DISCUSSION PAPER SERIES

IZA DP No. 5922

Classroom Grade Composition and Pupil Achievement

Edwin Leuven
Marte Rønning

August 2011

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Classroom Grade Composition and Pupil Achievement

Edwin Leuven
*University of Oslo, CEPR, CESifo,
Statistics Norway and IZA*

Marte Rønning
Statistics Norway (SSB)

Discussion Paper No. 5922
August 2011

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5922
August 2011

ABSTRACT

Classroom Grade Composition and Pupil Achievement[*]

This paper exploits discontinuous grade mixing rules in Norwegian junior high schools to estimate how classroom grade composition affects pupil achievement. Pupils in mixed grade classrooms are found to outperform pupils in single grade classrooms on high stake central exit tests and teacher set and graded tests. This effect is driven by pupils benefiting from sharing the classroom with more mature peers from higher grades. The presence of lower grade peers is detrimental for achievement. Pupils can therefore benefit from de-tracking by grade, but the effects depend crucially on how the classroom is balanced in terms of lower and higher grades. These results reconcile the contradictory findings in the literature.

JEL Classification: I2

Keywords: educational production, combination classes, class size, peer effects

Corresponding author:

Edwin Leuven
Department of Economics
University of Oslo
Postboks 1095 Blindern
0317 Oslo
Norway
E-mail: edwin.leuven@econ.uio.no

[*] We thank Adam Booij, Eric Bettinger, Julie Cullen, Pascaline Dupas, Tarjei Havnes, Magne Mogstad, Hessel Oosterbeek, Holger Sieg, David Sims and seminar participants for generous comments. A special thanks to Maria Fitzpatrick for providing us with descriptive statistics from SASS 2007. The usual disclaimer applies.

1 Introduction

What are the consequences of classroom grade composition for pupil achievement? Many children around the world find themselves in classrooms that group pupils from different ages and/or grades. These combination classes are not only common in many poor developing countries but are also often found in industrialized countries (Little, 2004).[1] In 2007, about 28 percent of schools in the United States report "using multi-age grouping to organize most classes or most pupils".[2] Similarly, in 2001 about 25 percent of primary school pupils were in mixed grade classrooms in Ontario (Fradette and Lataille-Démoré, 2003). The incidence of combination classes is also high in many European countries (Mulryan-Kyne, 2005). In France, for example, 37 percent of primary school pupils are in mixed grade classrooms.[3]

Although combination classes are sometimes advocated from an educational point of view, they typically arise because of economic constraints. When confronted with an increase or drop in enrollment, schools often group pupils from different grade levels to avoid an extra (costly) classroom. This explains why combination classes are also common in regular sized schools in cities, even though they are typically associated with small schools in rural and remote areas. Thirty-two percent of American public schools located in cities report using multi-age grouping, compared to 26 percent in rural areas.

There are several ways in which combination classes can affect pupil achievement. Classrooms constitute natural peer groups and grouping pupils from different grades in a single classroom changes the peer group relative to a single grade classroom. This may lead to direct negative or positive spillovers due to the presence of more or less able peers since a pupil's grade is positively correlated with her age and length of schooling, and therefore with cognitive development and achievement (f.e. Bedard and Dhuey 2006; Fredriksson and Öckert 2005; Leuven et al. 2010). In addition, peers from higher grades can serve as role models in terms of non-academic behavior, which can feed back to school achievement. Finally, classrooms' grade composition can also significantly affect teacher inputs and teaching methods.

There is surprisingly little solid causal evidence about the impact of combination classes on pupil achievement. Veenman (1995) surveyed 56 studies and concluded that pupils in mixed grade classrooms do typically no worse and sometimes better than

[1] Multi-grade and multi-age can correspond to different educational practices when age and grade do not coincide. In most industrialized countries there is a close correspondence between age and grade, in which case the distinction bears little practical meaning.

[2] Based on the NCES Schools and Staffing Survey (SASS), a large sample survey of America's elementary and secondary schools.

[3] Personal communication with Ministère d'Éducation Nationale.

pupils in classrooms that track pupils by grade. This conclusion was subsequently challenged by Mason and Burns (1997) who argued that existing studies failed to address sorting of both pupils and teachers into combination classes. This critique illustrates that any analysis of the effectiveness of combination classes needs to address the same identification problems as standard peer effects studies.

