Effects of Multi-grade Classes in Primary Schools on Educational Outcomes

Ilka Gerhardts, University of Munich (LMU) [*]
Uwe Sunde, University of Munich (LMU)
Larissa Zierow, ifo Institute and University of Munich (LMU)

March 2021

Abstract

While teaching more than one age cohort of pupils in one classroom has been advocated as a cost-effective way to provide education and a means of fostering academic achievements through intensified interactions, the evidence on the effects of multi-grade classes on performance is mixed. This article provides novel evidence on the causal effect of multi-grade teaching in primary schools on literacy skills by the end of primary school. The analysis is based on student test score data of more than 68'000 fourth-graders and exploits the staggered introduction of policies targeted at making entry to primary schools more flexible across German states between 2001 and 2016 for identification. The results from a difference-in-differences design document that attending multi-grade classes had negative effects on reading test scores and German grades. These negative effects are more pronounced for girls.

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

Keywords: education, multi-grade classes, gender

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

1 Introduction

Multi-grade teaching with more than one age group attending the same class is common practice in most countries around the world. A considerable proportion of pupils in primary schools in developing countries including India, Peru, Sri Lanka and Pakistan, but also in developed countries including Finland, the Netherlands, the UK and Germany, experience teaching of more than one age cohort of pupils in one classroom (see, e.g., Little, 2004; Mulkeen and Higgings, 2009, for an overview). The determinants of multi-grade teaching are diverse, and range from lack of teachers, rural depopulation, and adjusting to enrollment fluctuations in the context of demographic change, to pedagogical arguments related to peer effects. While multi-grade teaching has been advocated since the 1920s as a way to overcome disadvantages of single-class teaching and to foster the potential of pupil interactions, the evidence on the effects of multi-grade teaching on academic performance, particularly among primary school children, is mixed. Findings of potentially detrimental effects of attending multi-grade classes on subsequent outcomes has led to heated debates regarding the appropriateness of multi-grade teaching as a legitimate goal of education policies (see Carle and Metzen, 2014, for a recent survey of the pedagocial literature).

This paper provides novel evidence on the causal effect of multi-grade teaching in primary schools on literacy skills by the end of primary school. The analysis is based on student test score data of more than 68'000 fourth-graders from Germany. To measure educational outcomes, the analysis considers the performance of fourth-graders, which constitute an important determinant for sorting into the different secondary school tracks, which typically occurs after fourth grade. We combine data originally collected within the PIRLS framework (Progress in International Reading Literacy Study) and the *IQB Laendervergleich* (a German National Student Assessment). In particular, we use test scores for reading skills at the end of fourth grade, German grades at the end of fourth grade, teachers' recommendations for the secondary school track, as well as enjoyment of school as outcomes variables. The analysis makes use of the 2001, 2006, 2011 and 2016 cohorts of fourth-graders, for whom these outcomes are observed, and matches these students with self-collected information about the respective state reforms introducing multi-grade classes in primary school to obtain information about the treatment status.

The identification is based on the repeated comparison of fourth grade student cohorts from schools spread over all states of Germany. The identifying variation is the result of a natural experiment that occurred in the context of the staggered introduction of flexible school entrance levels across German states between 1997 and 2010. This experiment delivers quasi-random variation in the exposure to multi-grade teaching that rules out typical concerns related to selection. The staggered introduction provides variation in treatment exposure that allows us to eliminate state and time fixed effects. Maintaining the standard common trend assumption across states the regional variation in the treatment reveals the causal impact of the experience of multi-grade teaching along the lines of an intention-to-treat analysis.

The results document that exposure to multi-grade teaching has detrimental effects on educational outcomes measured at age 10 (end of fourth grade). On average, multi-grade teaching in the first years of primary school entails a significant and robust negative effect on reading test scores of about 6% of a standard deviation, and a significant negative effect on German grades by 1/9 of a standard deviation. The effects are more pronounced for girls and show little heterogeneity with respect to parental background characteristics.

The results of our study contribute to the literature in several ways. Early work on the effects of multi-grade teaching often fails to identify causal effects because of selection into multi-grade classes (Veenman, 1995; Mason and Burns, 1996).[1] To address this issue, Sims (2008) made use of an instrumental variable strategy based on class size caps imposed by the California Class Size Reduction Program and shows that multigrade classes negatively affect test scores in Grades 2 and 3. Relying on survey data and comparing non-random but observationally equivalent single-grade and mixed-age classes in Sweden, Lindström and Lindahl (2011) report a sizable negative impact. Recent work by Leuven and Ronning (2016) has made use of discontinuous grade mixing rules in Norwegian junior high schools (grade 7-9). Their results document positive effects of multi-grade teaching on young students, but negative effects on more mature students within a class. Using a minimum class size rule in Italy which leads to multi-grade classes, Checchi and De Paola (2018) find negative effects of multi-grade teaching on numeracy of fifth-graders. Our results add to this small number of studies that report causal evidence on the impact of multi-grade classes by using the setting of staggered German state reforms to identify the causal impact of multi-grade classes on performance of fourth-graders in Germany. In light of the ongoing debate among German education scientists, this evidence sheds new light on the effects in various dimensions.

Our evidence on the short-run effects of multi-grade teaching in Germany complements the findings of a companion study on the long-term effects of multi-grade classes (Gerhardts *et al.*, 2021). In Gerhardts *et al.* (2021), we find that the abolition of denominational schools implied the abolishment of multi-grade classes in the German state *Saarland* in 1969. Using this setting, we show that multi-grade teaching has a causal negative impact on the students' educational and labour market outcomes measured in adult age, which is especially pronounced for women. The results presented here are consistent with these finding of negative effects of multi-grade classes lasting into adulthood and document that the negative effects can be traced to the exposure to multi-grade teaching during primary school.

The remaining part of the paper is structured as follows. Section 2 describes the institu-

[1] For a comprehensive overview of the literature on multi-grade classes, see also Gerhardts *et al.* (2021).

tional background. Section 3 provides details on the two data sources we use and presents our identification strategy. Section 4 presents the main empirical results and shows the results of the subgroup analysis. Section 5 discusses the results of several robustness tests. Section 6 concludes.

