomes we run robustness checks restricting the sample to below 18 years, the legal threshold valid in 1987. This imitates what Lundborg *et al.* (2012) do facing the same problem.

However the case is much more complicated for individuals between 21 and 32 years old in 1987. They have been partially treated because they are still in lower secondary school when the reform is rolled out in 1969. With respect to multigrade exposure they fall into a transition period with exceptional schooling conditions due to fundamental restructuring. Therefore, we exclude them from our sample. Note that the seventeen-year elapse between both censuses is just short enough to preclude that parents of the post-cohorts have already been treated. Otherwise multi-generational class composition effects could accumulate, a channel established in Lundborg *et al.* (2012). Admittedly, the framework cannot rule out general equilibrium effects, a caveat that needs further investigation.

We furthermore restrict the sample to individuals for whom we have information on the outcomes of interest. In the end, our final dataset consists of 287,153 individuals when combining both age groups. When taking only into account the younger individuals of both censuses (aged between 15-20 years) the sample consists of 111,081 individuals.

6 Results

This section presents estimates of the impact of the abolition of multigrade schools on schooling and labor market outcomes. Our findings are in line with the literature suggesting a negative net effect from multigrade classes whenever other education inputs are not adapted accordingly. We show that results are robust to the inclusion of a wide range of individual characteristics and schooling covariates. Moreover we stratify the sample to investigate heterogeneity of the multigrade effect across subgroups. Throughout, we show (1) estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator, (2) estimates of the DDD estimation using the school-based multigrade indicator, (3) estimates of the DDD estimation using the student-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. The latter respects the number of students (school size) upon averaging. Both indicators are continuously defined over 0 and 1 and measure each individual's multigrade exposure/treatment probability precisely.

Overall Results

Schooling Outcomes

Table 5 presents the main results based on the whole sample. We show estimates of the two different continuous DDD specifications (using the student-based multigrade indicator and the school-based multigrade indicator respectively) as well as estimates of a DID specification (using the school-based multigrade indicator). For each specification, we show estimates of the baseline approach (not including any controls), the core controls approach (not including potentially bad controls, see Section 5) and the extended controls approach (including all controls).[27] DID as well as DDD regression results displayed in Table 5 suggest that the abolition of denominational schools favorably influenced degree attainment. This finding is remarkably robust across our different specifications. According to the estimated coefficients the change from a multigrade school system to a single-grade school system significantly raised the average probability of attaining an intermediate secondary degree (*Mittlere Reife* or *Abitur*) by 7-11 percentage points (ppt), depending on the specification. The effect on having attained a high-school degree (*Abitur*) is also positive and indicates that the switch to a single-grade school system led to an increase of students holding a *Abitur* of around 5 ppt. A natural explanation of this finding would be that individuals spend more time on schooling because single-grade classes improve basic training. This in turn makes superior educational attainment accessible.

Professional Outcomes

The estimates in Table 5 show that the reform did not change the overall probability of being employed. Yet, we observe a reform-induced increase of the likelihood of holding a white-collar job and a reform-induced reduction of the likelihood of becoming a non-participant in the labor market (in other words, in the case of women, becoming a housewife). Interestingly, the labor market estimates get more precise and larger when adding the control group of elder people, i.e. turning from the DID-estimation to the DDD-estimation. This indicates that the increased take-up of white-collar jobs is not due to a region-specific labor market trend. The global gain in white-collar employment seems to be partly driven by female labor market participation which is reflected in the housewife/non-labor-market-participation status declining by 3 ppt. Below, we discuss channels of gender-specific responsiveness to the treatment in more detail. In sum, results suggest that reform-induced higher educational attainment led to an increase of better qualified employment.[28]

Sensitivity Analysis

Table 11 in the appendix shows the results of the main regressions – using the core controls approach – when restricting the sample in two different ways. For the sample used for regressions in the upper part of Table 11, we only take into account individuals for whom the municipality where they went to school is definitely known, i.e. we can exclude migration in order to take up

[27] We present the overall results for all three approaches. In the cases of the sensitivity analysis and the subgroup analysis, however, we only display the results of regressions including the core controls. The results of the other specifications are available upon request.

[28] The importance to assess general equilibrium effects for policy recommendations is detailed in Heckman *et al.* (2014). As mentioned before the sizable period elapsing between pre- and post cohorts' outcomes heightens the probability that general equilibrium effects understate or overstate positive effects from improved education. Disentangling the partial effect we are interested in and the general effect offsetting it requires a joint estimation of skill supply and demand elasticity. The latter lies - for now - beyond the scope of our study.

Table 5: Overall Effects on Schooling and Labour Market Outcomes of 15-20-Year-Olds

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
DDD (pupil-based)						
Baseline	0.112***	0.0375	0.00277	-0.0165	0.0473	-0.0254
	[0.0367]	[0.0242]	[0.0395]	[0.0284]	[0.0311]	[0.0194]
Core Controls	0.112***	0.0407*	0.0187	0.00427	0.0463	-0.0457**
	[0.0363]	[0.0239]	[0.0446]	[0.0253]	[0.0304]	[0.0191]
Extended Controls	0.111***	0.0409*	0.0153	-0.00264	0.0463	-0.0391**
	[0.0367]	[0.0235]	[0.0423]	[0.0250]	[0.0287]	[0.0160]
DDD (school-based)						
Baseline	0.0903***	0.0520**	0.0296	-0.0285	0.0529**	-0.0194
	[0.0290]	[0.0229]	[0.0330]	[0.0241]	[0.0267]	[0.0171]
Core Controls	0.0898***	0.0534**	0.0371	-0.0162	0.0528**	-0.0320**
	[0.0286]	[0.0225]	[0.0348]	[0.0233]	[0.0262]	[0.0162]
Extended Controls	0.0912***	0.0529**	0.0396	-0.0193	0.0514**	-0.0280**
	[0.0287]	[0.0223]	[0.0333]	[0.0227]	[0.0242]	[0.0140]
DID (school-based)						
Baseline	0.0868***	0.0117	0.0515	-0.0242	0.0543*	-0.0164
	[0.0281]	[0.0122]	[0.0384]	[0.0271]	[0.0296]	[0.0130]
Core Controls	0.0823***	0.0115	0.0443	-0.00480	0.0358	-0.0187
	[0.0286]	[0.0124]	[0.0359]	[0.0264]	[0.0303]	[0.0116]
Extended Controls	0.0817***	0.0113	0.0324	-0.00750	0.0354	-0.0178**
	[0.0281]	[0.0121]	[0.0365]	[0.0261]	[0.0278]	[0.00794]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
N (DDDpupil)	287153	287153	287153	287153	287153	287151
N (DDDschool)	287153	287153	287153	287153	287153	287151
N (DID)	111081	111081	111081	111081	111081	111079
Cluster (DDDpupil)	337	337	337	337	337	337
Cluster (DDDschool)	337	337	337	337	337	337
Cluster (DID)	333	333	333	333	333	333
Adj.R2 (DDDpupil)	0.129	0.0797	0.276	0.289	0.0931	0.510
Adj.R2 (DDDschool)	0.129	0.0797	0.276	0.289	0.0931	0.510
Adj.R2 (DID)	0.181	0.0660	0.172	0.234	0.189	0.550