The lack of consensus about the effectiveness of combination classes reflects the difficulty of giving quantitative measure to peer effects highlighted by Manski (1993). To mitigate omitted variable bias most empirical peer-effects studies follow fixed-effect type approaches that rely on within school or grade variation in peer characteristics (f.e. Black et al. 2010; Hoxby 2000; Lavy et al. 2008; Ammermueller and Pischke 2009). This strategy is compromised if pupils are not randomly allocated to peers and teachers (as in Rothstein, 2010). Although an analysis at the grade rather than the classroom level may partially address this issue, it can also lead to bias because peer group characteristics are then subject to measurement error (Ammermueller and Pischke, 2009; Sojourner, 2008). A practical limitation of many fixed-effect type studies is that, by their nature, they often have little variation in peer group composition. An alternative approach is to rely on experiments which randomly allocate pupils to classes (Boozer and Cacciola, 2001; Duflo et al., 2008). Social experiments are however rare and have their own limitations (Heckman and Smith, 1995), and quasi-experiments are an interesting alternative (f.e. Angrist and Lang 2004).

Some recent studies have addressed the endogeneity of combination classes. Sims (2008) uses an instrumental variable approach and finds that a higher fraction of students in combination classes negatively affects performance for 2nd and 3rd graders. Thomas (2011) follows a fixed-effects and selection-on-observables approach to estimate the impact of combination classes on 1st-graders and finds positive effects. Although these papers do an arguably better job at correcting for selection bias than previous studies, their contradictory findings remain a puzzle. The role of class room grade composition also plays an important role in the discussion about the effectiveness of middle schools. Rockoff and Lockwood (2010) for example estimate negative effects of switching to middle school in New York.

This paper sets out to estimate how classroom grade composition affects pupil achievement, and presents a number of significant contributions to the literature. The first is related to identification and addressing confounding effects. To address endogeneity issues the analysis relies on institutional features in Norway that significantly change the grade composition of classrooms. Norwegian junior high schools are bound by national regulation that uses enrollment by grade level to determine classroom grade composition. These rules give rise to many discontinuities. Although sample size limitations prevent us from exploiting the local nature of this setup, the rules determine predicted grade mixing which we use as instruments for actual grade mixing. Previous papers have struggled to address potential negative spillovers of grade mixing to other single grade class rooms, and the confounding effects of class size. An important aspect of our study is that the first issue does not arise in our setup, and that we can control for and, more importantly, instrument for class size. The second contribution of this study is that we can separate the average effect of grade mixing into that of sharing the class room with lower grades vs. higher grades. We argue below that these results go a long way toward explaining the contradictory findings in the literature.

To briefly summarize our results, we find that a one year exposure to a classroom that combines two grade levels increases exam performance by about 9 percent of a standard deviation. Further analysis shows that this effect is driven by pupils benefiting from sharing the classroom with more mature peers from higher grades, whereas the presence of a lower grade is detrimental to achievement. By the time they matriculate from junior high school, most pupils in mixed grade classrooms in Norway have spend time with both higher and lower grades. The average effect is therefore the sum of these positive and negative effects. Since the positive effect of sharing the classroom with a higher grade is somewhat larger in size that the negative effect of sharing the classroom with a lower grade, the average effect is small and positive. This illustrates that, depending on the type of exposure, average effects of grade mixing can be negative, positive or close to zero.

In what follows we start by describing the institutional context and our data sources. After outlining our empirical approach in Section 4, we present our estimation results in Section 5 and discuss how classroom age composition affects pupil achievement on the short and longer term. Section 6 concludes.

2 Institutional settings and data

2.1 Institutions

Compulsory education in Norway consists of six years of primary school and three years of junior high school education. Schools at the primary and secondary level are essentially public – private schools amount for less than 3% of total enrollment – and there are no school fees. Schools are governed at the local school district level and have catchment areas, implying that parental school choice between schools for given residence is not allowed.[4]

Children start primary school the year they turn seven.[5] One defining feature of the Norwegian schooling system is that early/late starting and grade retention are extremely rare. In the current context this is important since we are interested in the effects of classroom age composition on school achievement. Grade retention is strongly related to maturity (e.g. Cahan and Cohen (1989)), and if schools practice grade retention then this would introduce an extra endogenous margin of classrooms ability composition. As shown in Bedard and Dhuey (2006) and Strøm (2004) however, there is no grade retention in Norway. As a consequence nearly everybody starts junior high school the year they turn fourteen.