2 Institutional Background

2.1 The German School System

In Germany, education policy is the responsibility of federal states. This implies that each of the country's 16 federal states is solely responsible for its respective school system. Although differences exist across states, the general structure is still rather uniform. Before school, the large majority of children attends kindergarten. While only about 35% of children aged 1–3 years receive day care, 92.5% of children aged 3–6 attend kindergarten or receive another form of day care.[2] Usually at the age of six, children are enrolled in primary school. After four years at primary school, i.e., typically at age 10, the school system tracks children into three secondary school tracks: lower secondary school (Hauptschule), intermediate secondary school (Realschule), and high track grammar school (Gymnasium) in which students attain the university entrance qualification (Abitur).[3] The selection into a particular track is based on ability. Teachers in primary school recommend the highest school track they deem to be suitable for the child.[4] In light of this, our analysis makes use of fourth-graders' test scores as well as of teacher recommendations for the track in secondary schools as outcomes for the analysis of the effects of multi-grade teaching on educational outcomes and opportunities.

2.2 The Reforms under Study

2.2.1 Reasons for the Introduction

In 1997, the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (*Kultusministerkonferenz*) discussed several national and international studies which showed that children with low socio-economic backgrounds were disadvantaged in the German school system (Wagener, 2014). Furthermore, since the 1990s an increase in the heterogeneity of abilities and skills at the time of enrollment was observed. As a result, among others, for 8–12% of children at school entrance age, enrollment to primary school

[2] See `https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Soziales/Kindertagesbetreuung/_inhalt.html`.

[3] In the city state Berlin and in Brandenburg, primary school lasts for six years before selection into higher tracks.

[4] In some federal states, this recommendation acts as a limit for the schooling available to the child. Parents are subsequently responsible for choosing their child's secondary school track from the (limited) set of available school tracks.

was postponed by one year. To counteract on this development, the ministers of education agreed on lowering the school entry age and introducing the so-called flexible school entry stage. This concept would have implied that children could enroll in primary school also without a school entry examination certifying their school readiness[5] and that their time in grade 1 and grade 2 could be set individually. Children of both grades would be mixed and some of them would stay in this stage for one year and others for up to three years. The pedagogical arguments for such a change included that the flexible entry stage allowed to provide special support to children at the beginning of their academic education. The hope was that multi-grade classes could take into account the substantial heterogeneity in abilities, social background, and interest of enrolling children. This was supposed to happen through the optimal support of both high and low achieving students by giving the former the possibility to keep pace with students in second grade and hereby foster their intrinsic motivation and supporting the latter via a more intense education as well as giving them more room for their personal development. The goal of the reform was thus to smooth out potential differences in knowledge and skills directly at the beginning of the school careers in order to ensure a good knowledge in basic skills for all children.

However, in the end not all states introduced this flexible entry stage as a mandatory system. Almost all states implemented pilot projects to test the idea of having a multi-grade classroom at the beginning of primary school. In sum, the flexible school entrance stage was present in approximately 20 percent of all primary schools nation-wide since the first reforms.

2.2.2 Reactions to the Reforms

Politicians, teachers, and parents reacted all very differently to the introduction of the flexible school entry stage. On the positive side, some hoped that age heterogeneous learning groups could especially help pupils with lower skills or knowledge, because older classmates explain topics more intuitively and less abstract than teachers. It was argued that also more advanced children could benefit from these learning groups. Furthermore, the lower level of competition among pupils due to different tasks and lacking comparison of their grades was regarded as a beneficial development.

On the negative side, however, the flexible school entry age implied more effort and preparation time for teachers. It turned out to be more difficult to teach a class with a more heterogeneous age and skill structure. Often, older pupils cannot or do not want to help their younger classmates because of lacking empathy or patience (Heinzel and Koch, 2017).

[5] The school entrance examination is a mandatory medical screening meant to promote children's health by diagnosing medical anomalies and providing necessary treatment as early as possible. It is conducted by pediatricians employed by the local health service who document children's development including their "readiness" for primary school. The examination takes place in the year prior to entering primary school when children are on average six years old.

To the best of our knowledge, no causal evaluations of the flexible school entrance stage in Germany have been conducted. The existing descriptive studies do not provide a clear result on whether multi-grade classrooms at the beginning of primary school have any beneficial effects and whether they reduce educational inequality (Helbig and Nikolai, 2015).

2.2.3 Reform Timing

After the *Kultusministerkonferenz* in 1997, all states except of one (the *Saarland*) introduced multi-grade teaching in a flexible school entrance system in some pilot schools. Yet, as Table 1 shows, only few states introduced a flexible school entrance stage as a mandatory system.

Saxony-Anhalt, in 2000, was the first state to introduce a mandatory flexible entry stage. Berlin followed in 2005 with a state-wide mandatory implementation, but reintroduced the choice between the traditional system and the multi-grade approach in 2010. Thuringia made the flexible school entrance mandatory from 2008 onwards and Brandenburg followed in 2010. All of the other states did not introduce a mandatory system of early multi-grade teaching. However, some of them made it optional for schools to establish a flexible school entrance stage: Bremen, Hesse, Lower-Saxony, and Berlin. The rest of the 16 states decided - after experimenting in some pilot schools - against a broader implementation of the flexible school entrance.

3 Data and Empirical Strategy

3.1 Data

Our analysis combines two data sources, which enables us to produce a longitudinal dataset that is representative for fourthgraders' performance and motivation in German primary schools. The two data sources are the *Progress in International Reading Literacy Study* (PIRLS) in 2001 and 2006, and the National Assessment Study in 2011 and 2016 by the *German Institut zur Qualitätsentwicklung im Bildungswesen* (IQB) (Institute of Quality Development in the Education System). Both data sources have in common that they provide state identifiers. Those are necessary for the linkage with our reform data.[6] The first source, the extended PIRLS assessment in Germany, not only includes reading test scores, but also students' grades in German, the recommendation for the next school track, as well as students' school enjoyment. We make also use of the information available on student and family background in order to control for factors that may impact students' education outcomes. In a robustness check, we also use available teacher and school characteristics as control variables. The second source, the German National Assessment Study, also assesses reading test scores of fourth-graders, comparable to the PIRLS

[6] Note that the PIRLS data for the years 2011 and 2016 do not include state identifiers anymore. Therefore, we have to rely on the IQB data. Since the IQB studies are very similar to the design of PIRLS, however, the combination of both data sources is possible.