Notes: This table shows in the upper part estimates of the DDD estimation using the student-based multigrade indicator, then it shows the estimates of the DDD estimation using the school-based multigrade indicator and in the bottom part it shows the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the table shows estimates of the baseline approach (not including any controls), estimates of the core controls approach (not including potentially bad controls) and estimates of the extended controls approach (including all controls).
Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.
Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

employment elsewhere. This implies that this group of individuals represents a negative selection – they might be more afraid to move away from home or do not have sufficiently good skills to get employed elsewhere. The results in Table 11 are in line with this negative selection argument. While we observe a similar reaction to the switch from multigrade to single-grade teaching in terms of the attainment of a higher secondary degree, the labor market response is much smaller than for the whole sample. For results in the bottom part of Table 11, we restrict the sample to those individuals who live in those municipalities in which a unique mapping between individual and school is possible (since there is at maximum one school per denomination prior to the reform). This restriction makes a clean attribution of school controls possible. The disadvantage of this restriction is that we are left with the very small municipalities, and face, again, the problem of negative selection: those students staying in small villages are probably less ambitious. The results in Table 11 are very similar to the overall findings in Table 5. In contrast to the upper part of Table 11 we also find a significant negative effect on the likelihood to become a housewife/non-participant in the labor market.

Subgroup Analysis

Related studies motivate robustness checks by gender and denomination which we present in the following.

Boys & Girls

While the reasons for gender-specific reactions to education policies are still debated their existence has been shown repeatedly. Along these lines Angrist and Lavy (1999) find incentives pushing college certification rates only for Israeli girls. Deming *et al.* (2014) document gender-dependent attainment gains in US post-secondary education where only girls respond to higher school quality. These findings are complemented by relatively higher female responsiveness to tracking (Duflo *et al.*, 2011). However Whitmore (2005) draws on the STAR experiment to single out gender-neutral gains by class size reduction. As shown in Table 12 in the appendix, Saarland's data confirm girls' final grade attainment to improve more strongly than that of boys in the case of secondary education. While the switch from a multigrade system to a single-grade system led to a 11-16 ppt increase in a girl's likelihood to attain at least a secondary degree, it increased a boy's likelihood to attain such a degree by only 5-8 ppt which is already strong. Regarding the probability of attaining at least a high-school degree (*Abitur*), however, girls fare somewhat worse. Interestingly, as regards labor market outcomes, we do not observe large differences across gender and, moreover, the coefficients are not significant when splitting the sample. Yet, results in Table 12 show that the switch from a multigrade school system to a single-grade school system decreased the likelihood of becoming a housewife/non-participant in the labor market significantly for girls, but not for boys. What are potential explanations for girls benefiting more than boys from the disappearance of multigrade teaching? One possibility refers to girls being on average higher achieving than boys. Analogously it could be that their trajectories of improved education inputs are steeper. The literature also suggests girls to be less competitive than boys (Leuven and Ronning, 2016). Thus learning in highly heterogeneous multigrade groups might be more demanding for them. Consequently, they profit more from the switch to single-grade classes.

Catholics & Protestants

Table 13 in the appendix shows the estimated coefficients for the sample stratified by denomination. Overall, it indicates that both groups of individuals benefited from the reform in terms of their educational outcomes. Surprisingly, Protestants seem to have gained by much more than Catholics did. Moreover, Table 13 shows insignificant and close-to-zero labor market effects for

Catholics, while it indicates large and significant reform-induced gains for Protestants. What are potential explanations of this finding? Again, as in the case of explaining larger benefits of the reform accruing to girls than to boys, it could be that Protestant students are on average higher achieving than Catholic students and are therefore responding more to an increase of inputs into their education production function. This touches upon the Weber Hypothesis of Protestants' inherently superior work ethics, see Becker and Woessmann (2009) who connect wide-spread literacy to Protestants' prosperity. In a follow-up version of this paper, we will offer more evidence to gain a deeper understanding of the reasons for the heterogeneity of our findings with respect to denomination.

7 Conclusion and Outlook

This paper addresses the question how attending a multigrade school affects school attainment and labor market outcomes, and whether there are any differences by gender or denomination in this effect. To answer this question our analysis exploits the abolition of Saarland's denominational schools as a natural experiment that overcomes the main challenges of impact evaluations for policy design (McEwan, 2008).