Our analysis focuses on integrated schools that manage both a primary and junior high school level (i.e. offer education from grade 1 to 9). More than half of the schools in Norway are integrated, most of which are located outside the four major cities.[6] Since these schools are relatively small and remote, it is common practice to combine multiple grades in a single classroom. All junior high schools in Norway – including the integrated schools – follow the same national curriculum, and all junior high school teachers are required to have completed teacher college. This has the important advantage that none of our results will be driven by differences in teacher education or curriculum.

2.2 Data

We use administrative enrollment data (provided by Statistics Norway) on all pupils who graduated from junior high school the school years 2001/02 and 2002/03. For these two cohorts we have complete information on their further schooling career (until 2009). We merge this data set with the school database GSI ("Grunnskolens Informasjonssystem") which, in addition to information on actual grade mixing, also contains information on number of pupils and classes per grade at the start of the school year. Norwegian administrative registers also provide us with information on the pupils' birth date and gender, socioeconomic characteristics such as mother's and father's education; whether parents cohabit; and whether the pupil has a non-western migrant background.

As measures of pupil performance we use test-score data from both teacher

[4] In specific cases parents can apply for exemptions to this rule, but this is very uncommon.

[5] Of the pupils in our data about two percent did not start primary school they year they turned 7, but one year earlier or later. School entry was lowered to age six as of 1997 when Norway increased compulsory schooling to 10 years. The official school starting age for the cohorts in our data was seven, and they had nine years of compulsory education.

[6] From the largest to the smallest these are: Oslo, Bergen, Trondheim and Stavanger. The last one having about 110,000 inhabitants at the time of our data.

set and graded tests in the final year, and centralized exit exams (from Statistics Norway). At the end of the final year in junior high, all pupils in Norway are required to take an exit exam. Although the curriculum includes many subjects, a written exit exam is only undertaken in one of three subjects: mathematics, Norwegian and English. The exams are centrally assigned and it is not known in advance what the exam topic will be, and are therefore beyond the control of schools, teachers and pupils. In the analysis we pool these three subjects and standardize them with zero mean and standard deviation one. The teacher tests as well as the exam scores are used to construct pupils' junior high school exit test scores which are important for secondary school choice.

The correlation between the teacher score and the exam score is 0.78. Although both the exam and teacher tests are supposed to measure learning of the same content (the junior high school curriculum), there are some differences that can affect their comparability. The exit exams are identical across schools and externally graded, which means that there are no comparability issues across schools. The teacher grades in these subjects on the other hand are based on tests set by students' teachers. It is therefore less clear to what extent these can be compared across schools. One advantage of the teacher tests scores is that they are based on multiple evaluations, and are therefore probably less noisy measures of achievement than the exam scores which are based on a single test. One important caveat regarding comparability arises if teachers engage in relative grading. This will not only make the teacher test scores less comparable, but can also be a source of bias if relative grading is affected by classrooms' grade composition. We will discuss this in more detail in the context of our results below.

Grade mixing mostly occurs outside the major cities in integrated schools. We therefore restrict our population of interest to these integrated schools outside the four largest cities. Since classroom information is recorded at the grade level and pupils are not necessarily randomly allocated to classrooms within a grade, we further restrict our sample to schools that have one 7th grade class room when pupils start junior high school. We drop 9 schools with missing information on predicted class size and schools where information on grade mixing is lacking, and 90 pupils with missing information on the exam score are also dropped.

Our analysis data set of small schools consists of 9,647 pupils and 388 schools. This amounts to about 10 percent of the pupil population and 1 out of 3 schools in Norway. In total 173 schools, about 1 out 6 of all junior high schools, combine grades in at least one school year. Figure A1 in the Appendix shows the location of the municipalities that have junior high schools combining grades, as well as the comparison group of municipalities with small schools that do not combine grades. The population of schools that we study not only represents an important fraction of the overall school population in Norway, but also provides good regional coverage.

Table 1 reports descriptive statistics for the pupils in small schools, and compares it to the total population of junior high school pupils. Relative age – which equals 0 for the youngest pupil (born December 31st) and 1 for the relatively oldest one (born January 1st) – is on average 0.5. This implies that pupils in their final year of junior high school in Norway are on average 16.5 years old. Differences with respect to individual and parental characteristics are mostly small: Compared to the whole population, parents of pupils in small schools are somewhat less educated, the mother and father are also slightly more often cohabiting.

