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

Ilka Gerhardts, University of Munich *
Uwe Sunde, University of Munich
Larissa Zierow, University of Munich

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, labor market participation and socioeconomic mobility. Notably, the negative impact is most pronounced in the outcomes of girls. Disentangling the confounding role of variation between Catholic and Protestant schools suggests that this effect might be driven by socialization early in life.

This is a preliminary draft. Please do not cite or distribute without permission of the authors.

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

1 Introduction

Many schools are operated by religious denominations on a basis of multigrade classes, which represent the typical way of teaching children in the context of limited resources. Such multigrade classes are a cost-effective way of providing children with education and 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).

In the face of the dramatic demographic change, multigrade classes have recently become a principal adjustment device for enrollment fluctuations also in many parts of Europe. However, 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, denominational affiliation has lost importance, and demographic change led to the abolition of denominational schools. Mixed empirical evidence regarding the effects of abolishing denominational schools with multigrade classes on subsequent outcomes fuels heated debates regarding the appropriate education policies.

This paper investigates the impact of denominational schools on the returns to education. The identification strategy exploits the natural experiment of the abolition of parochial 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 rural students, who were predominantly exposed to multigrade teaching before 1969, but not afterwards, compared to urban students, who were affected much less intensely by the abolition of denominational schools. The estimation uses an enhanced differences-in-differences approach to estimate the effects of the reform on schooling and labor market outcomes.

By exploring the heterogeneity of the effects across genders, 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, occupational choice and financial dependencies 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 size classes, 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. The availability of a wide range of schooling covariates allows us to control for channels like class size, school size, school consolidation, Catholic education style, gender composition, etc. that might confound the multigrade effects.

The empirical results suggest that multigrade classes have detrimental effects on final grade attainment and labor market participation. The effect is notably stronger for girls and pupils with less educated parents, findings that persist after controlling for potential mediators like marriage patterns. The latter seems to be particularly the case among Catholic children, not among Protestant children. The results therefore suggest an interplay of socialization, potentially based on religious denomination and stereotypes, 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. By contrast, those 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. 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, occupational choice and financial dependencies. 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 and discusses robustness with respect to sensitivity checks. 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 Bonesrønning (2015) and 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 Johansson and Lindahl (2008) 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-pupil 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 pupils *if* the pedagogical infrastructure 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

[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.

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 pupils. 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

A rather recent study on combination classes is the one by Johansson and Lindahl (2008). 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 Rønning (2011). 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 Rønning (2011) 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 Rønning (2011) 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 pupils to cut less able ones out of their circle leaving them with even worse academic attainments.

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 Johansson and Lindahl (2008) and Leuven and Rønning (2011) 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

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.

[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 Jacob et al. (2010) 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.

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 institutional 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 Johansson and Lindahl (2008)). 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, occupational choice and financial dependencies. 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. Finally, the institutional setting we study allows us to disentangle the impact of multigrade schools by gender and by denomination. This enables us to disentangle gender effects from socialization effects which has - up to now - never been explored in this context yet.

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, close to all *Volksschulen* sorted pupils by denomination. This allocative restriction created multigrade classes in low-density regions[9]. With few exceptions parochial 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]

[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 Bonesrønning (2015) explore spillovers in Norwegian elementary schools where special education happens to be intergrated 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 5). Thus the insights by Iversen and

Schools providing primary or lower secondary education were uniformly labeled *Volksschule*. Obliged to teach Catholics and Protestants separately[11] 75% of rural schools resolved to a multigrade structure shown in Table 11. By contrast church-maintained urban schools adhered to a single-grade structure. Taking into account that Saarlanders were predominantly Catholic and even more so in the villages there were more Catholic multigrade pupils. However the few rural Protestant schools collapsed more grade levels per class. Pupils' age distribution prior to the reform was right-skewed. Given schools target balanced class sizes higher grade levels were more likely to be mixed and multigrade peer effects were observed for teenagers rather than children. Another implication of denominational schools was that they rejected basically all non-Christian migrants. This made them popular with parents who favored German classmates for their offspring. With less than 5% foreign households residing in the Saarland this should not have been crucial. Finally staff selection took into account a teacher's denomination.

The reform directly impacted schools in basic education. Inducing a change in pupils' distribution across school types it also indirectly affected higher education though. When parochial 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. It shrank prevalence of multigrade schools by two thirds in less than a year and from 1974 onwards the share was negligible. Learning environment for village children changed 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 urban schools only moderately (Figure 2). This 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 pupils were taught in much more homogeneous classes. By construction the reform reshaped the educational infrastructure in multiple ways and also implied more pupils and more teachers per school in absolute terms (EENEE 2015). At first sight surprisingly, average class size shrank because the inflow of remote area children into urban school districts was mitigated by a demographic decline in enrollment. It drastically reduced overall class size from 39 (1964) to 19 (1986) pupils on average, but the relative change was identical for urban and rural regions. However commuting pupils coming from remote areas might have encountered higher quality peers

Bonesrønning (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).

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

from denser municipalities (Leuven and Rønning 2011). Overall school covariates did change in absolute terms, but looking at important ratios like the pupil-teacher ratio and the shares of girls and female teachers the changes seem modest.

All key features of schools are summarized in Table 13, partitioning the universe of *Volksschulen* into four groups, namely rural and urban schools, each before and after 1969. For this comparison to make sense a common trend between rural and urban 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 rural and urban 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. [13]

Diminishing age diversity and academic redshirting matters in overcoming grade retention problems (Faust 2006; Leuven and Rønning 2011).
Next *Compulsory Schooling Duration* requires that pupils 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 pupils 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. Angrist and Pischke (2008) show that the short school years cause a peak in grade repeating rates and thus promotes negative peer effects. Reassuringly a common trend in class size (Table 12) suggests that repetition rates spike similarly across regions. As skills and labor market opportunities are supposed to weakly increase in years of schooling (Angrist and Pischke 2008) the law precludes early dropouts, e.g. due to financially constrained parents requiring their children to support the family instead of attending class.
Then, *No Tuition Fees* guarantee basic education to be free of charge, independent of the school's being state- or church-maintained.[14] It limits yet again 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 pupils over schools based on catchment areas.[15] To choose a certain *Volksschule* by its reputation would require the household to move into that school's catchment area. However parents not teachers choose the type of school their child attends after

[13] §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).

[14] §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).

[15] §29 Satz 2 Schulordnungsgesetz vom 5. Mai 1965.

completing primary school because a teacher's recommendation for admission to one of the higher tracks is not yet required.[16] As not all types are provided within each catchment area there remains some freedom regarding school choice. Rothstein (2006) investigates parental preferences over school choice and establishes that peer groups matter even more than school's effectiveness. This underlines the importance of pupil 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. So from the Saarland's reintegration into Germany onwards we screen the most influential historical events for asymmetric impacts on rural and urban regions. A primary concern arises by fluctuations in economic activity centered in urban regions. The coal and steel crises depress the Saarland even more than the rest of Germany (Lichtblau 2009). They cause dramatic peaks in unemployment and overshadow 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. It is worth mentioning that inspite of those shocks the Saarland was politically nearly perfectly stable (Lichtblau 2009). Only the very last year of our study's time horizon is subjected to a different government, so 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 pupils in a multigrade environment to a single-grade enviroment, which is presumably less heterogeneous. We tackle this question estimating a triple differences (DDD) model that exploits exogeneous variation in the probability to be a multigrade pupil over time, region and age group.

Let Y_{1ircy} represent individual i's outcome in region r, cohort c and age group y if she attended a multigrade school and Y_{0ircy} otherwise. As a baseline, we estimate the 'reform effect' in a parsimonious regression with main effects $r \in \{Rural, Urban\}$, $c \in \{Pre, Post(Reform)\}$ and $y \in \{Young, Old\}$. In an additional analysis, we add

[16] Zeugnis- und Versetzungsordnung der Grundschulen im Saarland vom 01. Juni 1968 (GMBl. 1968 Seite 177).

a triple interaction, reflecting a DDD estimator,

$$\begin{aligned} Y_{ircy} &= \beta_0 + \beta_1 Rural_r + \beta_2 Post_c + \beta_3 Young_y \\ &+ \beta_{12} Rural_r Post_c + \beta_{13} Rural_r Young_y + \beta_{23} Post_c Young_r \\ &+ \beta \underbrace{Rural_r Post_c Young_y}_{D_{rcy}} + \varepsilon_{ircy} \end{aligned} \tag{1}$$

Identification is based on the contrasts across age groups, community types, and treatment status. We estimate the DDD baseline reform effect including just the main effects *Rural, Post, Young* and their interaction terms. *Rural* switches on for municipalities < 10,000 inhabitants. This definition is relevant for balancing tables (available upon request) and refined upon regressing by using a set of size class fixed effects instead. *Post* is 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 32 year-olds and above.

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, German* and *White-Collar Breadwinner*. *Young at School Entry* relates birth month and school entry cutoff date to indicate if a pupil is relatively young within her cohort. *White-Collar Breadwinner* refers to the socioeconomic status of a household's head serving to proxy parent education.[17] Combining this with administrative data from school records allows us to include additional controls. These include size-class level regressors *Class Size, School Size* (defined as the number of pupils) *, Girls' Share, Female Teachers' Share, Teachers Without Abitur, Catholic Schools' Share*. We are still waiting for regression output that accounts for *Potential Commuting Costs* which we define as the average distance to the nearest *Realschule* or *Gymnasium*.[18]

Estimation is conducted by OLS, estimating triple differences (DDD). The estimation sample excludes observations younger than fifteen or aged between 21 and 31 years. Results are robust to dropping observations whose residence of schooling is unknown. OLS standard errors are reported in parentheses.[19]

[17] In additional unreported robustness checks, we further include *Household Size, Single*, which might be endogenous and thus constitute bad controls. Nevertheless, in general the results are very similar in sign and significance albeit the size of the estimates may fluctuate somewhat. Results are available upon request.

[18] Further robustness checks are conducted on a continuous multigrade indicator $MissG \in \{-8, 0\}$ that denotes the difference between actual and required grade levels per school, averaged at the size class level. Detailed results are available upon request.

[19] Given just nine size classes, raw standard errors happen to be relatively larger than robust or clustered ones. Future work will construct bootstrapped clustered standard errors, but these are not available yet.

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.[20] 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 rural neighborhood quality induced by state-level interventions to counteract drift to the cities. The setup still requires unobservable asymmetries in teaching effectiveness and ability differences between rural and urban pupils to be time-constant, because with two periods region-specific outcome trends are not identified, a drawback detailed in Stephens and Yang (2014). Moreover we rely on the aforementioned pupil allocation via catchment areas to ensure that pupils do not choose their school, and thus their multigrade exposure. To sum up, for multidimensional 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. Given that multigrade studies are peer group studies the impact need not be uniformly signed. Reassuringly, in our subgroup analyses the multigrade impact ranges from strongly to negligibly negative, but never positive.