The reform produces a sharp treatment effect, in terms of the variation of the reduced probability to attend a multigrade class caused by an exogenous event, namely the abolition of denominational schools. Based on a legal change that is rapidly and comprehensively accomplished the setup provokes, if any, negligible anticipation or conditional-on-participation effects. Highly accurate school-level data allow us to control for rivaling changes in the educational infrastructure that are also implied by abandoning denominational tracking. The estimation approach based on triple differences plausibly identifies causal links between treatment and outcome candidates. Our results are remarkably robust across specifications and unambiguously suggest single-grade classes to be more beneficial for students' educational and labor market outcomes. Due to the reform treated students shift away from obtaining only a lower secondary degree (*Volksschulabschluss*) and a blue-collar job. Their probability to attain at least an intermediate secondary degree (*Realschulabschluss*) and to become a white-collar employee increases significantly when switching from a multigrade school system to a single-grade school system. Stratifying the main sample the emerging patterns line up with asymmetric treatment responses observed in related studies. Splitting the sample by denomination suggests that Protestant students profited more from the reform than Catholic students did. Moreover, we show that girls were more affected by the switch from a multigrade to a single-grade school system than boys. Our research approach provides external validity for the European context, which is particularly relevant in the light of the ongoing demographic change. To our knowledge this is the first study to exploit a large-scale experiment on multigrade classes in Germany. Policy interest in combination classes spans the globe but major empirical research is located in developing countries. Therefore, it suffers from limited external validity for the Eurpean context as third-world schooling bears many peculiarities. Saarland's data date back to the 1960s but the insights provided seem still easier adaptable for use in Europe. The village schools we observe are much more likely to produce positive peer effects than schools in developing countries doomed by overage-for-grade students. Our findings nevertheless suggest that a beneficial multigrade system needs strategic adjustments. We conclude that peer effects based on student collaboration alone are no panacea which refutes the argument that reallocation is a *costless* way to improve education.

Still, there are some open questions that we want to address in a follow-up version of this paper: Why do we observe stronger effects of the reform for Protestants? So far, we did not consider the pure effect of the abolition of *denominational* schools, but assume that the effects we find are the result of the disappearance of multigrade schools due to the abolition of denominational teaching. Yet, it might be that part of the *multigrade* effect is due to *denominational* teaching methods (that had a different impact in treated and control groups). Future research will thus try to disentangle the denominational effect from the multigrade effect. Furthermore, we will investigate in more depth why the shift from multigrade teaching to single-grade teaching has larger effects for girls. Using German data of the PIRLS study (Progress in International Reading Literacy Study) we will investigate whether gender-specific effects of multigrade teaching already arise at a young age. In particular, we will use the variation in the introduction of multigrade teaching in primary schools across German states between 2000 and 2010.

References

Abramitzky, R. and Lavy, V. (2011). How responsive is investment in schooling to changes in redistribution policies and in returns. *Econometrica*, **82**(4), 1241–1272.

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy*, **20**(2), 198–212.

Angrist, J. D. and Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, **114**(2), 533–575.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

Becker, S. O. and Woessmann, L. (2009). Was Weber wrong? A human capital theory of protestant economic history. *The Quarterly Journal of Economics*, **124**(2), 531–596.

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013). From Natural Variation to Optimal

Policy? The Importance of Endogenous Peer Group Formation. *Econometrica*, **81**(3), 855–882.

Checchi, D. and De Paola, M. (2018). The effect of multigrade classes on cognitive and non-cognitive skills. Causal evidence exploiting minimum class size rules in Italy. *Economics of Education Review*, **67**, 235–253.

Chetty, R., Hendren, N., Kline, P., Saez, E., and Turner, N. (2014a). Is the United States still a land of opportunity? Recent trends in intergenerational mobility. *American Economic Review*, **104**(5), 141–147.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, **104**(9), 2633–2679.

Clark, D. and Del Bono, E. (2016). The Long-Run Effects of Attending an Elite School: Evidence from the United Kingdom. *American Economic Journal: Applied Economics*, **8**(1), 150–176.

Coleman, J. S. (1968). Equality of Educational Opportunity Study (EEOS). *Equity & Excellence in Education*, **6**(5), 19–28.

Cortes, K. E. and Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of double-dose Algebra on student achievement. *The American Economic Review*, **104**(5), 400–405.

Cullen, J. B., Jacob, B. A., and Levitt, S. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica*, **74**(5), 1191–1230.

Deming, D. J., Hastings, J. S., Kane, T. J., and Staiger, D. O. (2014). School Choice, School Quality, and Postsecondary Attainment. *The American Economic Review*, **104**(3), 991–1013.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *The American Economic Review*, **101**(5), 1739–1774.

EENEE (2015). The Impact of School Size and Consolidations on Quality and Equity in Education. Technical Report 19, EENEE (European Expert Network on Economics of Education).

Gerhardts, I., Sunde, U., and Zierow, L. (2021). Class Composition and Educational Outcomes: Evidence from the Abolition of Denominational Schools. *mimeo, University of Munich (LMU).*

Glewwe, P., Kremer, M., and Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, **1**(1), 112–135.

Greenwood, J., Guner, N., Kocharkov, G., and Santos, C. (2016). Technology and the Changing Family: A Unified Model of Marriage, Divorce, Educational Attainment, and Married Female Labor-Force Participation. *American Economic Journal: Macroeconomics*, **8**(1), 1–41.

Hanushek, E. a., Kain, J. F., Markman, J. M., and Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, **18**(5), 527–544.

Hattie, J. A. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, **37**(5), 449–481.

Heckman, J., Lochner, L., and Taber, C. (2014). General Equilibrium Effects of Schooling Policy. Technical report.

Iversen, J. M. and Bonesrø nning, H. (2015). Conditional gender peer effects? *Journal of Behavioral and Experimental Economics*, **55**, 19–28.

Jones, S. (2013). Class size versus class composition: What matters for learning in East Africa? *WIDER Working Paper Series*, **065**.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, **114**(2), 497–532.

Lavy, V., Paserman, M. D., and Schlosser, A. (2012). Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom. *Economic Journal*, **122**(559), 208–237.

Leuven, E. and Ronning, M. (2016). Classroom Grade Composition and Pupil Achievement. *Economic Journal*, **126**(593), 1164–1192.

Lichtblau, K. (2009). 50 Jahre Saarland: Wirtschaft Saarland 1959 bis 2009 Wie hat sich das Saarland in den letzten 50 Jahren entwickelt - ein Bundesländervergleich. Technical report, Institut der deutschen Wirtschaft Köln (IW Consult GmbH), Köln.

Lindström, E.-A. and Lindahl, E. (2011). The Effect of MixedAge Classes in Sweden. *Scandinavian Journal of Educational Research*, **55**(2), 121–144.