Table 1: Multigrade Teaching in Flexible School Entrance Levels - Reform Overview

State	**Year**	**Flexible School Entrance Reform**	**Mandatory**	**Optional**
Baden-Wurttemberg	Since 1997	Model Projects in 82 schools	none	none
Bavaria	2010-2014	Model Projects in 20 schools	none	none
	2017	Number of schools gradually extended until 2017		
	2019	216 schools		
Berlin	Since 1992	Model Projects in 340 schools	cohort 2011	cohorts 2011,2016
	2005	State-wide implementation		
	2010	Choice between FSE and Traditional System		
Brandenburg	1992-1995	Pilot Projects in 2 schools	cohort 2016	cohort 2016
	1999-2002	Model Projects in 2 schools		
	2000-2004	Extension to 20 schools		
	2003	Extension to 139 schools		
	2010	State-wide implementation		
Bremen	1993-1995	Model Projects in 2 schools	none	cohorts 2011,2016
	2005	Optional for all primary schools		
Hamburg	1994-1996	Model Projects in 2 schools	none	none
Hesse	1994-1998	Model Projects in 6 schools	none	cohorts 2011,2016
	1998-2004	Extension to 29 schools		
	2007	Optional for all primary schools		
Lower-Saxony	1994-2002	Model Projects in 10 schools	none	cohorts 2006, 2011,2016
	2003	Optional for all primary schools		
Mecklenburg-Vorpommern	2005-2007	Model Projects in 16 schools	none	none
	2019	Optional in all primary schools		
North Rhine-Westphalia	1999-2004	Model Projects in 6 schools	none	none
Rhineland-Palatinate	1995-1998	Model Projects in 2 schools	none	none
		Gradual Extension to 20 schools		
Saarland		no flexible entrance reforms	none	none
Saxony	2001-2004	Model Projects in 25 schools	none	none
Saxony-Anhalt	1997-2000	Model Projects in 4 schools	cohorts 2006, 2011,2016	cohorts 2006, 2011,2016
	2000	State-wide implementation		
Schleswig-Holstein	1994-1997	Model Projects in 5 schools	none	none
		Gradual Extension to 12 schools		
Thuringia	1997	Optional in all primary schools	cohort 2016	cohort 2016
	1999-2003	Model Projects in 14 schools		
	2003-2008	Transfer Projects in 25 schools		
	Since 2008	Gradually region-wide implementation		

Own collection of information in legal documents and websites of the states' education ministries. The fourth and fifth columns indicate whether fourthgraders of the respective cohorts in the given state are part of a mandatory resp. optional flexible school entrance system. The category "optional" includes both mandatory and optional flexible school entrance rules.

and takes place at the end of primary school. It includes also information on the other outcomes of interest as well as the control variables in the same way as the PIRLS data. Our final sample thus comprises students in their fourth grade in 2001, 2006, 2011, or 2016; and who entered their first school year in 1998, 2003, 2008, or 2013 respectively. The combined data yields a sample of approximately 68,000 students.

3.2 Empirical Model

The combination of the different reforms in the various states with cross-sectional data of outcomes for four cohorts of fourth-graders (2001, 2006, 2011, and 2016) implies the following research design: a cohort of fourth-graders was part of a flexible school entrance system if the reform had been in place when the cohort entered first grade. For example, since Saxony-Anhalt introduced a mandatory school entrance stage in 2000, the cohort 2001 was not treated by the reform yet, but the cohorts 2006, 2011, and 2016 were treated. This is shown in the fourth column of Table 1 for every state. The fifth column shows the affected cohorts when we are not only considering mandatory flexible school entrance systems, but in addition all states that made it optional for schools. We use the *mandatory* definition for our main analysis of the effects of the flexible school entrance system. However, we use the *optional* definition in robustness checks.

We use the staggered implementation of the flexible entrance stage across German states to estimate the effect of multi-grade classrooms in a difference-in-differences framework. This approach exploits the variation in the exposure to a multi-grade class (i) across reforming and non-reforming states and (ii) between affected and unaffected cohorts within the same federal state. Our main specification is thus given as follows:

$$Y_{i,s,t} = \beta_0 + \beta_1 multigrade_{s,t} + X_i\beta_2 + \mu_s + \mu_t + \epsilon_{i,s,t}. \tag{1}$$

where $Y_{i,s,t}$ is the outcome variable for student i in cohort t attending school in state s. The dummy variable *multigrade* equals 1 for the treated states in the treatment period. In our baseline analysis we only define those states as treated if they introduced a *mandatory* flexible school entrance stage. This has the advantage that all students of a cohort who got enrolled in primary school during a treatment period were certainly experiencing a multi-grade setting during their first school years. The disadvantage is that the observed students in the control states could have been also treated if their states had an optional rule regarding the flexible school entry stage (see Table 1). This implies a mis-classification of treatment and control and might lead to a bias in the estimates towards zero. Therefore, as a robustness check we define all states as being treated which introduced mandatory or *optional* multi-grade classes. Since, however this latter definition does imply that very probably not all students in the treatment group are actually treated, it rather has to be interpreted as an intention-to-treat effect.

The vector X_i includes a set of control variables to account for students' demographic characteristics. Our baseline analysis controls for gender and age, kindergarten attendance, migrant background, and parental education. In a robustness check, we control for books at home instead of parental education. In further robustness checks, we additionally control for teacher and school characteristics.

State fixed effects μ_s control for time-invariant conditions in each state, including state capacity, local culture, or geography. Cohort fixed effects μ_t capture national trends in student cohorts' demographic composition, as well as general trends in the education sector or the labor market. $\epsilon_{i,s,t}$ is the error term. We cluster the standard errors at the state level as the treatment varies as the state level. Considering recent developments in the econometric literature we calculate p-values of two different types of clustering methods for each reform coefficient displayed in our tables. First, we use the standard clustering method which is conservative in our kind of setting and accounts for potential correlation of error terms across years within states (Athey and Imbens, 2018). Second, to account for the limited number of clusters (because there are only 16 German states) we calculate wild cluster bootstrap p-values (Roodman *et al.*, 2019).

Under the assumptions of the difference-in-differences framework, the coefficient β_1 represents the causal effect of the reform. Most importantly, the common trend assumption implies that - in absence of the treatment - reforming and non-reforming states would both lie on the same trend with respect to outcome variables. It is typically argued that this assumption is likely to be fulfilled if the pre-trends prior to the reform are the same in reforming and non-reforming states. Since our data only covers four points in time it is not possible to investigate pre-trends of the outcomes. A specific feature of our main analysis is that only states in East Germany introduced mandatory flexible school entry stages. We therefore, in a further specification, restrict our sample to only East German states. This makes it even more likely that control and treatment states have common trends. Interestingly (and reassuringly), the results do not differ much from the main specification.