Strictly speaking any school offering fewer classes than grade levels required by the degree it bestows is a multigrade school. A natural parameterization of municipality-level multigrade intensity[21] consists in

$$MissG_{rc} = \frac{1}{S_{rc}} \sum_{s=1}^{S_{rc}} ActualClasses_{src} - GradeLevels_{src} \leq 0. \tag{2}$$

We use this indicator in two ways. For the treatment-control scenario above we define municipalities < 10,000 inhabitants as rural, after observing the multigrade indicator and population density to be highly negatively correlated. The drawback of creating a treatment indicator D_{rcy} out of dummy interactions is that prior to the reform there have been some, much less extreme multigrade schools also in urban municipalities. Therefore urban municipalities fail as a control group in the strictest sense.

As a second approach we regress outcomes directly on $MissG_{rc}$ which is more accurate in terms of treatment intensity (see Acemoglu2004) but not suitable upon running balancing tests. For group comparisons some aggregation over the continuity of treatment groups is required. In the continuous case β measures the impact of collapsing one grade level less, i.e. being taught jointly with two instead of three levels. Apart from that let R_r denote the size class a municipality

[20] *CIA:* $\{Y_{0ircy}, Y_{1ircy}\} \perp\!\!\!\perp D_{rcy} \implies E[Y_{0ircy}|r,c,y,\mathbf{x_{ircy}},D_{rcy}] = E[Y_{0ircy}|r,c,y,\mathbf{x_{ircy}}]$

[21] For example, a small *Volksschule* with five classes is assigned a value of $5-9=-4$, while a large primary school with ten classes gets a zero as $10-4>0$.

is in with $r \in \{1,...,R\}$. This set of size class fixed effects captures region-specific time-constant unobservables more precisely. Instead of time fixed effects in form of cohort dummies, we stick with the binary post-period indicator but control for age in addition. We include further predetermined individual characteristics as well as a wide range of schooling covariates represented by the $k \times 1$ vector $\mathbf{x_{ircy}}$. Jointly these modifications give rise to the enhanced regression

$$\begin{aligned} Y_{ircy} = \beta_0 + \sum_{r=2}^{R} \beta_{1r} R_r + \beta_2 Post_c + \beta_3 Young_y \\ + \beta_{12} MissG_{rc} + \sum_{r=2}^{R} \beta_{13r} R_r Young_y + \beta_{23} Post_c Young_r \\ + \beta \underbrace{MissG_{rc} Young_y}_{D_{rcy}} + \mathbf{x'_{ircy}} \delta + \varepsilon_{ircy}. \qquad (3) \end{aligned}$$

5 Data

This section describes the data. Via municipality size classes we combine two population and one schools' survey, all of which are comprehensive, high-quality administrative datasets.[22]

Outcomes[23]

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 separate dummies as well as a combined, normalized index comprising Volksschulreife (Hauptschulabschluss), Mittlere Reife (Realschulabschluss) and Fach-/ Abitur (Gymnasialabschluss), assigned values of one to three imposing cardinality. 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 testscores 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 Rønning (2011). Above all peer effects trigger social competences not captured by test scores but

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

[23] 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).

perhaps reflected in post-schooling attainment.

Next we use labor market outcomes to assess lasting or reemerging effects of schooling similar to Chetty, Friedman, et al. (2014). To analyze labor market participation we use binary indicators on unemployment, housewife status and financial dependence, a global measure subsuming family support, unemployment benefits and social aids. Given labor market entry we distinguish further between blue- and white-collar occupations to capture socioeconomic status. Note that wages are not reported in the Census 1987[24].

Treatment Indicator

We determine each individual's likelihood for having been a multigrade pupil computing the indicator $MissG_{rc}$. Using data from Saarland's Land 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 pupils and teachers, the number of classes, school's type, denomination and address are given for each school on an annual basis yielding more than 7,500 school-year observations.[25] The school's adress enables us to average over schooling conditions in a given municipality[26] and via the municipality size class match these averages back to individual-level census data.

Controls

From the schools' records we compute pre- and post-reform municipality size class averages of class size, pupil-teacher ratio, school size (in terms of number of pupils), girls' share and female teachers' share. Class size, the principal rivaling input shows observational equivalence and a similar trend. This is not the case for less crucial confounders. The share of female pupils is much lower prereform, but is equilibrated postreform (Table 13).

Next census data provide a set of individual-level controls, most of which are commonly used and self-explanatory (Table 1). The established differences are in line with expectations depicting more Catholics, fewer migrants and larger households with lower educated income earners and teachers in rural regions. Here we briefly discuss those controls with non-standard implications. To begin with we recognize denomination as being essential in disentangling a 'parochial school effect' from the multigrade effect we are after. In the 1960s the concept of 'the Catholic girl educated to become a housewife' seems to be still prominent

[24] 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, Hendren, et al. (2014)

[25] 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.

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

in conservative villages shaping a life course perhaps rooted more in parent and teacher attitudes than in peer effects in class. We proxy the influence of Catholic teaching methods by the share of Catholic schools per municipality. Another issue is that 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. Pupils 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*.

Concerned with convergence in rural and urban teachers' professional formation we proxy teacher quality. Census information provides the highest degree of Saarland's teaching population employed in lower basic education. Again we impose a lower bound of 32 years to rule out that teachers themselves have been partially treated.

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[27]. We drop individuals younger than fifteen 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[28] such that their mobility is low. This matters because census data provide the municipality size class of current residence and not the size class of school attendance. Fortunately the residence-of-household definition ties children to their parents' address until they begin their own household. Nevertheless concerned with rural individuals moving reform-induced to urban regions we impose that underage restriction. It leaves us with a sample of main interest of five consecutive birth cohorts that are between fifteen and twenty years old in either census. All of them attend primary and lower secondary school either strictly before or after the reform takes place.

Second, although there is no panel structure at the individual level, observations of

[27] At the moment we still use a 10% subsample of the 1970 Census but the full sample will be available in March 2016 at the latest.

[28] 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.

the 1970 Census reappear in the survey of 1987. People older than 32 years 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. However the case is much more complicated for individuals younger than 32 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. 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.

6 Results

This section presents estimates of the multigrade impact 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[29] are robust to the inclusion of a wide range of individual characteristics and schooling covariates. Moreover we subject our estimates to rigorous subsample analyses. Throughout, results are put into perspective via baseline estimates of the pure reform effect.

Overall Results

Schooling Outcomes DID regressions suggest the abolition of parochial schools to favorably influence degree attainment (Table 2). A natural explanation could be that individuals spend more time on schooling because single-grade classes improve basic training. This in turn makes superior educational attainment accessible. First these gains are reflected by a sizable and robust 8 percentage points reduction in the probability to earn at most a *Volksschulabschluss*. Indeed the estimate is larger than the pre-reform attainment gap measured in the share of rural (72%) and urban (66%) lowest-educated teenagers as shown in Table 3. Second, Table 2 shows that the probability to obtain an intermediate-level degree rises due to the reform. The predicted change is about half the size of the decline of the lowest degree attainment and marginally significant. Estimated impacts on the probability to obtain the highest degree are insignificant and practically zero. Finally the multi-level degree index is up by 0.08 (se 0.04) of the degree distribution's standard

[29] As mentioned before, our analysis is confined to remote execution. As DDD results for the continuous multigrade indicator are not yet available, we discuss DID results here instead.

deviation which is normalized to 1 (see first column in Table 2). It confirms overall schooling attainment to rise notably due to the reform although much higher standard errors render it insignificant once core controls are included.
Exchanging the binary treatment indicator for a continuous one seems to slightly heighten precision (as shown in bottom part of Table 2). And it alters the outlined patterns in predicting a significantly negative effect on the likelihood to obtain a university entrance qualification. Recall that treatment intensity is measured as the saldo of pre- and post-reform mixed grade levels prevalent in a certain size class. Accordingly switching from e.g. three- instead of two-level to a single-grade class *decreases* probability to obtain an *Abitur* by 2.3 percentage points. The paradoxon lies with overall evidence suggesting multigrade classes to be detrimental.

Running a set of placebo baseline regressions (Table 4) proposes a significant shift from low to intermediate degree attainment also for untreated people. This is alarming insofar that it indicates the presence of region-specific reform-unrelated systematic outcome dynamics that affect both old and young people significantly. Note that estimates' signs switch and/or significance dissolve once core controls are included.
This motivates a triple differences model as well as it highlights that core covariates are doing a good job. For the combined degree index DDD regression predicts a significantly positive and stable estimate of about 0.11 of a standard deviation (as shown in upper part of Table 2). Estimating the impact on the basic lower secondary attainment predicts a likewise stable and significant decrease. Order of magnitude is very similar to the DID model. More intriguing is perhaps that the impacts beyond the basic track dissolve to zero as well as they loose significance. This suggests the negative impact associated with an *Abitur* degree observed in the DID model to be at least not schooling-specific.

Viewing this evidence jointly we conclude that single-grade classes do invoke a beneficial effect on educational attainment but its longevity seems debatable. Which mechanism offsets these gains convexly along the degree scale? A natural starting point for interpretation are changed *Potential Commuting Costs* that are not yet controlled for. Remote regions show a very small number of ability tracking schools. Eradicating their tiny village schools leads to reallocation of rural pupils to the next larger rural school which in most cases provides lower secondary education as well. The multigrade schools of peripheral pupils are wiped out likewise, however their alternatives differ. Both groups are forced to commute more but only in the foothills of urban municipalities this equalizes costs of attending any of the educational tracks. Ability tracking normally begins after primary school. Even those rural pupils striving for higher degrees tend to stay at a lower secondary school during compulsory schooling to postpone excessive commuting. Nevertheless to do so creates an additional barrier because as Pischke and Wachter (2005) states adjustment costs are so high that upward migration only occurs by exception. If at all switching into the intermediate track seems more feasible than jumping up into the top track. Apart from that admission barrier high-track schools are even more centered in the cities. Also the treatment groups' relative larger increment in potential costs needs to be multiplied by a duration of four instead of one additional year of schooling. Upon deciding on further schooling these additional commuting costs are traded off against benefits of better preparation by single-grade education received so far. Moreover forced commuting during compulsory schooling (*de facto* nine years only post-treatment) might fastidiate the rural post-cohort more whenever costs are convex. Jointly this could explain decreasing beneficial effects in spite of higher treatment gains due to even more increasing *Potential Commuting Costs*. Besides, the implementation of the *Berufsbildungsgesetz (BBiG)* in 1969 boosts reputation of the dual vocational education and training.[30] We need to look further into intermediate education opportunities possibly rivaling higher education more than before. Abramitzky and Lavy (2011) argue similarly with competing educational paths.