Little, A. W. (2001). Multigrade teaching: towards an international research and policy agenda. *International Journal of Educational Development*, **21**(6), 481–497.

Lundborg, P., Nilsson, A., and Rooth, D.-O. (2012). Parental Education and Offspring Outcomes : Evidence from the Swedish Compulsory Schooling Reform. *IZA Discussion Paper Series*, **6570**.

Mason, D. A. and Burns, R. B. (1996). 'Simply No Worse and Simply No Better' May Simply Be Wrong: A Critique of Veenman's Conclusion About Multigrade Classes. *Review of Educational Research*, **66**(3), 307–322.

McEwan, P. J. (2008). Evaluating multigrade school reform in Latin America. *Comparative Education*, **44**(4), 465–483.

Mulkeen, A. and Higgings, C. (2009). Multigrade Teaching in Sub-Saharan Africa. *World Bank Working Paper Series*, **173**.

Pischke, J. and Wachter, T. V. (2005). Zero returns to compulsory schooling in Germany: Evidence and interpretation. *The Review of Economics and Statistics*, **90(3)**, 592–598.

Rothstein, J. M. (2006). Good principals or good peers? Parental valuation of school characteristics, tiebout equilibrium, and the incentive effects of competition among jurisdictions. *The American Economic Review*, **96**(4), 1333–1350.

Russell, J. V., Rowe, K. J., and Hill, P. W. (1998). Effects of Multigrade Classes on Student Progress in Literacy and Numeracy: Quantitative Evidence and Perceptions of Teachers and School Leaders. In *Annual Meeting of the Australian Association for Research in Education*, Adelaide.

Sims, D. (2008). A Strategic Response to Class Size Reduction: Combination Classes and Student Achievement in California. *Journal of Policy Analysis and Management*, **27**(3), 457–478.

Stephens, M. and Yang, D.-Y. (2014). Compulsory Education and the Benefits of Schooling. *American Economic Review*, **104**(6), 1777–1792.

Thomas, J. (2012). Combination classes and educational achievement. *Economics of Education Review*, **31**(6), 1058–1066.

Veenman, S. (1995). Cognitive and Noncognitive Effects of Multigrade and Multi-Age Classes: A Best-Evidence Synthesis. *Review of Educational Research*, **65**(4), 319–381.

Vivalt, E. (2015). Heterogeneous Treatment Effects in Impact Evaluation. *American Economic Review*.

Whitmore, D. (2005). Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. *The American Economic Review*, **95**(2), 199–203.

Table A.1: Mixed Grade Levels by Treated and Control Status of Catholic Students

	PRE REFORM				POST REFORM			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	.986	5.571	4.585	(57.118)	.049	.226	.177	(10.436)
Not Mixing	.704	0	-.704	(-49.25)	.977	.834	-.143	(-16.189)
Mixing Two Levels	.1	.032	-.067	(-6.304)	.012	.107	.095	(13.586)
Mixing Three Levels	.048	.045	-.003	(-.296)	.006	.06	.053	(10.011)
Mixing Four Levels	.02	.072	.053	(6.119)	0	0	0	(-.572)
Mixing Five Levels	.027	.105	.078	(7.648)	.001	0	-.001	(-.991)
Mixing Six Levels	.03	.139	.109	(9.623)	.003	0	-.003	(-1.718)
Mixing Seven Levels	.038	.245	.207	(15.104)	0	0	0	(-.572)
Mixing Eight Levels	.025	.244	.218	(16.455)	0	0	0	(.)
Mixing All Levels	.008	.118	.109	(11.312)	0	0	0	(.)
Observations	1216	1021			2667	872		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Catholic students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Table A.2: Mixed Grade Levels by Treated and Control Status of Protestant Students

	PRE REFORM				POST REFORM			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	1.61	5.806	4.196	(29.635)	.087	.109	.023	(1.352)
Not Mixing	.58	0	-.58	(-24.854)	.94	.945	.005	(.571)
Mixing Two Levels	.078	.027	-.051	(-3.347)	.035	.034	-.001	(-.136)
Mixing Three Levels	.059	.038	-.021	(-1.402)	.024	.008	-.016	(-3.087)
Mixing Four Levels	.035	.056	.021	(1.431)	0	.001	.001	(1.673)
Mixing Five Levels	.08	.083	.002	(.124)	0	.003	.003	(2.901)
Mixing Six Levels	.067	.138	.072	(3.339)	.001	.008	.007	(3.513)
Mixing Seven Levels	.07	.252	.183	(7.165)	0	.001	.001	(1.673)
Mixing Eight Levels	.019	.25	.231	(9.919)	0	0	0	(.)
Mixing All Levels	.013	.156	.143	(7.302)	0	0	0	(.)
Observations	374	448			2607	932		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Protestant students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

A Appendix

Notes: A pupil is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational pupils and the schools they attended are considered.

Source: Schools' Index 1964-1986. Own calculations.

Table A.3: Mixed Grade Levels by Treated and Control Status of Non-Denominational Students

	PRE REFORM				POST REFORM			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	1.29	4.259	2.969	(8.423)	.071	.239	.168	(7.613)
Not Mixing	.623	0	-.623	(-9.717)	.95	.881	-.07	(-5.943)
Mixing Two Levels	.058	.121	.063	(1.25)	.031	.064	.033	(3.611)
Mixing Three Levels	.101	.121	.019	(.342)	.018	.031	.013	(1.89)
Mixing Four Levels	.043	.103	.06	(1.31)	0	.002	.002	(2.616)
Mixing Five Levels	.029	.155	.126	(2.563)	0	.007	.007	(4.54)
Mixing Six Levels	.101	.155	.054	(.905)	.001	.013	.012	(4.865)
Mixing Seven Levels	.029	.224	.195	(3.532)	0	.002	.002	(2.616)
Mixing Eight Levels	.014	.121	.106	(2.494)	0	0	0	(.)
Mixing All Levels	0	0	0	(.)	0	0	0	(.)
Observations	69	58			3087	452		