A second crucial assumption of our identification strategy is that the treatment effect does not represent any development simultaneously occurring to the multigrade reforms. To avoid this problem, we investigate whether other education reforms affecting primary school students were simultaneously introduced. Indeed, a reform abolishing numerical grades in the first years of primary school has been introduced in some of the states during a similar time frame, yet with a different timing pattern across states. We test the robustness of our results by controlling for the early grading reform, and show that our results are not affected.

As described in Section **??**, a major reason for the introduction of the reform was the heterogeneous school readiness of children at the beginning of primary school. If children with a lower school readiness stayed longer in kindergarten before the reform, but reduced time in kindergarten after the reform due to the integrative approach of the flexible school entrance level, this would threaten our identification strategy. It is well studied in the literature that years in kindergarten have a positive effect on child development and school performance. If the reform reduced years in kindergarten this could lead to a negative result, but the teaching in multi-grade classes would not be the cause for it. Therefore, we use years in kindergarten as placebo outcome. We find that the reform did not affect time in kindergarten.

Finally, since years in kindergarten could lead to being better prepared for following a multi-grade class, we interact the reform with kindergarten years, and indeed find heterogeneous effects.

To explore whether girls and children from a low socio-economic background are affected in different ways by being taught in a multi-grade classroom, we perform separate analyses for these subgroups and study heterogeneous effects by gender and by parental education.

3.3 Descriptives

Table 2 shows the descriptive statistics of our dataset.

Table 2: Descriptive Statistics

	Mean	St.Dev	Min	Max	Observations
Panel A: Reform					
Mandatory Multi-grade	0.12	0.32	0.00	1.00	72873
Optional Multi-grade	0.31	0.46	0.00	1.00	72873
Panel B: Outcomes					
Standardized Reading Testscore	0.00	1.00	-4.46	3.67	68453
German Grade (1=lowest, 6=highest)	4.47	0.90	1.00	6.00	63953
Recommendation for Gymnasium (Dummy)	0.34	0.47	0.00	1.00	69345
Enjoy School (Dummy)	0.68	0.46	0.00	1.00	55009
Panel C: Student Controls					
Student is a girl	0.49	0.50	0.00	1.00	72380
Age of student (in years)	10.47	0.50	6.42	12.92	72346
Low parental education (Dummy)	0.20	0.40	0.00	1.00	72873
Books at Home (1=(<10) to 5=(>200))	3.33	1.20	1.00	5.00	63233
First generation migrant	0.06	0.23	0.00	1.00	72873
Years spent in kindergarten	3.30	0.97	0.00	5.50	72873
Panel D: Teacher Controls					
Age of teacher (in years)	46.91	10.31	24.00	72.00	63578
Experience of teacher (in years)	20.83	12.31	0.00	57.00	63754
Teacher specialized in German (Dummy)	0.81	0.39	0.00	1.00	64116
Teacher works full-time (Dummy)	0.72	0.45	0.00	1.00	64456
Panel E: School Controls					
No. of students enrolled in school	276.48	151.35	12.00	2008.00	67490
School is a public School (Dummy)	0.96	0.20	0.00	1.00	69412
Experience as headmaster in this school (in years)	9.14	7.12	0.00	42.00	65712
Age of headmaster in years (in years)	52.34	7.42	22.00	71.00	66223
Headmaster is male (Dummy)	0.32	0.47	0.00	1.00	67978

Notes: The table shows the descriptive statistics of a quasi-panel of fourthgraders using data from the PIRLS assessment and the German National assessment (IQB) for the years 2001, 2006, 2011, and 2016.

Reform Variables. As described in Section **??** not all reforming states made the flexible school entrance stage, which introduced multi-grade classes, a mandatory policy for their schools. Linking the reform data from Table 1 to the students observed in our data, it shows that 12% of students experienced a system of a mandatory flexible school entrance stage, see Panel A of Table 2. In our main analysis we use this definition of being treated by the reform. The second row of Table 2 shows that our alternative definition of the treatment, i.e. being treated if the state has introduced an optional or mandatory flexible school entrance stage, leads to 31% of students belonging to the treatment group.

Outcome Variables. *Reading test scores.* Students' reading test scores are measured by the standardized reading tests provided by the PIRLS resp. IQB study. The test scores from all datasets are originally constructed to have a mean of 500 and standard deviation of 100, thereby facilitating nation-wide comparison. They are z-standardized for the purpose of this study (see first row of Panel B of Table 2).

Grades We use the information on the last grade student received for their performance in German. They are graded according to the German grading scale, which varies from 1 (*excellent, sehr gut*) to 6 (*insufficient, ungenügend*). We inverse the scale for the purpose of readability. The second row of Panel B of Table 2 shows that students in our dataset receive on average a grade between "good" and "satisfactory".

Recommendation for Gymnasium. As described in Section **??**, in Germany students are tracked into three differents tracks after primary school. In fourth grade, they receive a recommendation by their teacher on which track would be most suitable given the student's ability. We use this information to create a dummy variable indicating whether a student is recommended to enroll in the highest track (Gymnasium). The data show that a third of students in our sample receive this recommendation.

Enjoy going to school. Students were asked to what extent they agree that going to school is enjoyable for them. The answers include the four categories "strongly agree", "somewhat agree", "neither agree nor disagree", and "strongly disagree". We create a dummy variable for enjoying going to school, which takes value 1 if the student strongly or somewhat agrees, and 0 otherwise.

Control Variables. *Student Controls.* Panel C of Table 2 shows the individual controls we use in our main analysis. In our dataset, 49% of students are female. The average age is 10.47 years, which is the usual age of fourthgraders. 20% of students have low educated parents, i.e. their parents have at most a lower secondary degree. As an alternative measure for socio-economic background we use the number of books at home as proxy for parental educational background in a robustness check. On average, children's families have a bit more than 100 books at home (as category 3 contains 26-100 books, and category 4 contains 101-200 books). 6% of the students are first generation immigrants. On average, the students have spent 3.3 years in kindergarten prior to primary school.

Teacher Controls. Panel D of Table 2 shows teacher characteristics which we use as controls in a robustness check. On average, teachers are 46 years old and have 20 years of experience in the teaching profession. 81% of teachers are specialized in teaching German, and 72% of them work full-time.