Professional Outcomes The DID model (continuous treatment indicator, significant point estimate of 0.036) implies that sparing pupils from studying jointly with one more grade level translates into a 3.6 percentage points higher employment probability before turning 21 (Table 2). Findings change only negligibly upon using a binary indicator or by adding the control group of elder people. If this was a demographic trend only the predicted rise in employment is impressive insofar as it occurs inspite of delayed labor market entry by prolonged schooling. So chances on the labor market must have improved sizably which could be explained as follows. Better education produces more skilled rural junior workers. Ceteris paribus on the rural labor market they crowd-out those from the older generation. Next treated workers are also expected to be more mobile because they are more competitive and because they are used to commute. This opens up urban labor markets on top. The global gain in employment is partly driven by female labor market participation. It is reflected in housewife status declining by 2 percentage points, a highly significant point estimate that amounts to half of the full sample's pre-treatment mean. Channels of gender-specific responsiveness to treatment are detailed below. Further labor market outcomes suggest single-grade schooling to invoke a shift away from blue- to white-collar occupations. Running a DID specification for the overall sample of young adults these estimates are sized about 2 percentage points but lack precision. Intriguingly using a DDD model instead the probability to choose a blue-collar occupation falls drastically to -20%. The estimate is significant at the 1% level. This is the most spectacular

[30] http://www.bpb.de/themen/5Q9PD4,0,Die_duale_Ausbildung.html (24 August 2015). Not listed in Section 3 because the law does not directly address basic education.

example of how misleading the baseline specification can be. It suggests an insignificant coefficient of +1%. The predicted shrink fits the picture because neither an increased share of female workers nor enhanced academic attainment suggest self-selection into blue-collar professions. Meanwhile the DDD model's predicted impact on white-collar jobs turns negative albeit it stays insignificant. At first sight surprisingly the effects are reconciled by extended schooling duration. White-collar employment on average requires longer educational preparation. Thus the respective future employees might simply enter the labor market at a more advanced age than that monitored in our study. So longer investment in schooling leads up to white-collar professions and simultaneously complicates capturing this effect at such an early age. Prior to treatment urban white-collar employment exceeds its rural counterpart by 6 percentage points.

Financial Outcomes DID regressions robustly predict the share of people depending on any type of financial support to shrink about 3.5 percentage points (see right part of Table 2). The estimate is marginally significant and mirrors the increased employment probability that makes it more likely for the treated cohort to earn its own living. The global measure *Supported* pools recipients of unemployment benefits, social aid, etc. but also students maintained by their parents. We dispose of insufficient observations to evaluate unemployeds' developments separately. By contrast the dependence relation between parents and children occurs frequently enough. It yields a very similar decrease, in size and significance. This highlights yet again how strong the employment effect seems to be. It completely offsets that an individual's prolonged schooling postpones the period she earns her own living. Predictions based on binary and continuous treatment indicators are within a range of 0.2 percentage points of each other. Even more precisely measured DDD regression estimates back them up. Prior to treatment both dependency indicators are completely balanced.

Sensitivity Checks

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

15-17 Year-Olds & Adults > 31 Years of Age The need to check on a younger-than-eighteen subsample arises from the lowered threshold to adulthood faced by post-reform cohorts. As outlined in Section 5 it is much harder to determine for legally grown-ups if they are treatment or control observations. Concern is centered on selective migration of more able rural graduates to urban labor markets. Results generated by the restricted age group alone are less error-prone to moving. Therefore the restricted estimates (available upon request) serve as a benchmark whenever outcomes are relevant in both age ranges. Albeit for the subsample slightly less precisely measured, estimates are similar by sign and significance. The point estimate of the probability to obtain more than a *Volksschulabschluss* rises by 2 percentage points, the largest difference by order of magnitude. And it backs up the suspicion that some highly able treatment observations are mistaken for control ones in the full sample analysis. Then beneficial effects of single-grade classrooms are necessarily downward biased. Estimates on labor market and financial outcomes track full sample results more closely. This is intriguing because employment effects could be expected to be less pronounced observing so young an age group. However it seems logical that improved basic education plays also an important role for on-the-job skill acquirement, e.g. in form of vocational trainings.

Individuals older than 32 years represent a placebo group (introduced in Section 4) because there is no doubt they are all untreated. As already mentioned, the baseline regressions yield occasionally significant effects thus motivating DDD estimation in the first place. Restricting the sample to only elder people precludes triple differences by construction. Therefore in this paragraph we refer to DID results (Table 4). Including core controls predicted impacts are either insignificant or suggesting an opposing effect relative to the full sample. By contrast to the full sample the placebo group is very sensitive upon extending the set of controls, above all to aggregate-level regressors. Finally controlling for average educational attainment per size class always renders placebo estimates insignificant. Nevertheless emphasizing DDD instead of DID results seems advisable.

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 equilibrated gains by class size reduction.

Saarland's data confirm girls' final grade attainment to improve more strongly than that of boys. This holds true in the sense of more sizable estimates as well as statistical significance. For boys there does not exist a single outcome precisely enough estimated to obtain significance (Table 8). Standard errors repeatedly exceed those observed for girls, most notably for the per se sizable increase in employment probability (13 percentage points, se 0.09). The female subsample yields quite different evidence (Table 7). Concretely for six out of ten outcomes there exist at least marginally significant impacts. In size they tend to be similar or larger than coefficients predicted for the overall sample. The extremest gap of 3 percentage points is associated with obtaining no more than a *Volksschulabschluss* which is predicted to decrease by 8.9 percentage points for girls alone. What drives exclusively female benefits from single-grade classes? 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 Rønning 2011). Thus learning in highly heterogeneous multigrade groups might be more demanding for them. However most discriminative seem marriage patterns. Balancing tests report initially 10 percentage points less female than male singles (Table 1). Post-treatment the share is more equilibrated with 95 to 99 percentage points respectively. Controlling for marital status dissolves significance of coefficients on combined grade attainment and on the indicator of overall financial support. The specification also negates that single-grade schooling shrinks the frequency of becoming a housewife. By contrast impacts on *Volksschulabschluss*, *Blue-Collar Employee* and *Supported by Family* are robustly significant and reduced by at most 2 percentage points. Implications are twofold. On the one hand there exists stable evidence on girls' responsiveness in every outcome group. What is more girls drive impacts strongly enough to translate significance into the full sample's analysis. On the other hand marital status is revealed to be a mediator (Lundborg et al. 2012). It calls a direct causal relationship between combination classes and among others female labor market participation into question. DID regressions depict the marital status' mediator function less clearly which supports that the DDD model is more appropriate here. None of the other regressor extensions provoke substantial changes in the gender subsamples.

Blue- & White-Collar Parent The literature further proposes to check out varying impacts by pupils' financial backgrounds. Chetty, Friedman, et al. (2014) present evidence that teacher quality impacts are constant in terms of parent income. If similar mechanism apply to other schooling inputs likewise the higher tail of the income distribution might be most reactive to the abolition of parochial schools. Taking a different perspective Dahl and Lochner (2012) evaluate the impact of relaxing a household's monetary constraints on test scores of its children. Unsurprisingly the poorest children show relatively largest and significant improvements. In this case treatment and stratum are directly related. However Jones (2013) also speaks of particular responsiveness to treatment for pupils from financially constraint backgrounds. Nonlinear responsiveness might be explained by decreasing marginal returns of schooling inputs and/or substitution effects between schooling inputs and teacher quality. Taking this idea further positive effects observed in the Saarland might be magnified in a low-income stratum by global instead of German standards.

Socioeconomic asymmetries in our dataset are pronounced clearly enough for subgroup-specific patterns. For children of blue-collar parents beneficial effects from single-grade classes seem especially sustainable (Table 10). Concretely the decrease in predicted lowest degrees (6.4 percentage points, se 0.02) translates into a significant increase in intermediate achievement (3.2 percentage points, se 0.01). Notably impacts on *Mittlere Reife* are not significant for any other (sub)group. By contrast the white-collar stratum belongs to the few subsamples without any effect on degree indicators (Table 9). Which is not primarily due to a lack of precision but rather because the point estimates themselves are very small. Remarkably DDD estimation is needed to drive them down to zero whereas DID models predict a *larger* shrink in basic degree holders for white- than blue-collar backgrounds. DID estimates are mentioned here only for the sake of underlining importance of triple differences that are discussed by default.

Labor market indicators trace out a nearly one-to-one highly significant as well as huge decrease (about 27 percentage points) in the probability to become a blue-collar (white-collar) worker for pupils born to blue-collar (white-collar) parents. Theoretically this could depict a fierce crowding-out driven by all of sudden competetive rural blue-collar children. As shown for schooling outcomes educational opportunities translate into grade attainment which in turn conditions socioeconomic mobility. This should be more pronounced in professional trajectories of pupils who by family and neighborhood cannot compensate for poor schooling (Table 1). Remaining labor market and financial outcomes show positive effects that are more pronounced and more precisely estimated for disenfranchised children. Observable estimated benefits for these children could be driven by potential mediators such as smaller households, later marriages, higher educated teachers or more academic neighborhoods. All of them are suggested byproducts of rural convergence to urban standards.Given there are more rural pupils born with a blue-collar background these factors might push performance asymmetrically. However coefficients are stable across long and short regressions.

Catholics & Protestants Catholics mimic the fulls sample's pattern in every respect. Surprisingly Protestants seem to be robustly unaffected by the reform. At first sight this seems very surprising given Protestant schools merge on average more grade levels than Catholic ones.

Strictly speaking denominational tracking creates two separate treatment-control comparisons. All four groups experience different schooling inputs. By pooling them into common schools the reform introduces universal schooling conditions for all rural and all urban pupils respectively. Due to distinct pre-reform key features the shift parochial pupils are subjected to is not uniform. The majority of multigrade schools are Catholic because the share of Catholic versus Protestant pupils is even higher in rural regions. Thus the pooled post-reform trend in covariate schooling characteristics, i.e. class size, turns out to be close to that of Catholics alone. Simultaneously the reform evokes a discontinuity for Protestants. Due to that differencing without considering schooling regressors seems inappropriate.

These insights allow to disentangle possible mechanisms as follows. To begin with treatment-control *differences* in the number of multigrade levels are nearly equilibrated. The gaps span roughly 2.5 levels in both denomination strata. However for overall scarce Protestant pupils there exists no single-grade control group. Instead urban children are mostly taught in two-level classes. Underlying trends show that shortly before the legal change inures the gap between Catholic and Protestant schools increases which should actually heighten the reform's specific impact. One possible explanation are level-dependent gains from mixing a grade level less. Let single-grade classes represent a panacea while any multigrade structure produces severe disadvantages. Then the reduction from 4.5 to 2 collapsed grade levels faced by Protestants is located on a much flatter portion of the assumedly concave gains curve.