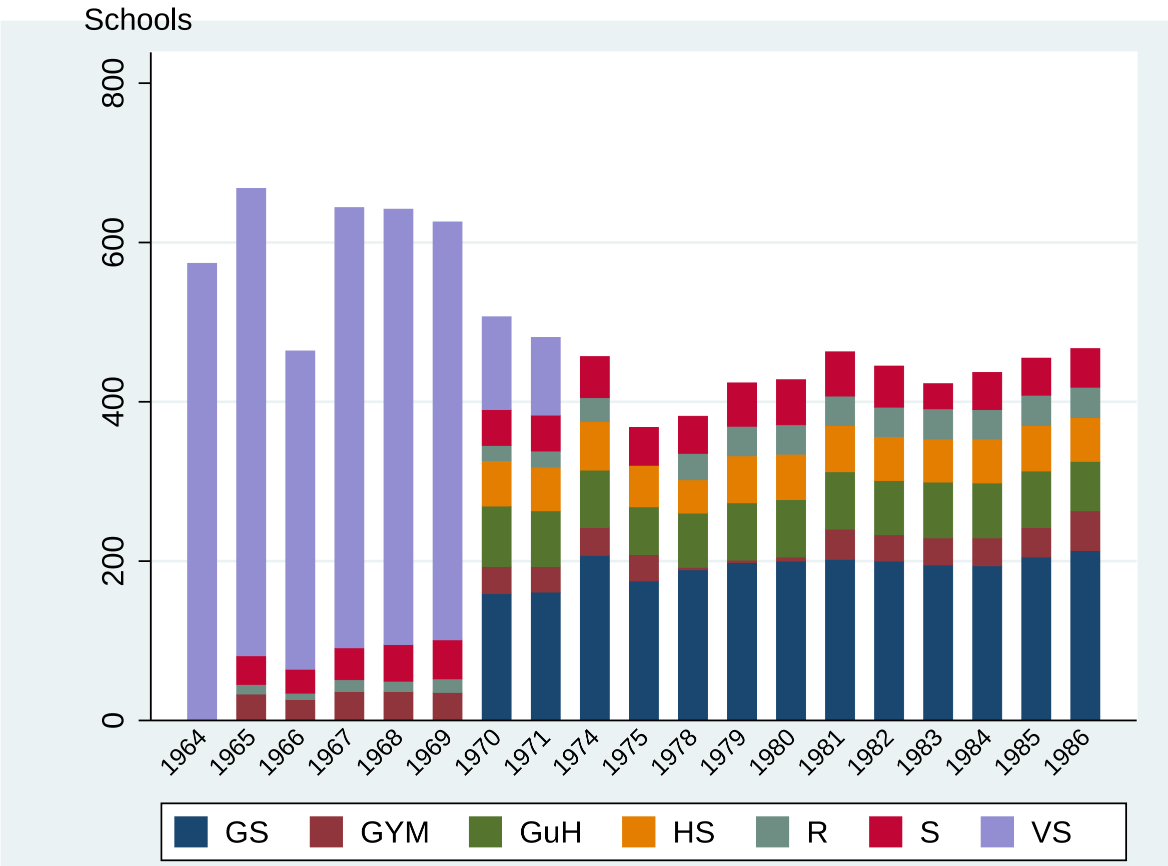
Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Table A.4: School Characteristics by Treated and Control Status of Non-Denominational Students

	PRE REFORM				POST REFORM			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Class Size	31.087	31.397	.31	(.478)	23.017	21.788	-1.23	(-5.216)
Pupils/Teacher	31.043	32	.957	(1.377)	20.312	20.252	-.06	(-.223)
Pupils/School	321.826	149.81	-172.016	(-7.972)	251.168	210.701	-40.467	(-5.127)
Girls' Share	.493	.473	-.02	(-2.031)	.483	.483	0	(.052)
Female Teachers' Share	.574	.437	-.137	(-5.42)	.522	.559	.037	(3.807)
Teachers/School	10.188	4.655	-5.533	(-8.269)	12.539	10.564	-1.975	(-5.042)
Observations	69	58			3087	452		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Figure A.1: Main School Types' Distribution Over Time



Note: Schools' Index 1964-1986 (Own calculations). Records on 1972/73 and 1976/77 are missing completely. In 1964 only the type *Volksschule* (VS) is reported. 1966 about 20% of all types are missing. 1975 there are no records for Realschule (R) and 1978-80 for Gymnasiun (GYM). GS=Grundschule, GuH=Grund- und Hauptschule, HS=Hauptschule, S=Sonderschule.

Table A.5: Effects on Schooling and Labour Market Outcomes – Alternative Sample Restrictions

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
CERTAIN RESIDENCE						
DDD (pupil-based)						
Core Controls	0.106***	0.0284	-0.00849	-0.00316	0.0417	-0.0308
	[0.0342]	[0.0199]	[0.0535]	[0.0337]	[0.0482]	[0.0390]
DDD (school-based)						
Core Controls	0.0660**	0.0284	0.00131	0.0169	0.00293	-0.0153
	[0.0285]	[0.0183]	[0.0422]	[0.0235]	[0.0356]	[0.0293]
DID (school-based)						
Core Controls	0.0940***	0.00913	0.0589**	0.00972	-0.00211	-0.00702
	[0.0289]	[0.0116]	[0.0258]	[0.0306]	[0.0359]	[0.0190]
UNIQUE MAPPING						
DDD (pupil-based)						
Core Controls	0.0840**	0.0251	-0.000134	-0.00807	0.0548*	-0.0439**
	[0.0352]	[0.0221]	[0.0457]	[0.0263]	[0.0314]	[0.0201]
DDD (school-based)						
Core Controls	0.0529*	0.0210	0.0114	-0.00196	0.0413	-0.0355**
	[0.0276]	[0.0184]	[0.0371]	[0.0220]	[0.0258]	[0.0162]
DID (school-based)						
Core Controls	0.0652**	0.0136	0.0155	-0.00340	0.0272	-0.0161
	[0.0259]	[0.0118]	[0.0355]	[0.0280]	[0.0311]	[0.0109]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
CERTAIN RESIDENCE						
N (DDDpupil)	132717	132717	132717	132717	132717	132716
N (DDDschool)	132717	132717	132717	132717	132717	132716
N (DID)	62445	62445	62445	62445	62445	62444
UNIQUE MAPPING						
N (DDDpupil)	125976	125976	125976	125976	125976	125975
N (DDDschool)	125976	125976	125976	125976	125976	125975
N (DID)	48836	48836	48836	48836	48836	48835