By construction larger differences are observed regarding the schools pupils are enrolled in. First, schools are on average 3.5 times larger in the whole population compared to the integrated schools outside the major cities that offer both primary and junior high school education. Class size in these schools is also smaller, and teacher hours per pupil, a common related measure for resource use, is larger. The table reports averages over pupils' time in junior high school.

When comparing the schools that mix grades to the reference population of small schools we observe some differences with respect to parental background, but these tend to be small and we cannot reject the null hypothesis that there are no difference (p=0.318). Again, and – as we will show below – by virtue of the institutional rules, the mixing schools are smaller with smaller classes.

3 Maximum class size rules

Junior high schools in Norway were subject to maximum class size rules (e.g. Angrist and Lavy, 1999). What makes these rules unique is that they interact in a systematic fashion with classrooms' grade composition. Section 8.3 of the Norwegian Education Act (Opplæringsloven) stated the following:

1. A class in junior high school cannot have more than
   - 18 pupils when there are three cohorts in the class
   - 24 pupils when there are two cohorts in the class
   - 30 pupils when there is one cohort in the class
2. When there are multiple cohorts in a class, they need to be adjacent if possible

Table 1. Descriptive statistics

	All schools		Small schools		Mixing schools	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Pupil characteristics						
Relative Age	0.51	(0.28)	0.51	(0.28)	0.51	(0.28)
Girl	0.49	(0.50)	0.48	(0.50)	0.47	(0.50)
Parental characteristics						
Mother's education						
Junior High school or less (≤10)	0.11	(0.31)	0.11	(0.32)	0.13	(0.33)
High schools (11-13)	0.56	(0.50)	0.64	(0.48)	0.63	(0.48)
College (14+)	0.30	(0.46)	0.23	(0.42)	0.22	(0.41)
Father's education						
Junior High school or less (≤10)	0.12	(0.32)	0.15	(0.35)	0.18	(0.39)
High schools (11-13)	0.54	(0.50)	0.62	(0.49)	0.60	(0.49)
College (14+)	0.28	(0.45)	0.19	(0.39)	0.17	(0.38)
Cohabiting	0.67	(0.47)	0.71	(0.45)	0.71	(0.45)
N observations	98,090		9,636		2,130	
School characteristics						
Integrated school	0.53	(0.50)	1		1	
School size	152.4	(119.3)	43.0	(23.0)	22.7	(8.2)
Class size	21.0	(5.9)	16.6	(5.3)	13.8	(2.9)
Teacher Hours per pupil	98.2	(38.6)	122.0	(33.2)	144.1	(31.1)
N schools	1,040		388		170	

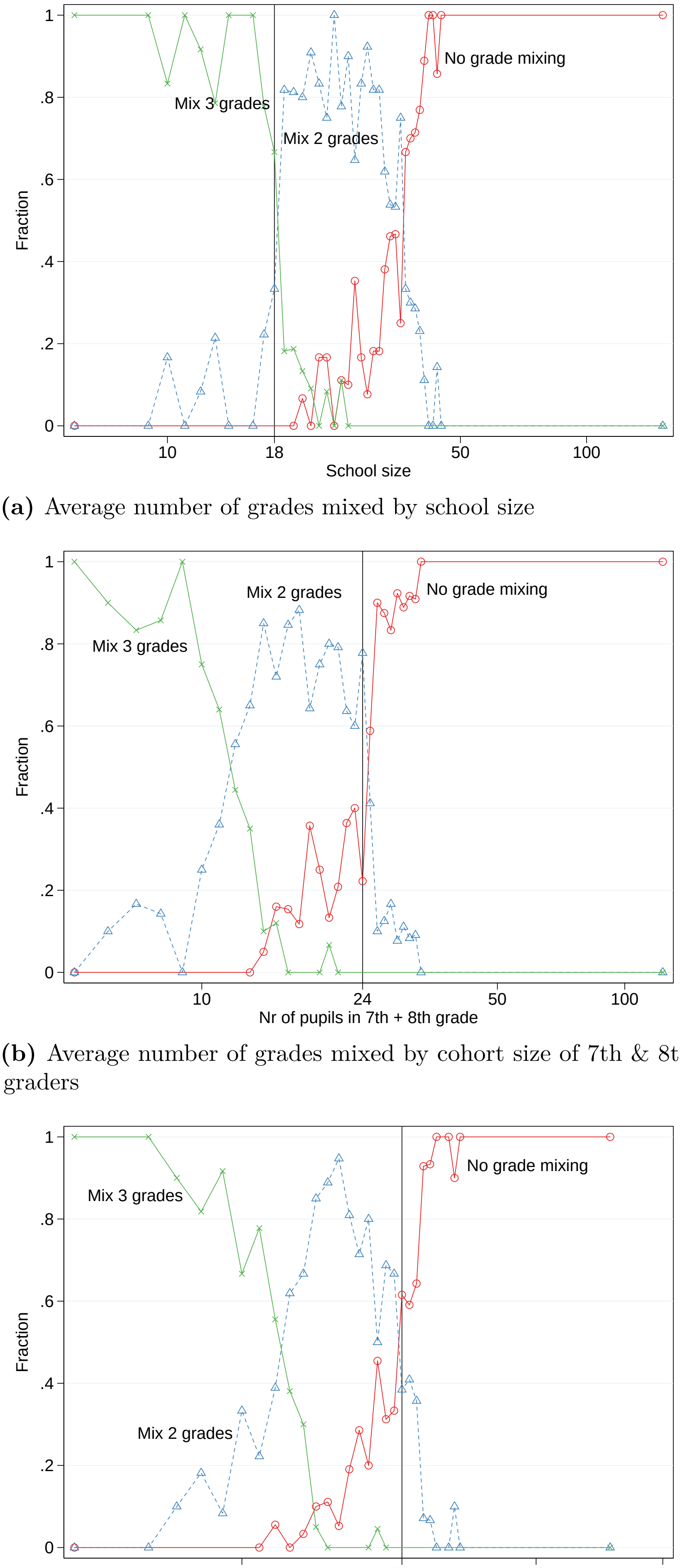
3. The school cannot simultaneously have mixed age and age-homogeneous classes within the same grade level, or parallel mixed age classes