Additionally, over time Protestant (Catholic) rural class sizes shrink by 0.9 (0.5)[31] pupils *less* than urban ones. This violation of the common trend assumption drives down positive treatment effects and relative more so for Protestants. If class size turned out to play a key role it would allow to quantify combination class costs and benefits in terms of class size reduction. It would lend a cardinal meaning to rather vague comparisons stated by policy makers claiming multigrade classes to be more/ less 'cost-effective' a device (Benveniste and McEwan 2000) for education expansion relative to diminishing class size (Jones 2013). So far, however, controlling for class size does not yield significant impacts on Protestant pupils although class size tends to overstate performance of the Catholic relative to the Protestant treatment group. Potential beneficial peer effects measured in terms of *Teachers per Schools* also advantages Catholic schools. However examining remaining key figures finds the twin indicator *Students/Teacher* to euphemize Protestants' performance. The same holds true for the *Girls' Share* shrinking in general more in the control groups but even more so in the Catholic one. This multitude of counteracting effects complicates to find out which effect possibly just offsets the gains from single-grade classes upon dissolving denominational tracking. But rerunning regressions including indicators from the Schools' Index does not result in significant coefficients for Protestants either.

Neither does the census-based teacher quality variable shed light on the phenomenon. Perhaps Protestant pedagogy is more advanced in the sense of adapted to the multigrade framework. There is no information on this in the data. Obviously these ideas and the resulting specifications intend to pin down some educational input explaining why Protestant multigrade schools seem at eye-level with

[31] Class Size Gap: 32.8-23.1 - (32.5-23.7) = 0.9 (Protestants) and 36.6-23.1 - (35.7-22.7) = 0.5 (Catholics).

post-reform single-grade common schools. This approach draws on Becker and Woessmann (2009) who connect wide-spread literacy to Protestants' prosperity. Recall that McEwan (2008) details the special importance of teachers' motivation upon dealing with multigrade classes. This touches upon the Weber Hypothesis of Protestants' inherently superior work ethics. Stratifying by denomination leads up to a very promising strand of Saarland's natural experiment.

7 Conclusion

This paper addressed the question how attending a multigrade school affected school attainment and labor market outcomes, and whether there are any gender differences in this effect. To answer this question our analysis has exploited the abolition of Saarland's parochial 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, namely the asymmetrically reduced probability to attend a multigrade class. 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. The most reliable results unambiguously suggest single-grade classes to be (weakly) preferable. Interestingly, effects are more frequently significant for labor market and corresponding financial outcomes (five out of six) than for those directly capturing education (one out of four). Treated pupils shift away from obtaining only a *Volksschulabschluss* and a blue-collar job. Their overall employment probability rises accompanied by a decreased need for parental support and less financial dependencies in general. Stratifying the main sample the emerging patterns line up with asymmetric treatment responses observed in related studies. For children with less educated parents the shrink in basic grade attainment transmits into a significant rise in intermediate degrees. Splitting the sample by denomination suggests that Protestant schools provide some benefit that so far escapes regression controls and persistently offsets the reform's impact.

Girls are most broadly affected albeit to some extent their results are driven by reform-induced postponed marriages. The latter seems to be particularly the case among Catholic girls, not among Protestant girls. Our results therefore suggest an interplay of socialization, potentially based on religious denomination and stereotypes (the *Catholic housewife*), and the mode of teaching in terms of multigrade classes (which is likely to be more competitive and therefore creates a more detrimental learning environment for girls).

The research approach also 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. 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 pupils. Preliminary findings nevertheless suggest that a beneficial multigrade system needs strategic adjustments. We conclude that peer effects based on pupil collaboration alone are no panacea which refutes the argument that reallocation is a *costless* way to improve education.

References

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

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

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

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

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

Benveniste, Luis and Patrick McEwan (2000). "Constraints to Implementing Educational Innovations: The Case of Multigrade Schools." *International Review of Education* 46(1/2), 31–48.

Carrell, Scott E, Bruce I Sacerdote, and James E West (2013). "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81(3), 855–882.

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

Chetty, Raj, Nathaniel Hendren, et al. (2014). "Is the United States still a land of opportunity? Recent trends in intergenerational mobility." *American Economic Review* 104(5), 141–147.

Clark, Damon and Emilia Del Bono (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, James S (1968). "Equality of Educational Opportunity Study (EEOS)." *Equity & Excellence in Education* 6(5), 19–28.

Cortes, Kalena E and Joshua S Goodman (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, Julie Berry, Brian A. Jacob, and Steven Levitt (2006). "The effect of school choice on participants: Evidence from randomized lotteries." *Econometrica* 74(5), 1191–1230.

Dahl, Gordon B and Lance Lochner (2012). "The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit." *The American Economic Review* 102(5), 1927–1956.

Deming, David J et al. (2014). "School Choice, School Quality, and Postsecondary Attainment." *The American Economic Review* 104(3), 991–1013.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (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*. Tech. rep. 19. EENEE (European Expert Network on Economics of Education).

Faust, Gabriele (2006). "Zum Stand der Einschulung und der neuen Schuleingangsstufe in Deutschland." *Zeitschrift für Erziehungswissenschaft* 9(3), 328–347.

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

Greenwood, Jeremy et al. (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, Eric a. et al. (2003). "Does peer ability affect student achievement?" *Journal of Applied Econometrics* 18(5), 527–544.

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

Iversen, Jon M and Hans Bonesrønning (2015). "Conditional gender peer effects?" *Journal of Behavioral and Experimental Economics* 55, 19–28.

Jacob, Brian A, Lars Lefgren, and David P Sims (2010). "The persistence of teacher-induced learning." *Journal of Human Resources* 45(4), 915–943.

Johansson, Elly-Ann and Erica Lindahl (2008). "The effects of mixed-age classes in Sweden Erica Lindahl." *IFAU Working Paper Series* 21.

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

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

Lavy, Victor, M Daniele Paserman, and Analia Schlosser (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 M Rønning (2011). *Classroom Grade Composition and Pupil Achievement*. Tech. rep. Norway: Statistics Norway (SSB) and Department of Economics, University of Oslo.

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

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

Lundborg, Petter, Anton Nilsson, and Dan-Olof Rooth (2012). “Parental Education and Offspring Outcomes : Evidence from the Swedish Compulsory Schooling Reform.” *IZA Discussion Paper Series* 6570.

Mason, DeWayne A and Robert B Burns (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, Patrick J (2008). “Evaluating multigrade school reform in Latin America.” *Comparative Education* 44(4), 465–483.

Mulkeen, Aidan and Cathal Higgings (2009). “Multigrade Teaching in Sub-Saharan Africa.” *World Bank Working Paper Series* 173.

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

Rothstein, Jesse 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, Jean V, Kenneth J Rowe, and Peter W Hill (1998). “Effects of Multigrade Classes on Student Progress in Literacy and Numeracy: Quantitative Evidence and Perceptions of Teachers and School Leaders.” *Annual Meeting of the Australian Association for Research in Education*. Adelaide.

Stephens, Melvin and Dou-Yan Yang (2014). “Compulsory Education and the Benefits of Schooling.” *American Economic Review* 104(6), 1777–1792.

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

Veenman, Simon (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, Diane (2005). “Resource and peer impacts on girls’ academic achievement: Evidence from a randomized experiment.” *The American Economic Review* 95(2), 199–203.

Appendices

A Balancing & Regression Tables

Table 1: Characteristics 15-20 Year-Olds

	(1) Control PRE	(2) Treatment PRE	(3) Diff. T-C (*t*-stat)	(4) Control POST	(5) Treatment POST	(6) Diff. T-C (*t*-stat)
Female	0.49	0.47	-0.02**	0.49	0.49	0.00
	[0.50]	[0.50]	(-2.45)	[0.50]	[0.50]	(0.74)
Age	17.6	17.5	-0.10***	17.8	17.7	-0.05**
	[1.71]	[1.71]	(-2.91)	[1.69]	[1.69]	(-2.34)
Young at School Entry	0.41	0.42	0.01	0.40	0.41	0.01*
	[0.49]	[0.49]	(1.05)	[0.49]	[0.49]	(1.66)
Catholic	0.68	0.82	0.14***	0.75	0.84	0.10***
	[0.47]	[0.38]	(16.60)	[0.44]	[0.36]	(22.99)
German	0.96	0.98	0.01***	0.95	0.97	0.03***
	[0.19]	[0.15]	(3.63)	[0.22]	[0.16]	(13.24)
Single	0.93	0.93	-0.01	0.97	0.97	0.00
	[0.25]	[0.26]	(-1.05)	[0.17]	[0.16]	(1.42)
Household Size	4.47	4.67	0.21***	3.88	4.12	0.24***
	[1.94]	[1.82]	(5.57)	[1.34]	[1.23]	(16.60)
White Collar Breadwinner	0.47	0.35	-0.12***	0.43	0.39	-0.04***
	[0.50]	[0.48]	(-8.30)	[0.49]	[0.49]	(-6.35)
Commuting Time (1-5)	2.75	2.92	0.17***	2.99	3.04	0.05***
	[0.98]	[1.06]	(8.21)	[0.83]	[0.84]	(4.74)
0-15 Minutes	0.26	0.22	-0.04***	0.27	0.26	-0.01***
	[0.44]	[0.41]	(-4.37)	[0.44]	[0.44]	(-2.69)
15-30 Minutes	0.42	0.37	-0.05***	0.44	0.44	-0.00
	[0.49]	[0.48]	(-5.19)	[0.50]	[0.50]	(-0.51)
30-60 Minutes	0.17	0.24	0.07***	0.24	0.25	0.01*
	[0.38]	[0.43]	(8.27)	[0.43]	[0.43]	(1.71)
Observations	1,681	2,560	4,241	57,979	6,029	64,008

Note: *Census 1970 and 1987 (Own calcualtions). See main text for details. Standard errors in brackets.*
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2: Outcomes 15-20 Year-Olds