Notes: This table shows the results when restricting the sample in two different ways. In the upper part, only those individuals are taken into account for whom the municipality where they went to school is definitely known, i.e. we can exclude migration in order to take up employment elsewhere. In the bottom part, the sample is restricted to those individuals who live in those municipalities in which a unique mapping between individual and school is possible (since there is at maximum one school per denomination prior to the reform). In each part, first estimates of the DDD estimation using the student-based multigrade indicator are shown, then the estimates of the DDD estimation using the school-based multigrade indicator and then the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the estimates of the core controls approach (not including potentially bad controls) are shown.
Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.
Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Table A.6: Effects on Schooling and Labour Market Outcomes – Stratified by Gender

	Schooling		Labor Market			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
BOYS						
DDD (pupil-based)						
Core Controls	0.0712**	0.0467	0.0236	-0.0286	0.0248	0.00828*
	[0.0355]	[0.0285]	[0.0465]	[0.0346]	[0.0328]	[0.00442]
DDD (school-based)						
Core Controls	0.0693**	0.0681**	0.0404	-0.0494	0.0486	0.00624
	[0.0318]	[0.0277]	[0.0361]	[0.0371]	[0.0358]	[0.00433]
DID (school-based)						
Core Controls	0.0554*	0.0135	0.0671*	0.00760	0.00457	-0.00439**
	[0.0326]	[0.0160]	[0.0407]	[0.0346]	[0.0353]	[0.00191]
GIRLS						
DDD (pupil-based)						
Core Controls	0.159***	0.0384	0.00721	0.0365	0.0655	-0.0953***
	[0.0469]	[0.0243]	[0.0689]	[0.0390]	[0.0534]	[0.0343]
DDD (school-based)						
Core Controls	0.116***	0.0411*	0.0304	0.0172	0.0542	-0.0666**
	[0.0390]	[0.0241]	[0.0500]	[0.0269]	[0.0422]	[0.0325]
DID (school-based)						
Core Controls	0.116***	0.00641	0.0179	-0.0202	0.0762*	-0.0378
	[0.0350]	[0.0144]	[0.0511]	[0.0369]	[0.0452]	[0.0232]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
BOYS						
N (DDDpupil)	146633	146633	146633	146633	146633	146631
N (DDDschool)	146633	146633	146633	146633	146633	146631
N (DID)	58042	58042	58042	58042	58042	58040
GIRLS						
N (DDDpupil)	140520	140520	140520	140520	140520	140520
N (DDDschool)	140520	140520	140520	140520	140520	140520
N (DID)	53039	53039	53039	53039	53039	53039

Notes: This table shows the results when stratifying the sample by gender. For each subgroup, first estimates of the DDD estimation using the student-based multigrade indicator are shown, then the estimates of the DDD estimation using the school-based multigrade indicator and then the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the estimates of the core controls approach (not including potentially bad controls) are shown.
Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.
Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Table A.7: Effects on Schooling and Labour Market Outcomes – Stratified by Denomination

	Schooling		Labor Market			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
CATHOLICS						
DDD (pupil-based)						
Core Controls	0.0726**	0.0239	0.0132	0.0158	0.0120	-0.0292
	[0.0360]	[0.0244]	[0.0424]	[0.0277]	[0.0335]	[0.0199]
DDD (school-based)						
Core Controls	0.0658**	0.0354*	0.0328	-0.0000945	0.0192	-0.0173
	[0.0274]	[0.0202]	[0.0330]	[0.0238]	[0.0273]	[0.0164]
DID (school-based)						
Core Controls	0.0800***	0.00828	0.0430	-0.00836	0.0352	-0.0142
	[0.0282]	[0.0126]	[0.0358]	[0.0271]	[0.0304]	[0.0104]
PROTESTANTS						
DDD (pupil-based)						
Core Controls	0.299***	0.0335	-0.216	-0.0993	0.150	0.0131
	[0.0888]	[0.0543]	[0.174]	[0.0818]	[0.103]	[0.0664]
DDD (school-based)						
Core Controls	0.110*	0.0735	-0.0744	-0.108*	0.138**	0.00421
	[0.0588]	[0.0553]	[0.0985]	[0.0617]	[0.0609]	[0.0371]
DID (school-based)						
Core Controls	0.0332	0.0328	0.138	0.0658	-0.0303	-0.0366
	[0.0836]	[0.0386]	[0.114]	[0.0684]	[0.0796]	[0.0390]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
CATHOLICS						
N (DDDpupil)	217373	217373	217373	217373	217373	217371
N (DDDschool)	217373	217373	217373	217373	217373	217371
N (DID)	86288	86288	86288	86288	86288	86286
PROTESTANTS						
N (DDDpupil)	61070	61070	61070	61070	61070	61070
N (DDDschool)	61070	61070	61070	61070	61070	61070
N (DID)	22837	22837	22837	22837	22837	22837

Notes: This table shows the results when stratifying the sample by denomination. For each subgroup, first estimates of the DDD estimation using the student-based multigrade indicator are shown, then the estimates of the DDD estimation using the school-based multigrade indicator and then the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the estimates of the core controls approach (not including potentially bad controls) are shown.
Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.
Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Table 8: Mixed Grade Levels by Treated and Control Status of Protestant Pupils

	PRE REFORM				POST REFORM			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	1.61	5.806	4.196	(29.635)	.087	.109	.023	(1.352)
Not Mixing	.58	0	-.58	(-24.854)	.94	.945	.005	(.571)
Mixing Two Levels	.078	.027	-.051	(-3.347)	.035	.034	-.001	(-.136)
Mixing Three Levels	.059	.038	-.021	(-1.402)	.024	.008	-.016	(-3.087)
Mixing Four Levels	.035	.056	.021	(1.431)	0	.001	.001	(1.673)
Mixing Five Levels	.08	.083	.002	(.124)	0	.003	.003	(2.901)
Mixing Six Levels	.067	.138	.072	(3.339)	.001	.008	.007	(3.513)
Mixing Seven Levels	.07	.252	.183	(7.165)	0	.001	.001	(1.673)
Mixing Eight Levels	.019	.25	.231	(9.919)	0	0	0	(.)
Mixing All Levels	.013	.156	.143	(7.302)	0	0	0	(.)
Observations	374	448			2607	932		