School are funded based on the number of classrooms they are supposed to operate according to these rules. Consequently there is little scope to permanently deviate from these rules since schools cannot levy fees or seek additional funding to avoid combining grades. Since the rules affect both classroom grade composition and class size, our empirical analysis will take into account potential endogeneity on both margins. We explain our identification strategy in detail in the next section, after showing how these rules were applied in practice.

Figure 1a illustrates the contemporaneous relationship between school size and multiple grade classrooms in 9th grade generated by the institutional setup. The x-axis in Figure 1 is on a logarithmic scale to improve the readability of the graph. The vertical line at 18 pupils marks the threshold above which schools are not supposed to combine all three grades. There is a close relation between observed grade mixing and the grade mixing rule. We observe a sharp drop in the propensity to combine these grades of 0.5. Where to the left of the first threshold schools essentially mixes all three grades, for schools larger than 18 pupils the picture is somewhat more complicated and schools tend to mix two adjacent grades. At first schools are bound by rules regarding the combination of two adjacent grades, and for schools larger than 50 pupils there is no longer any grade mixing taking place.

Schools are supposed to combine two grades when either the combined enrollment of 7th and 8th grade, or the combined enrollment of 8th and 9th grade drops below 24. These rules are illustrated in Figures 1b and 1c. We again see sharp drops in the incidence, this time of double grade classrooms. Note however, that these two discontinuities interact. To better understand how schools go from a single to a double grade classroom Figure 2 shows actual grade mixing as a function of the relevant cohort sizes. The four quadrants correspond to different enrollment regimes distinguished by the institutional rules. In the top-left quadrant the combined enrollment of 8th and 9th grade exceeds 24, and there grades should therefore not be combined. Enrollment of 7th + 8th grade on the other hand is 24 or less, and the rules therefore say that schools should combine these two grades. This is what we observe, although some schools deviate from the rules. In the bottom-right quadrant we observe a similar pattern, but then with respect to combining 8th and 9th grade.

Most schools find themselves in the top-right bottom-left quadrant. The top-right quadrant corresponds to the situation where the combined enrollment of both 7th + 8th and 8th + 9th grade exceeds 24. In this case the rules stipulate that grades are not to be combined which is indeed what we observe in the data, and these are