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DDD										
Baseline	0.0443	-0.0413***	0.0182*	-0.00205	-0.00408	0.0118	-0.0267	-0.0143	-0.00333	-0.0148
	[0.0238]	[0.0104]	[0.00828]	[0.00738]	[0.0118]	[0.0148]	[0.0162]	[0.0110]	[0.0113]	[0.0108]
Core Controls	0.107**	-0.0665***	0.00630	0.00844	0.0377*	-0.203***	-0.0433	-0.0319*	-0.0500***	-0.0554***
	[0.0395]	[0.0156]	[0.0131]	[0.0117]	[0.0146]	[0.0450]	[0.0348]	[0.0125]	[0.0143]	[0.0141]
Aggregate Controls	0.107**	-0.0662***	0.00568	0.00882	0.0381**	-0.205***	-0.0471	-0.0321**	-0.0505***	-0.0559***
	[0.0395]	[0.0156]	[0.0131]	[0.0117]	[0.0146]	[0.0451]	[0.0349]	[0.0125]	[0.0143]	[0.0141]
DID (binary)										
Baseline	0.0667*	-0.0620***	0.0492***	-0.00832	0.0385**	-0.0336	0.0371*	-0.0156***	-0.0376**	-0.0402**
	[0.0288]	[0.0141]	[0.0128]	[0.00828]	[0.0133]	[0.0179]	[0.0178]	[0.00416]	[0.0134]	[0.0138]
Core Controls	0.0796	-0.0835***	0.0365*	-0.00823	0.0328*	-0.0284	0.0171	-0.0235***	-0.0323	-0.0369*
	[0.0425]	[0.0191]	[0.0184]	[0.0114]	[0.0166]	[0.0284]	[0.0277]	[0.00573]	[0.0167]	[0.0166]
Aggregate Controls	0.0681	-0.0854***	0.0456*	-0.0213	0.0344	-0.0502	0.0332	-0.0271***	-0.0336	-0.0369*
	[0.0482]	[0.0214]	[0.0206]	[0.0128]	[0.0187]	[0.0323]	[0.0315]	[0.00644]	[0.0187]	[0.0187]
DID (continuous)										
Baseline	0.0574*	-0.0715***	0.0688***	-0.0275***	0.0176	-0.0764***	0.0600***	-0.0200***	-0.0153	-0.000214
	[0.0282]	[0.0136]	[0.0123]	[0.00797]	[0.0129]	[0.0177]	[0.0176]	[0.00403]	[0.0130]	[0.0134]
Core Controls	0.0615	-0.0707***	0.0455**	-0.0249*	0.0363*	-0.0460	0.0240	-0.0283***	-0.0351*	-0.0364*
	[0.0419]	[0.0183]	[0.0176]	[0.0109]	[0.0160]	[0.0281]	[0.0274]	[0.00552]	[0.0160]	[0.0160]
Aggregate Controls	0.0713	-0.0798***	0.0473**	-0.0223*	0.0398*	-0.0438	0.0220	-0.0300***	-0.0386*	-0.0407*
	[0.0424]	[0.0186]	[0.0180]	[0.0111]	[0.0163]	[0.0283]	[0.0276]	[0.00561]	[0.0163]	[0.0163]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DDD)	412461	414297	414297	414297	460441	277555	277555	460441	460441	460441
N (DID)	55611	57408	57408	57408	77419	35646	35646	77419	77419	77419

Standard errors in brackets. See main text for details.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Outcomes' Pre-Reform Benchmark 15-20 Year-Olds

	(1) Control PRE	(2) Treatment PRE	(3) Diff. T-C (*t*-stat)
Volksschulabschluss	0.66	0.72	0.06***
	[0.47]	[0.45]	(6.67)
Mittlere Reife	0.091	0.074	-0.02***
	[0.29]	[0.26]	(-3.12)
Abitur	0.039	0.023	-0.02***
	[0.19]	[0.15]	(-4.62)
Employed	0.46	0.47	0.01
	[0.50]	[0.50]	(0.66)
Blue-Collar Employee	0.43	0.46	0.03***
	[0.49]	[0.50]	(2.91)
White-Collar Employee	0.47	0.41	-0.06***
	[0.50]	[0.49]	(-5.13)
Housewife	0.043	0.046	0.00
	[0.20]	[0.21]	(0.74)
Financially Supported	0.53	0.53	-0.01
	[0.50]	[0.50]	(-0.58)
Supported by Family	0.52	0.52	0.01
	[0.50]	[0.50]	(0.72)
Observations	2,874	4,481	7,355

Note: *Census 1970 (Own calculations).*
Standard errors in brackets. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4: Outcomes Placebo Group

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DID (binary)										
Baseline	0.000300	-0.00983*	0.0165***	-0.00849*	0.0151**	-0.00860	0.0311***	0.000990	0.000382	0.0103*
	[0.0106]	[0.00478]	[0.00369]	[0.00351]	[0.00570]	[0.00827]	[0.00928]	[0.00567]	[0.00543]	[0.00513]
Core Controls	-0.0546*	0.0114	0.0123	-0.0258***	-0.0252**	0.0372	-0.00168	0.0163	0.0353***	0.0373***
	[0.0221]	[0.00987]	[0.00807]	[0.00772]	[0.00906]	[0.0532]	[0.0359]	[0.00852]	[0.00884]	[0.00858]
Aggregate Controls	-0.0416	-0.00413	0.0343***	-0.0322***	-0.0366***	0.0324	-0.0131	0.0238*	0.0507***	0.0527***
	[0.0251]	[0.0112]	[0.00917]	[0.00876]	[0.0103]	[0.0560]	[0.0378]	[0.00968]	[0.0100]	[0.00975]
DID (continuous)										
Baseline	-0.0119	-0.0128**	0.0324***	-0.0210***	0.0264***	-0.0281***	0.0447***	0.0123*	0.00567	0.0224***
	[0.00999]	[0.00452]	[0.00349]	[0.00332]	[0.00542]	[0.00763]	[0.00856]	[0.00536]	[0.00517]	[0.00488]
Core Controls	-0.0583**	0.00505	0.0309***	-0.0365***	-0.0409***	0.0284	-0.0657*	0.0298***	0.0542***	0.0563***
	[0.0217]	[0.00970]	[0.00793]	[0.00758]	[0.00891]	[0.0443]	[0.0299]	[0.00838]	[0.00870]	[0.00844]
Aggregate Controls	-0.0674**	0.0101	0.0269***	-0.0377***	-0.0409***	0.0400	-0.0639*	0.0298***	0.0542***	0.0565***
	[0.0220]	[0.00982]	[0.00804]	[0.00768]	[0.00903]	[0.0471]	[0.0318]	[0.00849]	[0.00881]	[0.00855]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DID)	356850	356889	356889	356889	383022	241909	241909	383022	383022	383022

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986 (Own calculations).*

Table 5: Outcomes Catholics

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DDD										
Baseline	0.0436	-0.0424***	0.0203*	-0.00311	-0.00139	-0.00917	0.00730	-0.0135	-0.0101	-0.0198
	[0.0261]	[0.0115]	[0.00925]	[0.00795]	[0.0136]	[0.0171]	[0.0186]	[0.0128]	[0.0131]	[0.0126]
Core Controls	0.131**	-0.0698***	0.00768	0.0165	0.0395*	-0.205***	-0.0389	-0.0342*	-0.0538**	-0.0603***
	[0.0441]	[0.0176]	[0.0148]	[0.0128]	[0.0171]	[0.0540]	[0.0419]	[0.0145]	[0.0167]	[0.0165]
Aggregate Controls	0.131**	-0.0691***	0.00695	0.0167	0.0398*	-0.209***	-0.0421	-0.0345*	-0.0542**	-0.0607***
	[0.0441]	[0.0176]	[0.0148]	[0.0128]	[0.0171]	[0.0543]	[0.0420]	[0.0145]	[0.0167]	[0.0165]
DID (binary)										
Baseline	0.0663*	-0.0640***	0.0473**	-0.00682	0.0479**	-0.0505*	0.0633**	-0.0163***	-0.0477**	-0.0468**
	[0.0326]	[0.0162]	[0.0147]	[0.00926]	[0.0154]	[0.0204]	[0.0203]	[0.00466]	[0.0155]	[0.0159]
Core Controls	0.0903	-0.0872***	0.0368	-0.00668	0.0358	-0.0389	0.0286	-0.0267***	-0.0353	-0.0410*
	[0.0484]	[0.0220]	[0.0212]	[0.0129]	[0.0193]	[0.0322]	[0.0314]	[0.00645]	[0.0193]	[0.0193]
Aggregate Controls	0.102	-0.101***	0.0406	-0.0133	0.0306	-0.0593	0.0418	-0.0270***	-0.0296	-0.0338
	[0.0574]	[0.0255]	[0.0246]	[0.0150]	[0.0224]	[0.0385]	[0.0375]	[0.00750]	[0.0225]	[0.0224]
DID (continuous)										
Baseline	0.0837*	-0.0857***	0.0741***	-0.0211*	0.0408*	-0.103***	0.0923***	-0.0188***	-0.0396*	-0.0165
	[0.0345]	[0.0168]	[0.0153]	[0.00965]	[0.0161]	[0.0218]	[0.0217]	[0.00488]	[0.0162]	[0.0167]
Core Controls	0.0989	-0.0943***	0.0431	-0.0161	0.0371	-0.0585	0.0316	-0.0325***	-0.0355	-0.0379
	[0.0523]	[0.0229]	[0.0221]	[0.0134]	[0.0201]	[0.0350]	[0.0341]	[0.00674]	[0.0202]	[0.0201]
Aggregate Controls	0.100	-0.0973***	0.0446*	-0.0156	0.0396*	-0.0584	0.0316	-0.0339***	-0.0380	-0.0409*
	[0.0524]	[0.0230]	[0.0222]	[0.0135]	[0.0202]	[0.0350]	[0.0341]	[0.00676]	[0.0202]	[0.0202]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DDD)	300822	302199	302199	302199	336427	198856	198856	336427	336427	336427
N (DID)	42772	44125	44125	44125	58777	27703	27703	58777	58777	58777