School Controls. Finally, Panel E of Table 2 displays the descriptive statistics of school characteristics. As in case of the teacher controls, we use these as control variables in a robustness check. On average, the primary schools under study have 276 enrolled students. 96% of the schools are public schools (note that private schools are uncommon in Germany). The headmasters of the respective schools are on average 52 years old, have 9 years of experience as headmaster, and 32% of them are male.

4 Results

4.1 Main Results

Table 3 shows the results of estimating Equation 1 with reading test scores of fourthgraders as outcome variable. In each column, we add one of the individual control variables. The specification in column (1) only uses state and cohort fixed effects. The multigrade coefficient is negative, but small and not significant. Column (2) adds gender and age as controls, which leads to a larger multigrade coefficient, which is still not significant, however. Interestingly, when adding years spent in kindergarten as control variable in column (3), the multigrade coefficient is significant at the 1%-level and equals -7.6% of a standard deviation. Adding a control for being a first generation immigrant in column (4) and having low-educated parents in column (5) leaves the multigrade coefficient significant and economically meaningful. According to the estimation results in column (5), being in a cohort which experienced reform-induced multi-grade teaching in the first years of primary school leads to a decline in reading test scores of 6.1% of a standard deviation. The specification of model (5) serves as our main specification in the next steps of the analysis as it has the highest explanatory power (measured by R^2).

Table 4 shows the main effects of early grading on two other achievement measures – the most recent grade in German and the recommendation for the high track, as well as on the motivational outcome – measured as enjoying school. Column (1) displays the negative effect on reading testscores described above. Column (2) shows that the multigrade reform had also a significant negative effect on the grade in German which equals approximately 1/9 of a standard deviation (0.108 divided by the sample standard deviation 0.9, see Table 2). Despite the negative

Table 3: Effect of Multigrade Class on Reading Test Scores of Fourth Graders

	Reading Skills				
	(1)	(2)	(3)	(4)	(5)
Multigrade	-0.019	-0.041	-0.076***	-0.094***	-0.061***
	(0.029)	(0.040)	(0.017)	(0.024)	(0.020)
Female		0.169***	0.160***	0.157***	0.163***
		(0.010)	(0.010)	(0.010)	(0.010)
Age		-0.446***	-0.410***	-0.388***	-0.354***
		(0.019)	(0.016)	(0.017)	(0.016)
Kindergarten Attendance (years)			0.130***	0.117***	0.098***
			(0.014)	(0.013)	(0.011)
Migration Background				-0.337***	-0.342***
				(0.026)	(0.024)
Low SES					-0.430***
					(0.029)
State FE	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes
R-squared	0.077	0.135	0.167	0.180	0.203
WCB P-Value	0.625	0.638	0.072	0.145	0.153
Observations	68,453	68,453	68,453	68,453	68,453

*Notes:*The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

impact on both test scores and grades, neither the high-track recommendation nor enjoyment of school are significantly affected by the multigrade reform, even though the coefficients in columns (3) and (4) are also negative.

4.2 Effect Heterogeneity

4.2.1 Boys & Girls

Earlier findings by Leuven and Ronning (2016) and Gerhardts *et al.* (2021) indicate heterogeneity of effects of multi-grade teaching by gender and parental education. The findings of Gerhardts *et al.* (2021) suggest that the negative effect of multi grade classes is stronger for girls than for boys, and document a more pronounced negative effect on children of blue-collar workers.

Table 5 shows that girls are significantly negatively affected by multigrade classes in terms of their reading test scores (-0.08), their grades in German (-0.123) and their enjoyment of school (-0.025). Boys, on the contrary, do not seem to be as harmed by being taught in a multigrade classroom. The effect on their reading test scores are smaller (-0.04) and not significant, and whether they enjoy going to school is also not affected. Boys' grades in German, however, are significantly affected, but less than in the case of girls (-0.092).

Table 4: Effect of Multigrade Class on Further Outcomes of Fourth Graders

	Reading Test Score	German Grade	High-Track Recommendation	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.061***	-0.108***	-0.046	-0.019
	(0.020)	(0.034)	(0.038)	(0.011)
Female	0.163***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.354***	-0.391***	-0.144***	-0.010**
	(0.016)	(0.010)	(0.012)	(0.004)
Migration Background	-0.342***	-0.263***	-0.076***	0.023***
	(0.024)	(0.020)	(0.014)	(0.006)
Kindergarten Attendance (years)	0.098***	0.081***	0.038***	0.000
	(0.011)	(0.006)	(0.005)	(0.003)
Low SES	-0.430***	-0.397***	-0.204***	-0.024**
	(0.029)	(0.016)	(0.016)	(0.008)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.203	0.195	0.138	0.044
WCB P-Value	0.167	0.059	0.699	0.376
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Consequently, these subgroup results are in line with the evidence of our study on the reforms in the Saarland several decades before.

4.2.2 Parental Education Background

Table A.1 in the Appendix shows that children with high parental education are significantly negatively affected by multigrade classes in terms of their reading test scores (-0.088) as well as their grades in German (-0.123). Surprisingly, children with low parental education are not significantly affected by the multigrade reform, the reform coefficient is negative but rather small (-0.021), see Table A.2 in the Appendix. Their grades in German, however, are significantly affected, but less than in the case of children with high-educated parents (-0.091). There is no significant effect on the high-track recommendation or the enjoyment of school for neither of both groups.

Finding worse results for children with more advantaged family backgrounds is in contrast to our findings on the effects of the reform in the Saarland (Gerhardts *et al.*, 2021). In some way, the result indicates that educational inequality could decrease due to the multigrade classes. Unfortunately, this seems to come at the cost of deteriorating skills of the more advantaged

Table 5: Effect of Multigrade Class on Girls

	Reading Test Score (1)	German Grade (2)	High-Track Recommendation (3)	Enjoy School (4)
Multigrade	-0.080***	-0.123***	-0.057	-0.025*
	(0.026)	(0.028)	(0.038)	(0.013)
Age	-0.358***	-0.384***	-0.141***	-0.018**
	(0.018)	(0.012)	(0.013)	(0.007)
Migration Background	-0.319***	-0.234***	-0.062***	0.008
	(0.028)	(0.027)	(0.017)	(0.013)
Kindergarten Attendance (years)	0.096***	0.079***	0.037***	-0.001
	(0.011)	(0.007)	(0.005)	(0.003)
Low SES	-0.442***	-0.417***	-0.218***	-0.024**
	(0.026)	(0.021)	(0.017)	(0.010)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.188	0.174	0.141	0.024
WCB P-Value (Reform)	0.089	0.055	0.537	0.333
N	33,763	31,541	34,144	27,363

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of female 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

group of students rather than through more enhanced skills of the disadvantaged group.