Notes: A pupil is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational pupils and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Table 9: School Characteristics by Treated and Control Status of Non-Denominational Pupils

	PRE REFORM				POST REFORM			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	1.29	4.259	2.969	(8.423)	.071	.239	.168	(7.613)
Not Mixing	.623	0	-.623	(-9.717)	.95	.881	-.07	(-5.943)
Mixing Two Levels	.058	.121	.063	(1.25)	.031	.064	.033	(3.611)
Mixing Three Levels	.101	.121	.019	(.342)	.018	.031	.013	(1.89)
Mixing Four Levels	.043	.103	.06	(1.31)	0	.002	.002	(2.616)
Mixing Five Levels	.029	.155	.126	(2.563)	0	.007	.007	(4.54)
Mixing Six Levels	.101	.155	.054	(.905)	.001	.013	.012	(4.865)
Mixing Seven Levels	.029	.224	.195	(3.532)	0	.002	.002	(2.616)
Mixing Eight Levels	.014	.121	.106	(2.494)	0	0	0	(.)
Mixing All Levels	0	0	0	(.)	0	0	0	(.)
Observations	69	58			3087	452		

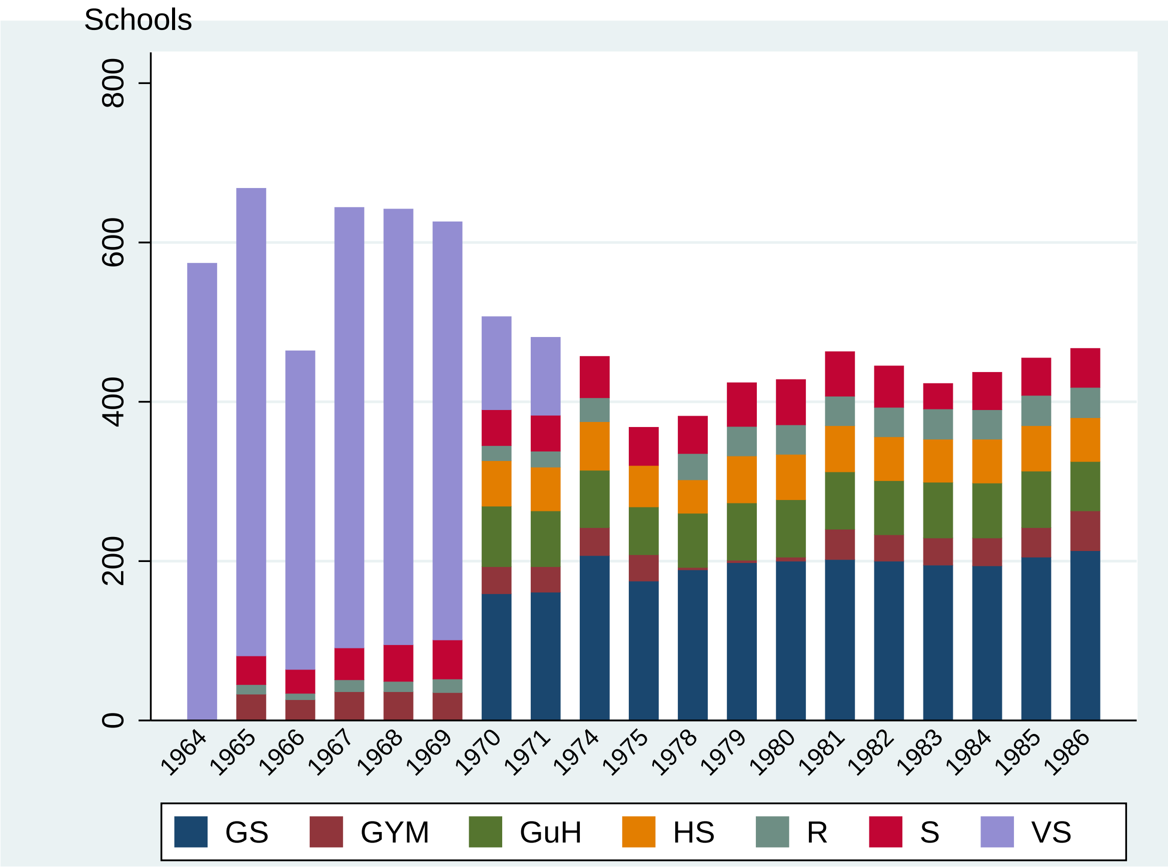
Notes: A pupil is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational pupils and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Table 10: Characteristics of Non-Denominational Schools in Treated and Control Municipalities

	PRE REFORM				POST REFORM			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Class Size	31.087	31.397	.31	(.478)	23.017	21.788	-1.23	(-5.216)
Pupils/Teacher	31.043	32	.957	(1.377)	20.312	20.252	-.06	(-.223)
Pupils/School	321.826	149.81	-172.016	(-7.972)	251.168	210.701	-40.467	(-5.127)
Girls' Share	.493	.473	-.02	(-2.031)	.483	.483	0	(.052)
Female Teachers' Share	.574	.437	-.137	(-5.42)	.522	.559	.037	(3.807)
Teachers/School	10.188	4.655	-5.533	(-8.269)	12.539	10.564	-1.975	(-5.042)
Observations	69	58			3087	452		

Notes: A pupil is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational pupils and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Figure 2: Main School Types' Distribution Over Time



Note: *Schools' Index 1964-1986 (Own calculations). Records on 1972/73 and 1976/77 are missing completely. In 1964 only the type Volksschule (VS) is reported. 1966 about 20% of all types are missing. 1975 there are no records for Realschule (R) and 1978-80 for Gymnasiun (G). GS=Grundschule, GuH=Grund- und Hauptschule, HS=Hauptschule, S=Sonderschule.*