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986 (Own calculations).*

Table 6: Outcomes Protestants

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DDD										
Baseline	-0.0541	0.00805	-0.0173	-0.0161	-0.0429	0.0859**	-0.132***	-0.00951	0.0263	0.0232
	[0.0569]	[0.0250]	[0.0201]	[0.0179]	[0.0262]	[0.0331]	[0.0374]	[0.0249]	[0.0249]	[0.0237]
Core Controls	-0.0638	-0.0189	-0.0143	-0.0330	-0.00754	-0.148	-0.0852	-0.00694	0.00309	0.00697
	[0.0890]	[0.0352]	[0.0297]	[0.0269]	[0.0313]	[0.0897]	[0.0709]	[0.0270]	[0.0304]	[0.0299]
Aggregate Controls	-0.0642	-0.0188	-0.0147	-0.0329	-0.00723	-0.147	-0.0898	-0.00693	0.00278	0.00665
	[0.0890]	[0.0352]	[0.0297]	[0.0269]	[0.0313]	[0.0898]	[0.0710]	[0.0270]	[0.0304]	[0.0299]
DID (binary)										
Baseline	0.0282	-0.0231	0.0210	-0.00709	-0.00605	0.0499	-0.0633	-0.0101	0.00860	0.00858
	[0.0666]	[0.0320]	[0.0287]	[0.0198]	[0.0293]	[0.0407]	[0.0407]	[0.00923]	[0.0294]	[0.0304]
Core Controls	0.0321	-0.0522	0.0233	-0.00842	0.00260	0.0385	-0.0472	-0.00855	-0.00197	0.00271
	[0.0962]	[0.0419]	[0.0403]	[0.0264]	[0.0354]	[0.0638]	[0.0629]	[0.0123]	[0.0354]	[0.0356]
Aggregate Controls	-0.0102	-0.0361	0.0425	-0.0308	0.0139	0.0181	-0.0307	-0.0190	-0.0134	-0.00642
	[0.102]	[0.0444]	[0.0427]	[0.0279]	[0.0375]	[0.0672]	[0.0663]	[0.0131]	[0.0376]	[0.0378]
DID (continuous)										
Baseline	-0.00636	-0.0271	0.0431	-0.0315*	-0.0225	-0.0270	-0.000890	-0.0252***	0.0243	0.0364
	[0.0538]	[0.0254]	[0.0228]	[0.0157]	[0.0236]	[0.0327]	[0.0328]	[0.00745]	[0.0237]	[0.0246]
Core Controls	-0.0146	-0.00999	0.0488	-0.0410	0.0222	-0.000263	-0.0108	-0.0216*	-0.0219	-0.0152
	[0.0785]	[0.0336]	[0.0323]	[0.0211]	[0.0286]	[0.0517]	[0.0510]	[0.00996]	[0.0287]	[0.0288]
Aggregate Controls	0.0272	-0.0352	0.0412	-0.0268	0.0154	0.0250	-0.0349	-0.0162	-0.0148	-0.00861
	[0.0848]	[0.0367]	[0.0353]	[0.0231]	[0.0311]	[0.0558]	[0.0550]	[0.0108]	[0.0311]	[0.0313]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DDD)	91270	91705	91705	91705	102249	62742	62742	102249	102249	102249
N (DID)	11177	11597	11597	11597	16162	7025	7025	16162	16162	16162

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986 (Own calculations).*

Table 7: Outcomes Girls

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DDD										
Baseline	0.0852**	-0.0754***	0.0483***	-0.00264	-0.0143	-0.00204	-0.0788**	-0.0281	0.00859	-0.00212
	[0.0293]	[0.0139]	[0.0119]	[0.00874]	[0.0136]	[0.0236]	[0.0253]	[0.0154]	[0.0171]	[0.0172]
Core Controls	0.0937*	-0.0894***	0.0219	0.00146	0.0393	-0.182***	-0.0552	-0.0506*	-0.0532*	-0.0602**
	[0.0466]	[0.0201]	[0.0177]	[0.0133]	[0.0218]	[0.0528]	[0.0542]	[0.0216]	[0.0231]	[0.0233]
Aggregate Controls	0.0936*	-0.0892***	0.0216	0.00156	0.0397	-0.180***	-0.0618	-0.0509*	-0.0537*	-0.0606**
	[0.0466]	[0.0201]	[0.0177]	[0.0133]	[0.0218]	[0.0530]	[0.0545]	[0.0216]	[0.0231]	[0.0233]
DID (binary)										
Baseline	0.0728	-0.0735***	0.0688***	-0.0151	0.0389*	0.000859	0.000386	-0.0368***	-0.0390*	-0.0445*
	[0.0417]	[0.0207]	[0.0193]	[0.0126]	[0.0187]	[0.0235]	[0.0250]	[0.00833]	[0.0188]	[0.0196]
Core Controls	0.119*	-0.112***	0.0485	0.00219	0.0219	-0.00573	0.0133	-0.0492***	-0.0214	-0.0301
	[0.0586]	[0.0275]	[0.0273]	[0.0170]	[0.0231]	[0.0422]	[0.0450]	[0.0112]	[0.0232]	[0.0234]
Aggregate Controls	0.0783	-0.109***	0.0562	-0.0150	0.0226	-0.0113	0.0150	-0.0533***	-0.0220	-0.0305
	[0.0668]	[0.0309]	[0.0307]	[0.0191]	[0.0260]	[0.0477]	[0.0510]	[0.0126]	[0.0261]	[0.0263]
DID (continuous)										
Baseline	0.0727	-0.0957***	0.100***	-0.0377**	0.0138	-0.0471*	0.0414	-0.0416***	-0.0111	-0.000558
	[0.0405]	[0.0197]	[0.0184]	[0.0120]	[0.0179]	[0.0229]	[0.0243]	[0.00800]	[0.0180]	[0.0188]
Core Controls	0.0712	-0.102***	0.0550*	-0.0204	0.0268	-0.0440	0.0414	-0.0536***	-0.0257	-0.0325
	[0.0584]	[0.0265]	[0.0263]	[0.0163]	[0.0223]	[0.0411]	[0.0440]	[0.0108]	[0.0224]	[0.0226]
Aggregate Controls	0.0901	-0.116***	0.0587*	-0.0154	0.0293	-0.0441	0.0430	-0.0583***	-0.0282	-0.0360
	[0.0590]	[0.0269]	[0.0267]	[0.0166]	[0.0227]	[0.0417]	[0.0445]	[0.0110]	[0.0228]	[0.0230]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DDD)	202874	203632	203632	203632	230743	85276	85276	230743	230743	230743
N (DID)	27539	28261	28261	28261	38039	15462	15462	38039	38039	38039

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986 (Own calculations).*

Table 8: Outcomes Boys

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DDD										
Baseline	-0.00379	-0.00813	-0.0108	-0.00447	-0.00351	0.0191	-0.0239	0.00874	-0.00488	-0.0187**
	[0.0374]	[0.0153]	[0.0115]	[0.0118]	[0.0168]	[0.0192]	[0.0210]	[0.00454]	[0.0108]	[0.00713]
Core Controls	0.0630	-0.0147	-0.0803	0.0286	0.130	-0.0292	0.0116	0.000633	-0.00678	-0.00759
	[0.267]	[0.113]	[0.0918]	[0.0925]	[0.0841]	[802.2]	[398.4]	[0.0265]	[0.0649]	[0.0576]
Aggregate Controls	0.0618	-0.0143	-0.0798	0.0276	0.130	0	0	0.000519	-0.00693	-0.00774
	[0.267]	[0.113]	[0.0918]	[0.0925]	[0.0841]	[.]	[.]	[0.0265]	[0.0649]	[0.0576]
DID (binary)										
Baseline	0.0626	-0.0540**	0.0300	-0.00123	0.0402*	-0.0468*	0.0437*	-0.00109	-0.0383*	-0.0382*
	[0.0392]	[0.0190]	[0.0166]	[0.0108]	[0.0189]	[0.0239]	[0.0207]	[0.00159]	[0.0190]	[0.0194]
Core Controls	0.0524	-0.0653*	0.0289	-0.0169	0.0437	-0.0317	0.0217	-0.000880	-0.0430	-0.0435
	[0.0617]	[0.0263]	[0.0247]	[0.0152]	[0.0238]	[0.0372]	[0.0343]	[0.00239]	[0.0238]	[0.0236]
Aggregate Controls	0.0671	-0.0680*	0.0379	-0.0262	0.0469	-0.0699	0.0513	-0.00182	-0.0459	-0.0438
	[0.0697]	[0.0294]	[0.0276]	[0.0170]	[0.0267]	[0.0425]	[0.0393]	[0.00267]	[0.0267]	[0.0265]
DID (continuous)										
Baseline	0.0441	-0.0509**	0.0388*	-0.0171	0.0218	-0.0938***	0.0649**	-0.00261	-0.0200	-0.000411
	[0.0387]	[0.0185]	[0.0161]	[0.0105]	[0.0185]	[0.0239]	[0.0207]	[0.00156]	[0.0185]	[0.0189]
Core Controls	0.0488	-0.0425	0.0373	-0.0286*	0.0457*	-0.0446	0.0167	-0.00211	-0.0444	-0.0402
	[0.0601]	[0.0252]	[0.0236]	[0.0146]	[0.0228]	[0.0372]	[0.0343]	[0.00229]	[0.0229]	[0.0227]
Aggregate Controls	0.0500	-0.0486	0.0380	-0.0280	0.0499*	-0.0398	0.0125	-0.00196	-0.0485*	-0.0451
	[0.0609]	[0.0257]	[0.0240]	[0.0148]	[0.0232]	[0.0375]	[0.0346]	[0.00233]	[0.0233]	[0.0231]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DDD)	209587	210665	210665	210665	229698	192279	192279	229698	229698	229698
N (DID)	28072	29147	29147	29147	39380	20184	20184	39380	39380	39380

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986 (Own calculations).*
Running regressions for the Housewife indicator represents a placebo test here.

Table 9: Outcomes Children Born to White-Collar Parent

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DDD										
Baseline	0.0235	-0.0131	-0.0204	-0.00955	0.0355	0.0943	-0.290***	-0.0193	-0.0209	-0.0305
	[0.0832]	[0.0293]	[0.0245]	[0.0241]	[0.0273]	[0.0494]	[0.0567]	[0.0245]	[0.0257]	[0.0255]
Core Controls	-0.00679	0.000518	-0.0263	-0.0172	0.0233	0.0875	-0.280***	-0.0279	-0.0411	-0.0512*
	[0.0804]	[0.0286]	[0.0244]	[0.0234]	[0.0228]	[0.0492]	[0.0562]	[0.0199]	[0.0224]	[0.0221]
Aggregate Controls	-0.00788	0.00120	-0.0272	-0.0170	0.0237	0.0742	-0.269***	-0.0282	-0.0417	-0.0518*
	[0.0804]	[0.0286]	[0.0244]	[0.0234]	[0.0228]	[0.0494]	[0.0564]	[0.0199]	[0.0224]	[0.0221]
DID (binary)										
Baseline	0.0939	-0.0723*	0.0315	-0.0173	0.00748	0.0527	-0.0703	-0.0377***	-0.00707	-0.0158
	[0.0887]	[0.0325]	[0.0318]	[0.0243]	[0.0275]	[0.0572]	[0.0592]	[0.00654]	[0.0276]	[0.0283]
Core Controls	0.0904	-0.0851**	0.0352	-0.00660	0.0191	0.0577	-0.0645	-0.0369***	-0.0187	-0.0284
	[0.0810]	[0.0310]	[0.0316]	[0.0227]	[0.0252]	[0.0494]	[0.0508]	[0.00650]	[0.0253]	[0.0257]
Aggregate Controls	0.0825	-0.0885**	0.0532	-0.0208	0.0217	0.0222	-0.0352	-0.0397***	-0.0211	-0.0280
	[0.0879]	[0.0335]	[0.0340]	[0.0245]	[0.0272]	[0.0534]	[0.0549]	[0.00701]	[0.0273]	[0.0277]
DID (continuous)										
Baseline	0.0744	-0.0503	0.0540*	-0.0322	0.00846	-0.0478	0.0239	-0.0323***	-0.00739	-0.00726
	[0.0785]	[0.0282]	[0.0276]	[0.0210]	[0.0239]	[0.0503]	[0.0521]	[0.00568]	[0.0239]	[0.0246]
Core Controls	0.0842	-0.0594*	0.0562*	-0.0242	0.0216	-0.0372	0.0138	-0.0320***	-0.0204	-0.0213
	[0.0717]	[0.0269]	[0.0274]	[0.0197]	[0.0219]	[0.0435]	[0.0447]	[0.00565]	[0.0220]	[0.0223]
Aggregate Controls	0.105	-0.0729**	0.0538	-0.0177	0.0241	-0.0131	-0.00926	-0.0375***	-0.0230	-0.0269
	[0.0746]	[0.0283]	[0.0288]	[0.0207]	[0.0230]	[0.0453]	[0.0466]	[0.00593]	[0.0231]	[0.0235]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DDD)	187786	188758	188758	188758	211468	125735	125735	211468	211468	211468
N (DID)	20885	21828	21828	21828	32515	12362	12362	32515	32515	32515