5 Robustness Checks

5.1 Years in Kindergarten as Placebo Outcome

In this section, we test the robustness of our main results presented in Section 4 and discuss some of the assumptions explained in Section 3.2 in more depth.

A major reason for the introduction of the reform was the heterogeneous school readiness of children at the beginning of primary school. If children with a lower school readiness stayed longer in kindergarten before the reform, but reduced time in kindergarten after the reform due to the integrative approach of the flexible school entrance level, this would threaten our identification strategy. Therefore, we use years in kindergarten as placebo outcome. Table A.3 in the Appendix shows that there is no effect of the reform on time spent in kindergarten. Looking at the control variables, the familiar pattern of socio-economic selection into kindergarten is visible. Both migrant background as well as low parental education are negatively associated with the intensive margin of kindergarten attendance.

Table 6: Effect of Multigrade Class on Boys

	Reading Test Score (1)	German Grade (2)	High-Track Recommendation (3)	Enjoy Reading (4)
Multigrade	-0.040	-0.092**	-0.036	-0.014
	(0.035)	(0.042)	(0.039)	(0.019)
Age	-0.352***	-0.397***	-0.146***	-0.002
	(0.017)	(0.013)	(0.012)	(0.005)
Migration Background	-0.364***	-0.287***	-0.089***	0.035***
	(0.030)	(0.020)	(0.014)	(0.008)
Kindergarten Attendance (years)	0.100***	0.082***	0.039***	0.002
	(0.013)	(0.007)	(0.004)	(0.004)
Low SES	-0.420***	-0.376***	-0.189***	-0.026**
	(0.036)	(0.019)	(0.015)	(0.010)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.204	0.174	0.132	0.016
WCB P-Value (Reform)	0.545	0.005	0.801	0.569
N	34,690	32,412	35,201	27,646

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of male 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

5.2 Interaction of Reform with Years in Kindergarten

As a longer preparation for school children receive in kindergarten could enable them to cope with the situation in multigrade classes better, we investigate the interaction of the reform with years spent in kindergarten. Table A.4 in the appendix shows that the interaction is significantly positive. This implies that spending more years in kindergarten before primary school mitigates the negative effect of being taught in a multigrade classroom. Interestingly, adding the interaction shows that children who spend less time in kindergarten not only experience significant negative effects in terms of their test scores and grade, but also in terms of their high-track recommendation and school enjoyment (columns (3) and (4)).

5.3 Controlling for Another Reform in Primary Schools

A second crucial assumption of our identification strategy is that the treatment effect does not represent any development simultaneously occurring to the multigrade reforms. To avoid this problem, we investigate whether other education reforms affecting primary school students were simultaneously introduced. Indeed, a reform abolishing numerical grades in the first years of primary school has been introduced in four of the states during a similar time frame, yet with a different timing pattern across states (Hesse in 1999, Saarland in 2000, Brandenburg in 2001, Berlin in 2006). We test the robustness of our results by controlling for the early grading reform in Table A.5 in the Appendix. The table shows that our results are not affected.

5.4 Sample Restricted to East German States

A specific feature of our main analysis is that only states in East Germany introduced mandatory flexible school entry stages. We therefore, in a further specification, restrict our sample to only East German states. This makes it even more likely that control and treatment states have common trends. Table 7 shows that the results are robust and do not differ much from the main specification.

Table 7: Effect of Multigrade Class on Students' Outcomes - East Germany

	Reading Test Score	German Grade	High-Track Recommendation	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.057**	-0.061***	-0.029	-0.018
	(0.020)	(0.013)	(0.017)	(0.010)
Female	0.196***	0.292***	0.027**	0.146***
	(0.013)	(0.015)	(0.009)	(0.008)
Age	-0.382***	-0.386***	-0.122***	-0.015
	(0.029)	(0.019)	(0.023)	(0.008)
Migration Background	-0.288***	-0.216***	-0.026	0.040***
	(0.041)	(0.038)	(0.023)	(0.009)
Kindergarten Attendance (years)	0.082***	0.087***	0.027**	0.006
	(0.018)	(0.009)	(0.007)	(0.004)
Low SES	-0.411***	-0.349***	-0.171***	-0.039**
	(0.083)	(0.033)	(0.024)	(0.015)
Student Controls	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.186	0.180	0.137	0.044
WCB P-Value	0.320	0.011	0.443	0.366
Observations	24,366	24,681	25,177	21,094

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of only East German 4th-grade students using data from the PIRLS assessment and a German National assessment (IQB). Reading scores are z-standardized. Standard errors are clustered at the state level. Grades are on a scale from 1 ("insufficient") to 6 ("very good"). High track is a dummy being one if teacher recommends high track. Enjoy school is a dummy equaling one if student fully agrees to enjoy going to school. Different number of observations due to differing availability of outcome variable. * denotes statistical significance at the 10% level, ** at the 5% level and *** at the 1% level.

5.5 Parental Background Control

Family background is often measured by parental education, as we do in our main specification. However, many studies have shown that also the variable "books at home" is a very reliable proxy for the socio-economic status of a family. Therefore, in a robustness check, we control for this variable instead of parental education. The results of both specifications are very similar, as Table A.6 in the Appendix shows.

5.6 Teacher and School Characteristics as Further Controls

To alleviate the assumption of common trends of treated and control states a little bit we add teacher and school characteristics as control variables. If the composition of teachers or organizational patterns of the schools changed along with the multigrade reforms, adding these controls would make a difference for our estimates. Table A.7 in the Appendix shows no important differences in comparison to our main specification, however. In addition, the explanatory power (R2) does not increase much by adding these further controls.

5.7 Definition of Treatment Status

Finally, we check the robustness of our results with respect to the definition of the treatment status of the cohorts in our sample. In our baseline analysis we only define those states as treated which introduced a *mandatory* flexible school entrance stage. This has the advantage that all students of a cohort who got enrolled in primary school during a treatment period were certainly experiencing a multi-grade setting during their first school years. The disadvantage is that the observed students in the control states could have been also treated if their states had an optional rule regarding the flexible school entry stage (see Table 1) and they happen to be in a school that makes use of this option.[7] This is likely to lead to a downward bias in our estimates. Therefore, as a robustness check we define all states as being treated which introduced mandatory or *optional* multi-grade classes. Table 8 shows that the effects on reading test scores and grades stay significant using the new definition, the coefficients are (in absolute terms) larger which is in line with the argument stated above – moving the states with optional flexible school entrance systems to the treatment group removes all potentially treated observations out of the control group.