(a) Average number of grades mixed by school size

(b) Average number of grades mixed by cohort size of 7th & 8th graders

(c) Average number of grades mixed by by cohort size of 8th & 9th graders

Figure 1. Grade mixing discontinuities in 9th grade

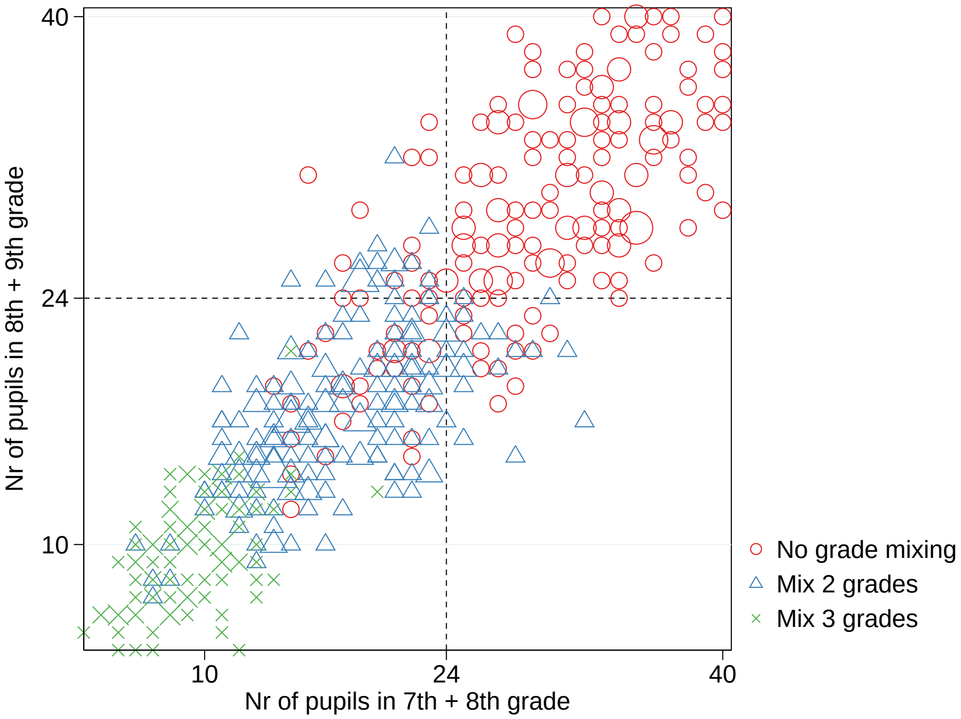


Figure 2. The interaction of grade mixing rules – 9th grade enrollment

regular single grade classrooms. In the bottom-left quadrant schools should combine two grades, unless the combined enrollment of 7th to 9th grade drop below 18 in which case schools are supposed to combine these three grades in a single classroom.

4 Empirical strategy

Pupils in classes with more than one grade level are exposed to more heterogeneous classrooms than those in single grade classes. The first question we set out to investigate in this paper is whether it is more beneficial to be in combination classes than in single grade classrooms. We do so by estimating the achievement effect of the number of different grades in the classroom using the following equation:

$$y_i = \alpha \cdot g_i + \gamma \cdot ssize_i + x_i'\beta + \varepsilon_i \qquad (1)$$

where y_i is pupil i's achievement at the end of junior high. Our main variable of interest, g_i, is the average number of grade levels in the classroom that a pupil was exposed to during junior high school. So for pupils who have never been in mixed grade classrooms $g = 1$. If they were mixed in 7th grade and not mixed in grades 8 and 9, then $g = (2 + 1 + 1)/3 = 4/3$, etc. We also add school and family control variables in x_i, which include parental education, whether parents are living together,

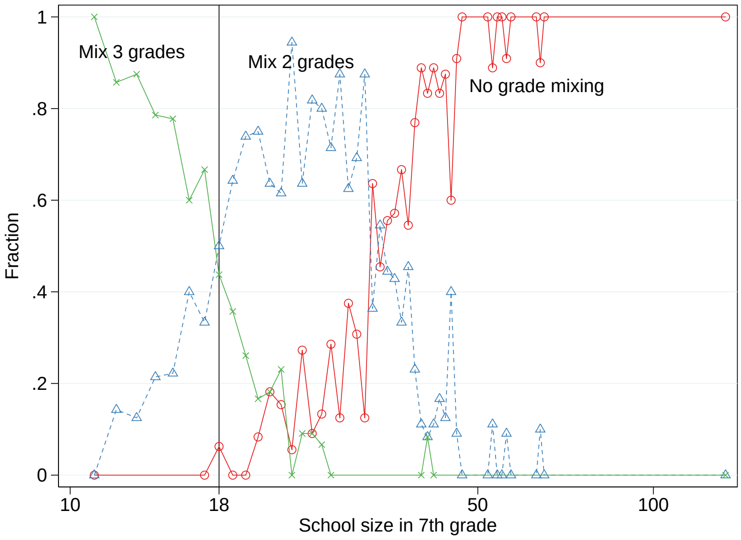


Figure 3. Grade mixing discontinuities in 9th grade – 7th grade enrollment

pupils gender and relative age.[7]