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

***Note:** Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986 (Own calculations).*

Table 10: Outcomes Children Born to Blue-Collar Parent

	Schooling				Labor Market				Financial Support	
	Index	Volkschule	Mittlere Reife	Abitur	Employed	Blue-Collar	White-Collar	Housewife	Supported	Family
DDD										
Baseline	0.0959*	-0.0665***	0.0354*	0.0108	0.0326	-0.276***	0.0429	-0.00421	-0.0181	-0.0199
	[0.0373]	[0.0167]	[0.0139]	[0.0102]	[0.0233]	[0.0667]	[0.0440]	[0.0205]	[0.0220]	[0.0218]
Core Controls	0.0937*	-0.0639***	0.0323*	0.0111	0.0411*	-0.273***	0.0149	-0.0271	-0.0486**	-0.0507**
	[0.0372]	[0.0167]	[0.0139]	[0.0101]	[0.0192]	[0.0662]	[0.0381]	[0.0161]	[0.0187]	[0.0183]
Aggregate Controls	0.0928*	-0.0635***	0.0318*	0.0112	0.0415*	-0.275***	0.0176	-0.0273	-0.0491**	-0.0512**
	[0.0372]	[0.0167]	[0.0139]	[0.0101]	[0.0192]	[0.0665]	[0.0382]	[0.0161]	[0.0187]	[0.0183]
DID (binary)										
Baseline	0.106*	-0.0838***	0.0609**	0.00118	0.0145	-0.113**	0.113**	-0.0100	-0.0138	-0.0113
	[0.0505]	[0.0250]	[0.0229]	[0.0119]	[0.0247]	[0.0391]	[0.0378]	[0.00884]	[0.0247]	[0.0251]
Core Controls	0.0898	-0.0764**	0.0526*	0.00239	0.0258	-0.0739*	0.0663*	-0.0109	-0.0252	-0.0234
	[0.0483]	[0.0243]	[0.0226]	[0.0116]	[0.0221]	[0.0349]	[0.0330]	[0.00860]	[0.0222]	[0.0219]
Aggregate Controls	0.0881	-0.0732*	0.0597*	-0.00213	0.0157	-0.0895*	0.0809*	-0.0120	-0.0148	-0.0115
	[0.0568]	[0.0285]	[0.0264]	[0.0136]	[0.0260]	[0.0411]	[0.0389]	[0.0101]	[0.0260]	[0.0257]
DID (continuous)										
Baseline	0.0941	-0.0801**	0.0699**	-0.00862	0.0156	-0.0935*	0.0869*	-0.0175	-0.0143	-0.0111
	[0.0536]	[0.0263]	[0.0241]	[0.0125]	[0.0261]	[0.0415]	[0.0401]	[0.00933]	[0.0261]	[0.0265]
Core Controls	0.0774	-0.0721**	0.0616**	-0.00794	0.0217	-0.0634	0.0549	-0.0175	-0.0204	-0.0173
	[0.0512]	[0.0256]	[0.0237]	[0.0122]	[0.0233]	[0.0370]	[0.0350]	[0.00907]	[0.0234]	[0.0231]
Aggregate Controls	0.0808	-0.0760**	0.0633**	-0.00740	0.0238	-0.0645	0.0559	-0.0179*	-0.0225	-0.0195
	[0.0513]	[0.0256]	[0.0238]	[0.0122]	[0.0234]	[0.0370]	[0.0350]	[0.00909]	[0.0234]	[0.0232]
Size Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (DDD)	224675	225539	225539	225539	248973	151820	151820	248973	248973	248973
N (DID)	34726	35580	35580	35580	44904	23284	23284	44904	44904	44904

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986 (Own calculations).*

Table 11: Multigrade Structure - Degree & Prevalence

	(1) Control PRE	(2) Treatment PRE	(3) Diff. T-C (*t*-stat)	(4) Control POST	(5) Treatment POST	(6) Diff. T-C (*t*-stat)
Multigrade School	0.25	0.68	0.43***	0.013	0.035	0.02***
	[0.43]	[0.47]	(24.86)	[0.11]	[0.18]	(4.37)
Mixed Levels/School	1.44	4.02	2.59***	0.034	0.13	0.09***
	[2.27]	[2.87]	(26.89)	[0.23]	[0.52]	(7.18)
Mixing Two Levels	0.086	0.056	-0.03***	0.016	0.046	0.03***
	[0.28]	[0.23]	(-2.83)	[0.12]	[0.21]	(5.40)
Mixing Three Levels	0.060	0.045	-0.01		0.025	0.02***
	[0.24]	[0.21]	(-1.62)		[0.16]	(3.71)
Mixing Four Levels	0.026	0.053	0.03***		0.00044	0.00
	[0.16]	[0.23]	(3.89)		[0.021]	(1.00)
Mixing Five Levels	0.048	0.077	0.03***		0.0013	0.00*
	[0.21]	[0.27]	(3.25)		[0.036]	(1.73)
Mixing Six Levels	0.056	0.10	0.05***		0.0040	0.00***
	[0.23]	[0.30]	(4.58)		[0.063]	(3.01)
Mixing Seven Levels	0.060	0.17	0.11***			
	[0.24]	[0.38]	(10.30)			
Mixing Eight Levels	0.033	0.17	0.13***			
	[0.18]	[0.37]	(13.72)			
Observations	923	2,263	3,186	1,276	2,262	3,538

Standard errors in brackets.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Schools' Index 1964 - 1986 (Own calculations). Reported are statistics for pre- and post-reform differences in means between treatment (municipality < 10,000 inhabitants) and control groups. 'Multigrade School' is a binary indicator switching on for lower secondary (primary) schools offering six (two) classes or less. This definition maximizes the difference in the reform's impact cross-sectionally. It is enhanced by the more precise 'Mixed Levels/School' that counts the number of missing grade levels. Running from zero to eight e.g. eight indicates that all nine grade levels are taught jointly. In addition each level of 'Mixed Levels/School' is subjected to an individual balancing test via separate dummies omitting the baseline category.*

Table 12: Treatment-Control Specific Time Trends In Schooling Conditions

	Class Size		Multigrade Levels	
	(1)	(2)	(3)	(4)
I. Linear Trend Model				
Linear Trend	-0.856***	-0.855***	-0.0787***	-0.0778***
Rural	-16.03	51.80	297.1***	229.3***
Rural*Trend	0.00798	-0.0311	-0.150***	-0.114***
II. Cohort Dummies Model				
Rural	-0.769		-0.00356	
Rural* 1964	-2.232	-1.523	2.375***	1.702***
Rural* 1965	0.827	1.576	2.501***	1.812***
Rural* 1966	0.926	1.775	2.902***	2.092***
Rural* 1967	1.162	1.754	2.395***	1.806***
Rural* 1968	1.398	2.078*	2.519***	1.890***
Rural* 1969	0.876	1.489	3.031***	2.397***
Rural* 1970	0.284	0.591	2.074***	1.672***
Rural* 1971	0.423	0.713	1.781***	1.337***
Rural* 1974	0.578	0.663	0.0351	-0.0803
Rural* 1978	0.874	0.894	0.108	-0.00309
Rural* 1979	0.595	0.618	0.101*	0.0451
Rural* 1980	0.615	0.866	0.104*	-0.125
Rural* 1981	0.639	0.663	0.125*	0.0494
Rural* 1982	0.600	0.618	0.165***	0.0830
Rural* 1983	0.553	0.659	0.0651	-0.0827
Rural* 1984	0.237	0.225	0.134	0.0717
Rural* 1985	0.475	0.436	0.0896*	0.0538
Rural* 1986	0.485	0.504	0.0972*	-0.0191
Size Class FE	No	Yes	No	Yes
Cohort FE	Model II Yes	Model II Yes	Model II Yes	Model II Yes
N	7517	7517	7517	7517
N Clusters	364	364	364	364

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Schools' Index 1964-1986 (Own calculations). Presented are linear regression estimates with standard errors corrected for clustering by municipalities defined as prior to the aggregation in 1974. Standard errors are available upon request. Model I allows for diverging linear trends in the development of class size and multigrade schools across rural and urban regions. Model II enables non-linear divergence between both regions, with separate interactions for any cohort. All models include a constant.*

Table 13: Standard Schooling Characteristics

	(1) Control PRE	(2) Treatment PRE	(3) Diff. T-C (*t*-stat)	(4) Control POST	(5) Treatment POST	(6) Diff. T-C (*t*-stat)
Class Size	34.9	34.7	-0.17	23.1	22.7	-0.39**
	[6.82]	[7.54]	(-0.63)	[4.78]	[4.65]	(-2.36)
Students/Teachers	33.7	34.6	0.88***	20.0	20.5	0.41**
	[7.22]	[7.83]	(3.03)	[5.22]	[5.37]	(2.23)
Catholic School	0.59	0.68	0.10***			
	[0.49]	[0.46]	(5.25)			
Parochial School	0.93	0.97	0.03***			
	[0.25]	[0.17]	(3.87)			
Girls' Share	0.53	0.50	-0.04***	0.48	0.48	0.00
	[0.19]	[0.099]	(-5.35)	[0.056]	[0.053]	(1.53)
Female Teachers' Share	0.52	0.43	-0.09***	0.59	0.49	-0.10***
	[0.21]	[0.15]	(-11.11)	[0.17]	[0.19]	(-16.54)
Teachers per School	9.16	5.48	-3.68***	13.5	11.6	-1.89***
	[4.16]	[4.60]	(-21.96)	[7.24]	[8.01]	(-7.16)
Pupils per School	310	197	-112.67***	269	233	-35.99***
	[143]	[172]	(-18.95)	[147]	[161]	(-6.75)
Observations	808	1,848	2,656	1,236	2,160	3,396