6 Conclusion

This paper provides novel evidence about the impact of exposure to multi-grade teaching in primary school on educational outcomes. The results of a difference-in-differences approach that exploits the staggered implementation of flexible school entrance levels across German states between 1997 and 2010 reveal a significant negative effect of multi-grade teaching on educational outcomes such as reading test scores and grades, but no effect on teacher recommendations or subjective perceptions of pupils. This partly rationalizes the mixed evidence in the literature by documenting that multi-grade teaching does not exhibit negative effects throughout. Instead, the effects emerge for skills that can be measured in comparable metrics. The effects are more pronounced for girls, complementing earlier evidence from other studies in different

[7] Note that the assignment to primary schools is based on school catchment areas in Germany, so that sorting to schools dependent on whether they introduced a flexible school entrance stage is not possible.

Table 8: Effect of Optional Multigrade Class on Students' Outcomes

	Reading Test Score (1)	German Grade (2)	High-Track Recommendation (3)	Enjoy School (4)
	Reading Test Score	German Grade	High-Track Recommendation	Enjoy School
Optional Multigrade	-0.086*	-0.100*	-0.044	-0.023
	(0.047)	(0.050)	(0.040)	(0.019)
Female	0.163***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.354***	-0.389***	-0.143***	-0.010**
	(0.016)	(0.010)	(0.012)	(0.004)
Migration Background	-0.342***	-0.264***	-0.076***	0.022***
	(0.024)	(0.021)	(0.014)	(0.006)
Kindergarten Attendance (years)	0.099***	0.081***	0.039***	0.000
	(0.012)	(0.007)	(0.005)	(0.003)
Low SES	-0.430***	-0.398***	-0.204***	-0.024***
	(0.030)	(0.015)	(0.015)	(0.008)
Student Controls	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.203	0.195	0.138	0.044
WCB P-Value	0.125	0.101	0.285	0.295
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and a German National assessment (IQB). Treatment is going to school in a state which introduced mandatory *or* optional multigrade classes. Reading scores are z-standardized. Standard errors are clustered at the state level. Grades are on a scale from 1 (”insufficient”) to 6 (”very good”). High track is a dummy being one if teacher recommends high track. Enjoy school is a dummy equaling one if student fully agrees to enjoy going to school. Different number of observations due to differing availability of outcome variable. * denotes statistical significance at the 10% level, ** at the 5% level and *** at the 1% level.

contexts. The evidence also shows that spending more years in kindergarten before primary school mitigates the negative effects of exposure to multi-grade teaching.

In light of these findings, more work is needed to reveal the mechanisms underlying these effects.

References

Athey, S. and Imbens, G. W. (2018). Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. *NBER Working Paper*, **24963**.

Carle, U. and Metzen, H. (2014). Literaturbersicht zum Stand derForschung, der praktischen Expertise und der pädagogischen Theorie. Eine wissenschaftliche Expertise des Grundschulverbandes. Frankfurtam Main: Grundschulverband (Wissenschaftliche Expertisen. In *Grundschulverband: Wissenschaftliche Expertisen*. Grundschulverband, Frankfurt am Main.

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

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

Heinzel, F. and Koch, K. (2017). *Individualisierung im Grundschulunterricht*. Springer, Wiesbaden.

Helbig, M. and Nikolai, R. (2015). Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949. page 383.

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

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

Little, A. W. (2004). Learning and teaching in multigrade settings. *Paper commissioned for the EFA Global Monitoring Report 2005, The Quality Imperative*.

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

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

Roodman, D., Nielsen, M. O., MacKinnon, J. G., and Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal*, **19**(1), 4–60.

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

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

Wagener, M. (2014). *Gegenseitiges Helfen. Soziales Lernen im jahrgangsgemischten Unterricht*. Springer, Wiesbaden.

A Appendix

Table A.1: Effect of Multigrade Class on Children with High Parental Education

	Reading Test Score	German Grade	High-Track Recommendation	Enjoy Reading
	(1)	(2)	(3)	(4)
Multigrade	-0.088***	-0.123***	-0.048	-0.015
	(0.028)	(0.040)	(0.036)	(0.011)
Female	0.175***	0.269***	0.033***	0.150***
	(0.011)	(0.009)	(0.005)	(0.007)
Age	-0.370***	-0.398***	-0.150***	-0.008
	(0.017)	(0.012)	(0.013)	(0.005)
Migration Background	-0.367***	-0.294***	-0.092***	0.024***
	(0.026)	(0.022)	(0.016)	(0.007)
Kindergarten Attendance (years)	0.105***	0.083***	0.043***	0.002
	(0.012)	(0.006)	(0.005)	(0.003)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.188	0.195	0.142	0.041
WCB P-Value (Reform)	0.163	0.056	0.677	0.410
N	53,927	50,407	55,197	43,572

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students with high-educated parents using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table A.2: Effect of Multigrade Class on on Children with Low Parental Education

	Reading Test Score (1)	German Grade (2)	High-Track Recommendation (3)	Enjoy Reading (4)
Multigrade	-0.021	-0.091***	-0.011	-0.011
	(0.040)	(0.023)	(0.028)	(0.019)
Female	0.121***	0.216***	0.011	0.157***
	(0.015)	(0.020)	(0.007)	(0.010)
Age	-0.293***	-0.359***	-0.120***	-0.017
	(0.014)	(0.015)	(0.012)	(0.011)
Migration Background	-0.247***	-0.147***	-0.018	0.011
	(0.029)	(0.032)	(0.014)	(0.017)
Kindergarten Attendance (years)	0.071***	0.072***	0.028***	-0.004
	(0.010)	(0.006)	(0.006)	(0.006)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.283	0.146	0.078	0.055
WCB P-Value (Reform)	0.596	0.042	0.720	0.571
N	14,526	13,546	14,148	11,437

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students with low-educated parents using data from the PIRLS assessment and the German National assessment (IQB). Reading test scores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table A.3: Effect of Multigrade Class on Placebo Outcome: Kindergarten Attendance