As documented above, grade mixing is governed by the rules set by the Ministry of Education. We use predicted grade mixing to construct instruments for actual grade mixing to take any remaining endogeneity into account. Endogeneity is potentially an issue, especially close to the thresholds where schools more often deviate from the rules. One example of endogenous grade mixing arises when school's grade mixing in year t depends on the (perceived) success of grade mixing in year $t-1$, rather than the rule.

Although one can think of the institutions as generating a regression discontinuity (RD) design, sample size and support limitations prevent us from performing a standard RD analysis. Note that the discontinuities are defined defined at the grade level. This is illustrated by Figure 3 which graphs grade mixing when the pupils are in 9th grade as a function of enrollment two years earlier (when the pupils were in 7th grade). Since the grade mixing rules rely on contemporaneous enrollment, using prior enrollment masks the discontinuity because year to year enrollment changes smooth out the discontinuities. This means that a proper RD analysis would need to consider all the permutations of the discontinuities in Figure 1 across the three grades in junior high school.

Since we have insufficient data for a standard RD analysis, we follow an instrumental variable approach in the spirit of Angrist and Lavy (1999) where we instrument observed grade mixing with predicted grade mixing. Using the enrollment

[7] We also estimated specifications where we instrument actual age using relative age as in Bedard and Dhuey 2006; Black et al. 2010. This does not affect our results. We report estimation results from reduced form models with respect to age for simplicity.

of 7th, 8th and 9th graders in a given school year we can determine the predicted grade mixing according to the rules. For each pupil we calculate the predicted grade mixing separately for each grade level when she was in junior high school. In a slight abuse of notation define predicted grade mixing for student i in year t, $E[g_{it}]$, as follows

$$E[g_{it}] = \begin{cases} 1 & \text{if } n_{it}^7 + n_{it}^8,\ n_{it}^8 + n_{it}^9 > 24 \\ 3 & \text{if } n_{it}^7 + n_{it}^8 + n_{it}^9 < 19 \\ 2 & \text{otherwise} \end{cases}$$

where n_{it}^j is the number of j-th graders in student i's school in year t. These conditions correspond to the official rules, also illustrated by Figures 1 and 2 above.

We construct predicted grade mixing for when a student was in 7th grade (year t), 8th grade ($t+1$) and 9th grade ($t+2$). In our 2SLS estimation we use six predicted grade mixing dummies, one for each grade and value of $E[g_{it}]$, leaving out the reference group of no grade mixing. The first stage thus becomes

$$g_i = \sum_{j=7}^{9} \sum_{n=2}^{3} \delta_{jn} \mathbb{1}_{\left[E[g_{i,\,t(i,\,j)}]=n\right]} + \delta_s \cdot ssize_{i,\,t(i,\,7)} + x_i' \delta_x + u_i \tag{2}$$

where $t(i,\, j)$ is the year pupil i was in j-th grade. We control throughout for school size – the combined enrollment of 7th, 8th and 9th grade when the pupil started junior high school – in the analysis. School size can be thought of as a running variable and potential confounder.[8]

To further investigate whether it matters to be mixed with higher or lower grade pupils, we also decompose the number of grades in a classroom into number of higher and lower grades as follows

$$g_i = 1 + g_i^+ + g_i^-$$

where g_i^+ is the average number of higher grade levels in the pupil's classroom while she was in junior high school. For example when mixed with 8th and 9th when in 7th grade, and not mixed afterward then $g_i^+ = (2+0+0)/3 = 2/3$, when mixed with 8th graders in 7th grade, 9th graders in 8th grade and not mixed in the final grade then $g_i^+ = (1+1+0)/3 = 2/3$, etc. Similarly, g_i^- is the average number of lower grade levels a pupil shared the classroom with. This leads to the following equation

$$y_i = \alpha_+ g_i^+ + \alpha_- g_i^- + \lambda \cdot ssize_i + x_{ij}' \beta + \varepsilon_1 \tag{3}$$

where we instrument both g_i^+ and g_i^- with the same set of instruments as in equation

[8] The relationship between school size and test scores in the total population is linear with a slope coefficient close to zero.