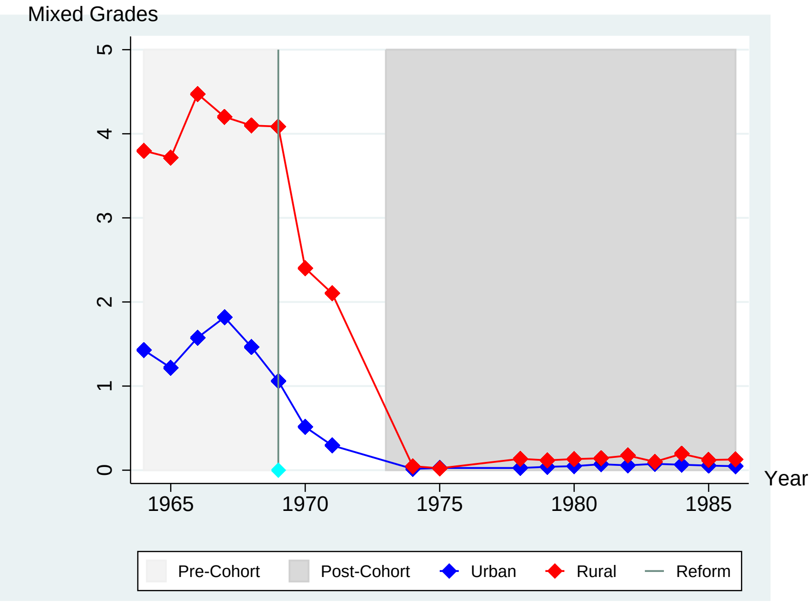
Standard errors in brackets.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: *Schools' Index 1964-1986 (Own calculations). Rural < 10,000 Inh./Mun.*

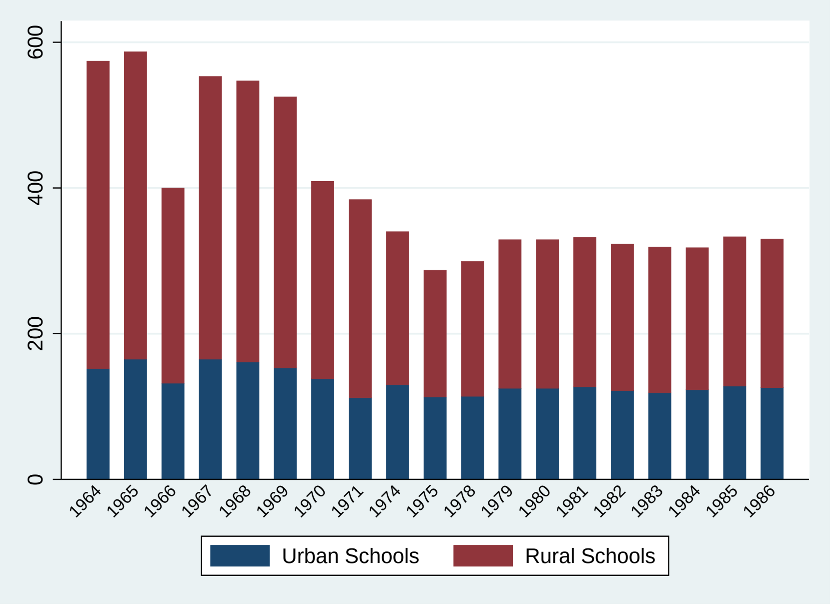
B Figures

Figure 1: Multigrade Structure over Time



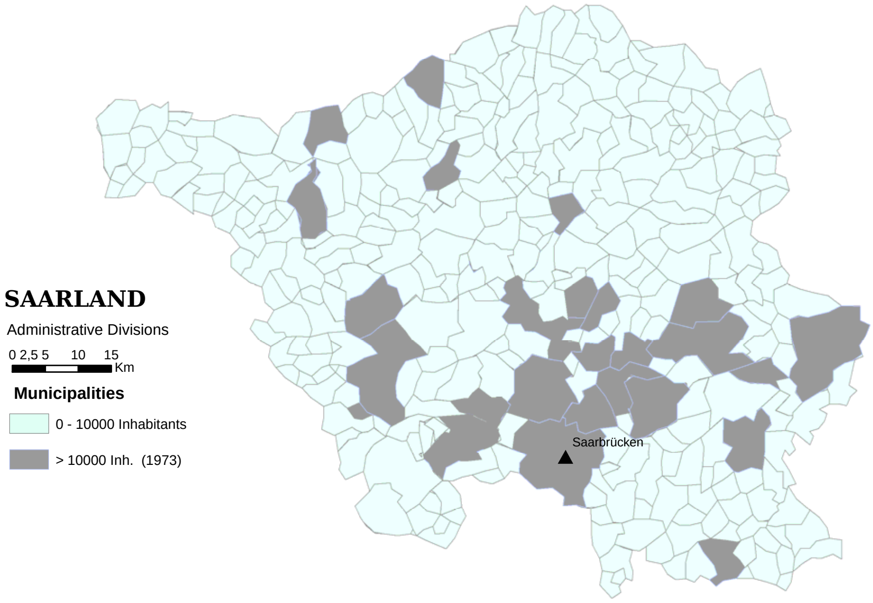
***Note:** Schools' Index 1964-1986 (Own calculations). Rural < 10,000 Inh./Mun.*

Figure 2: Time Trends in Frequency of Rural & Urban Schools



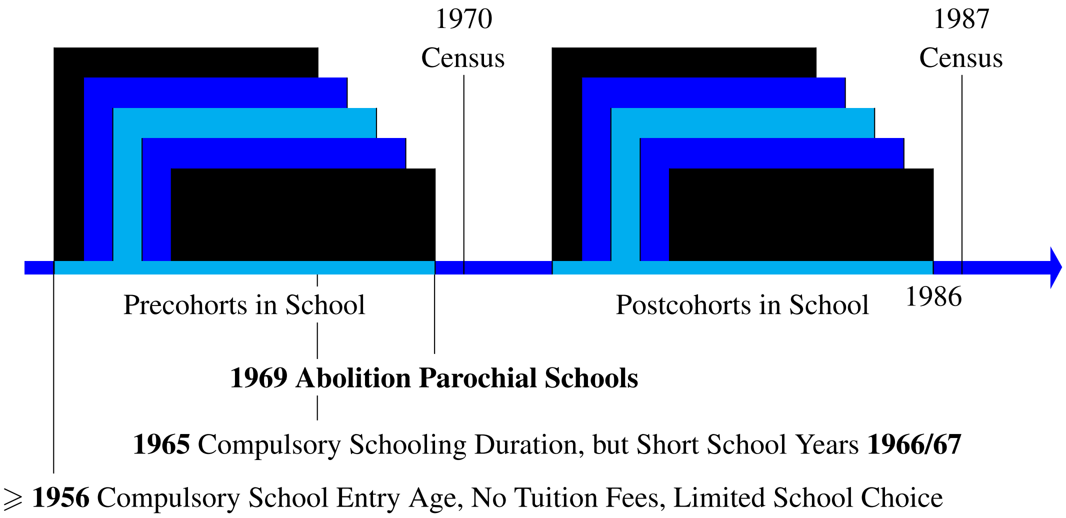
***Note:** Schools' Index 1964-1986 (Own calculations). Rural < 10,000 Inh./Mun.*

Figure 3: Rural vs. Urban Municipalities in the Saarland



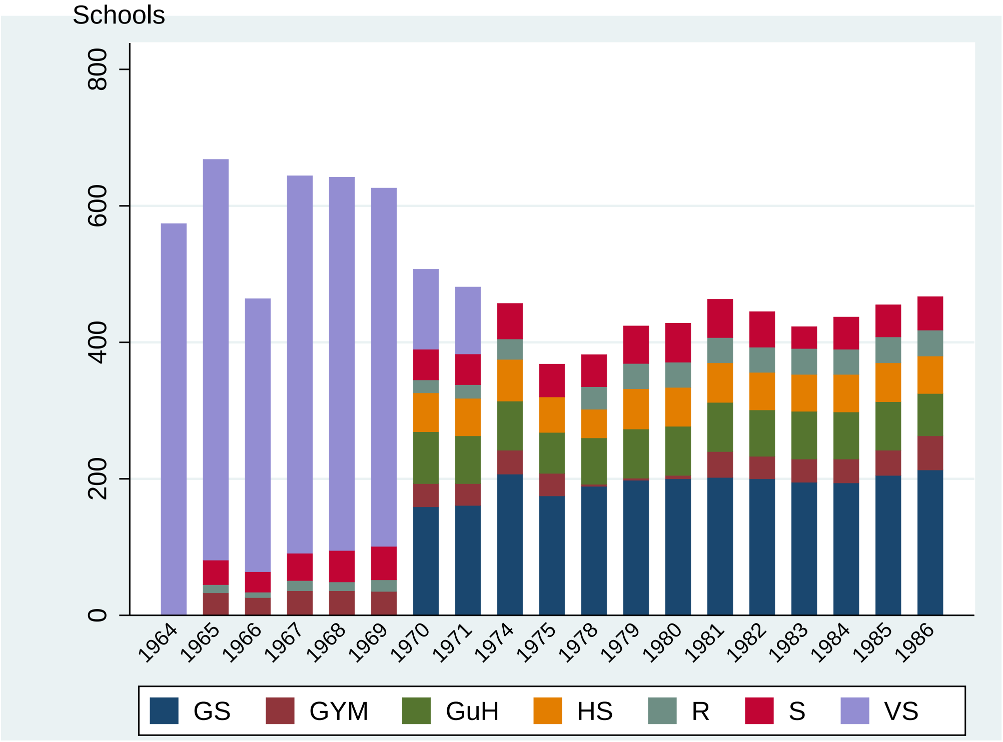
***Note:** Based on population data the map (own creation) depicts the spatial distribution of municipalities surpassing the threshold of 10,000 inhabitants. Administrative borders correspond to the system of 364 autonomous municipalities as in place prior to 1974.*

Figure 4: Timeline of the schooling reform(s) and selection of treatment groups


1970
Census
1987
Census
Precohorts in School
Postcohorts in School
1986
1969 Abolition Parochial Schools
1965 Compulsory Schooling Duration, but Short School Years 1966/67
⩾ 1956 Compulsory School Entry Age, No Tuition Fees, Limited School Choice

***Note:** The timeline is our own creation and depicts the five consecutive pre- and post-reform pupil cohorts we use out of the census observations. Each rectangle represents a basic schooling cycle (1st to 9th grade).*

Figure 5: 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.*