Table 11: Effects on Schooling and Labour Market Outcomes – Alternative Sample Restrictions

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant LM
CERTAIN RESIDENCE						
DDD (pupil-based)						
Core Controls	0.106***	0.0284	-0.00849	-0.00316	0.0417	-0.0308
	[0.0342]	[0.0199]	[0.0535]	[0.0337]	[0.0482]	[0.0390]
DDD (school-based)						
Core Controls	0.0660**	0.0284	0.00131	0.0169	0.00293	-0.0153
	[0.0285]	[0.0183]	[0.0422]	[0.0235]	[0.0356]	[0.0293]
DID (school-based)						
Core Controls	0.0940***	0.00913	0.0589**	0.00972	-0.00211	-0.00702
	[0.0289]	[0.0116]	[0.0258]	[0.0306]	[0.0359]	[0.0190]
UNIQUE MAPPING						
DDD (pupil-based)						
Core Controls	0.0840**	0.0251	-0.000134	-0.00807	0.0548*	-0.0439**
	[0.0352]	[0.0221]	[0.0457]	[0.0263]	[0.0314]	[0.0201]
DDD (school-based)						
Core Controls	0.0529*	0.0210	0.0114	-0.00196	0.0413	-0.0355**
	[0.0276]	[0.0184]	[0.0371]	[0.0220]	[0.0258]	[0.0162]
DID (school-based)						
Core Controls	0.0652**	0.0136	0.0155	-0.00340	0.0272	-0.0161
	[0.0259]	[0.0118]	[0.0355]	[0.0280]	[0.0311]	[0.0109]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
CERTAIN RESIDENCE						
N (DDDpupil)	132717	132717	132717	132717	132717	132716
N (DDDschool)	132717	132717	132717	132717	132717	132716
N (DID)	62445	62445	62445	62445	62445	62444
UNIQUE MAPPING						
N (DDDpupil)	125976	125976	125976	125976	125976	125975
N (DDDschool)	125976	125976	125976	125976	125976	125975
N (DID)	48836	48836	48836	48836	48836	48835

Notes: Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.

Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Table 12: Effects on Schooling and Labour Market Outcomes – Stratified by Gender

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant LM
BOYS						
DDD (pupil-based)						
Core Controls	0.0712**	0.0467	0.0236	-0.0286	0.0248	0.00828*
	[0.0355]	[0.0285]	[0.0465]	[0.0346]	[0.0328]	[0.00442]
DDD (school-based)						
Core Controls	0.0693**	0.0681**	0.0404	-0.0494	0.0486	0.00624
	[0.0318]	[0.0277]	[0.0361]	[0.0371]	[0.0358]	[0.00433]
DID (school-based)						
Core Controls	0.0554*	0.0135	0.0671*	0.00760	0.00457	-0.00439**
	[0.0326]	[0.0160]	[0.0407]	[0.0346]	[0.0353]	[0.00191]
GIRLS						
DDD (pupil-based)						
Core Controls	0.159***	0.0384	0.00721	0.0365	0.0655	-0.0953***
	[0.0469]	[0.0243]	[0.0689]	[0.0390]	[0.0534]	[0.0343]
DDD (school-based)						
Core Controls	0.116***	0.0411*	0.0304	0.0172	0.0542	-0.0666**
	[0.0390]	[0.0241]	[0.0500]	[0.0269]	[0.0422]	[0.0325]
DID (school-based)						
Core Controls	0.116***	0.00641	0.0179	-0.0202	0.0762*	-0.0378
	[0.0350]	[0.0144]	[0.0511]	[0.0369]	[0.0452]	[0.0232]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
BOYS						
N (DDDpupil)	146633	146633	146633	146633	146633	146631
N (DDDschool)	146633	146633	146633	146633	146633	146631
N (DID)	58042	58042	58042	58042	58042	58040
GIRLS						
N (DDDpupil)	140520	140520	140520	140520	140520	140520
N (DDDschool)	140520	140520	140520	140520	140520	140520
N (DID)	53039	53039	53039	53039	53039	53039

Notes: Standard errors are clustered at the municipality level and are shown in parenthesis: *p $<$ 0.10, ** p$<$0.05, ***p$<$0.010.

Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Table 13: Effects on Schooling and Labour Market Outcomes – Stratified by Denomination

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant LM
CATHOLICS						
DDD (pupil-based)						
Core Controls	0.0726**	0.0239	0.0132	0.0158	0.0120	-0.0292
	[0.0360]	[0.0244]	[0.0424]	[0.0277]	[0.0335]	[0.0199]
DDD (school-based)						
Core Controls	0.0658**	0.0354*	0.0328	-0.0000945	0.0192	-0.0173
	[0.0274]	[0.0202]	[0.0330]	[0.0238]	[0.0273]	[0.0164]
DID (school-based)						
Core Controls	0.0800***	0.00828	0.0430	-0.00836	0.0352	-0.0142
	[0.0282]	[0.0126]	[0.0358]	[0.0271]	[0.0304]	[0.0104]
PROTESTANTS						
DDD (pupil-based)						
Core Controls	0.299***	0.0335	-0.216	-0.0993	0.150	0.0131
	[0.0888]	[0.0543]	[0.174]	[0.0818]	[0.103]	[0.0664]
DDD (school-based)						
Core Controls	0.110*	0.0735	-0.0744	-0.108*	0.138**	0.00421
	[0.0588]	[0.0553]	[0.0985]	[0.0617]	[0.0609]	[0.0371]
DID (school-based)						
Core Controls	0.0332	0.0328	0.138	0.0658	-0.0303	-0.0366
	[0.0836]	[0.0386]	[0.114]	[0.0684]	[0.0796]	[0.0390]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
CATHOLICS						
N (DDDpupil)	217373	217373	217373	217373	217373	217371
N (DDDschool)	217373	217373	217373	217373	217373	217371
N (DID)	86288	86288	86288	86288	86288	86286
PROTESTANTS						
N (DDDpupil)	61070	61070	61070	61070	61070	61070
N (DDDschool)	61070	61070	61070	61070	61070	61070
N (DID)	22837	22837	22837	22837	22837	22837

Notes: Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.

Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.