	Kindergarten Attendance				
	(1)	(2)	(3)	(4)	(5)
Multigrade	0.152	0.149	0.161	0.160	0.173
	(0.237)	(0.238)	(0.242)	(0.243)	(0.239)
Female		-0.030***	-0.021***	-0.023***	-0.019**
		(0.007)	(0.007)	(0.007)	(0.007)
Age		-0.040***	-0.065***	-0.039**	-0.021
		(0.013)	(0.014)	(0.014)	(0.013)
Migration Background				-0.484***	-0.480***
				(0.035)	(0.035)
Low SES					-0.213***
					(0.028)
State FE	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes
R-squared	0.274	0.274	0.288	0.300	0.307
WCB P-Value	0.650	0.614	0.610	0.603	0.558
Observations	72,873	72,873	72,873	72,873	72,873

*Notes:*The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). The placebo outcome used here is kindergarten attendance (years). Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table A.4: Effect of Multigrade Class on Students' Outcomes - Interaction term: Reform and time spent in kindergarten

	Reading Test Score (1)	German Grade (2)	High-Track Recommendation (3)	Enjoy School (4)
Multigrade	-0.079**	-0.132***	-0.072*	-0.047**
	(0.035)	(0.014)	(0.039)	(0.016)
Interaction Kindergarten	0.037	0.042*	0.036*	0.034**
	(0.021)	(0.023)	(0.018)	(0.014)
Kindergarten	0.166***	0.124***	0.042***	-0.012**
	(0.018)	(0.010)	(0.009)	(0.005)
Female	0.162***	0.257***	0.028***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.358***	-0.394***	-0.145***	-0.010**
	(0.016)	(0.011)	(0.012)	(0.004)
Migration Background	-0.370***	-0.287***	-0.089***	0.022***
	(0.025)	(0.019)	(0.014)	(0.006)
Low SES	-0.440***	-0.405***	-0.209***	-0.025***
	(0.032)	(0.017)	(0.016)	(0.008)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.200	0.192	0.135	0.044
WCB P-Value	0.193	0.140	0.083	0.126
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. Kindergarten is a dummy measuring one if a child spent more than 3 years in child care before school. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table A.5: Effect of Multigrade Class on Students' Outcomes - Controlling for Early Grading Reform

	Reading Test Score	German Grade	High-Track Recommendation	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.060***	-0.108***	-0.046	-0.020*
	(0.016)	(0.036)	(0.037)	(0.011)
Early Grading	0.054	-0.006	0.016	-0.020
	(0.043)	(0.083)	(0.038)	(0.034)
Female	0.163***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.355***	-0.391***	-0.144***	-0.009*
	(0.016)	(0.011)	(0.012)	(0.004)
Migration Background	-0.342***	-0.263***	-0.076***	0.023***
	(0.024)	(0.020)	(0.014)	(0.006)
Kindergarten Attendance (years)	0.099***	0.081***	0.038***	0.000
	(0.012)	(0.006)	(0.005)	(0.003)
Low SES	-0.430***	-0.397***	-0.204***	-0.024**
	(0.029)	(0.016)	(0.016)	(0.008)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.203	0.195	0.138	0.044
WCB P-Value	0.136	0.083	0.729	0.491
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). In addition to the main specification, we control for a reform which introduced early numerical grading in some of the German states between 1999 and 2006. Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table A.6: Effect of Multigrade Class on Students' Outcomes - ”Books at home” as Parental Background Control

	Reading Test Score	German Grade	High-Track Recommendation	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.082***	-0.120***	-0.053	-0.021*
	(0.018)	(0.032)	(0.037)	(0.011)
Female	0.155***	0.244***	0.025***	0.152***
	(0.009)	(0.009)	(0.004)	(0.006)
Age	-0.281***	-0.341***	-0.128***	-0.006
	(0.014)	(0.012)	(0.012)	(0.005)
Migration Background	-0.239***	-0.186***	-0.035**	0.028***
	(0.025)	(0.022)	(0.013)	(0.007)
Kindergarten Attendance (years)	0.071***	0.066***	0.031***	-0.002
	(0.009)	(0.005)	(0.004)	(0.003)
Books at home	0.249***	0.192***	0.092***	0.015***
	(0.011)	(0.006)	(0.006)	(0.002)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.245	0.213	0.160	0.046
WCB P-Value	0.103	0.095	0.545	0.371
Observations	61,071	56,828	60,935	52,452

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). In contrast to the main specification, we use ”books at home” instead of parental education as proxy for socio-economic status of the family. Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table A.7: Effect of Multigrade Class on Students' Outcomes - Teacher and Schools Characteristics as Controls

	Reading Test Score	German Grade	High-Track Recommendation	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.101***	-0.107***	-0.045	-0.017
	(0.019)	(0.029)	(0.039)	(0.012)
Female	0.162***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.339***	-0.383***	-0.141***	-0.011**
	(0.014)	(0.011)	(0.012)	(0.004)
Migration Background	-0.294***	-0.243***	-0.070***	0.017**
	(0.027)	(0.024)	(0.014)	(0.006)
Kindergarten Attendance (years)	0.085***	0.077***	0.037***	0.002
	(0.009)	(0.005)	(0.004)	(0.003)
Low SES	-0.393***	-0.380***	-0.197***	-0.028***
	(0.020)	(0.013)	(0.016)	(0.008)
Teacher is female	0.025	-0.010	-0.003	0.022**
	(0.026)	(0.023)	(0.012)	(0.010)
Age of teacher	0.002	-0.001*	0.000	-0.001*
	(0.002)	(0.001)	(0.001)	(0.001)
Experience of teacher	-0.001	-0.001	-0.001*	0.001**
	(0.002)	(0.001)	(0.001)	(0.001)
Teacher specialized	0.033*	0.041**	-0.003	0.016**
	(0.017)	(0.017)	(0.008)	(0.007)
Teacher works full-time	-0.019	-0.033**	-0.020**	0.006
	(0.013)	(0.015)	(0.009)	(0.008)
No. of students enrolled	0.000	-0.000	0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
School is a public School (Dummy)	-0.171***	-0.161***	-0.077***	-0.014
	(0.034)	(0.027)	(0.016)	(0.009)
Experience as headmaster	-0.001	-0.001	-0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.000)
Age of headmaster	0.001	0.001	0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.000)
Headmaster is male	-0.041**	-0.022	-0.020**	0.010
	(0.019)	(0.015)	(0.009)	(0.007)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.226	0.202	0.143	0.046
WCB P-Value	0.143	0.060	0.773	0.537
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). In contrast to the main specification, we add controls for teacher and school characteristics. Reading testscores are z-standardized. Grades in German are on a scale from 1 (insufficient) to 6 (very good). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.