Table 2. Classroom count of observed grade mixing sequences

Sequence	N	Sequence	N	Sequence	N
111	408	221	69	311	1
112	4	222	23	321	5
121	47	223	8	322	5
122	38	231	1	323	6
123	9	232	3	331	2
132	2	233	10	332	3
133	3			333	63
				Total	710

Note: The 1st/2nd/3rd number in the shown sequences denotes mixing in 7th/8th/9th grade, where 1 = single grade classroom (no grade mixing), 2 = two grade classroom, 3 = three grade classroom.

(2).[9]

The variation that allows us to separately estimate α_+ and α_- is illustrated in Table 2, where we see that there are many different observed grade mixing sequences in our sample. Whether pupils were in a mixed grade classroom at one point during junior high school can therefore correspond to very different peer groups. Some pupils might have been mixed with lower grade peers, whereas others might be mixed with pupils from higher grades. Many sequences also differ with respect to the timing of grade mixing. Some grade mixing sequences are very common, such as being in a classroom that mixes all three grades ('333') throughout junior high school, being with 8th graders in 7th grade and with 7th graders in 8th grade ('221') or with 9th graders in 8th grade and with 8th graders in 9th grade ('122').

In addition to affecting the grade level composition of the class room, grade mixing also influences class size. Using the same data sources as this paper but excluding the integrated schools, Leuven et al. (2008) find that class size has no effect on pupil achievement in Norwegian junior high schools. This suggests that we do not need to control for class size. The variation in class size is however at smaller class size levels (average class size in schools that combine grades is 14), and class size can also affect achievement differently heterogeneous grade classrooms. We therefore take class size into account, and use predicted class size on junior high school start in 7th grade as an instrument for actual class size (average class size when in junior high school - the same class size measure as in Leuven et al., 2008).

[9] We also estimated specifications based on binary variables for being mixed, and being mixed with pupils from lower cq. higher grades. The results from these estimations – which represent average effects in our sample – are qualitatively similar to the ones we report. The estimates reported in in the text are scaled in terms of number of grades and are therefore more straightforward to interpret quantitatively.

Predicted class size is defined as follows

$$E[csize_i] = n^7_{it} + (n^8_{it} + n^9_{it}) \cdot \mathbb{1}_{[E[g_{it}]=3]} + 0.5n^8_{it} \cdot \mathbb{1}_{[E[g_{it}]=2]} \tag{4}$$

where $t = t(i, 7)$. Equation (4) implies that the expected class size on junior high school start is n^7_{it} in a single grade class, and $n^7_{it} + n^8_{it} + n^9_{it}$ in a three grade class. When two grades are predicted to be combined this can either be 7th and 8th grade or 8th and 9th grade. In the first case the expected class size is $n^7_{it} + n^8_{it}$, and in the second case n^7_{it}. We assume that these events have equal probability (0.5) which gives the expected class size in (4). The class size effect is therefore identified through an interaction between the predicted grade mixing rules and adjacent cohort sizes.

Since we are instrumenting class size we will estimate an additional first-stage for class size and augment the first stage (2), and the first-stages for g^+_i and g^-_i with (4). Our results below confirm our earlier findings for larger schools in Leuven et al. (2008), namely that there is no evidence of significant class size effects in Norwegian lower secondary schools. Our effect estimates of grade composition therefore do not change when we do not control for class size.

Finally, because we exploit the rules documented above as instrumental variables, we investigate their validity in two ways. First we check whether parents and/or schools position themselves in non-random ways around the points where schools are supposed to change the classroom grade composition. A second concern are alternative confounding changes of related school inputs. We discuss each in turn.

4.1 Sorting

We can distinguish between two main sources of sorting. The first is supply side sorting which arises when school or local education authorities manipulate enrollment relative to the discontinuities. The main reason for doing so is typically related to funding. In some countries, for example in Sweden, local education authorities are known to sometimes redraw school catchment areas in such a way as to avoid opening a new classroom when maximum class-size rules would dictate this. This is however not an issue here since catchment areas are fixed in Norway.

The second potential source of sorting is the demand side. When parents prefer mixing or non-mixing classrooms they might decide to enroll their children in a different school. If for example more advantaged families sort in different ways than disadvantaged families, the underlying pupil population at both sides of the discontinuities are no longer comparable. The implicit exclusion restriction in the IV design then breaks down and we would no longer recover reliable estimates.

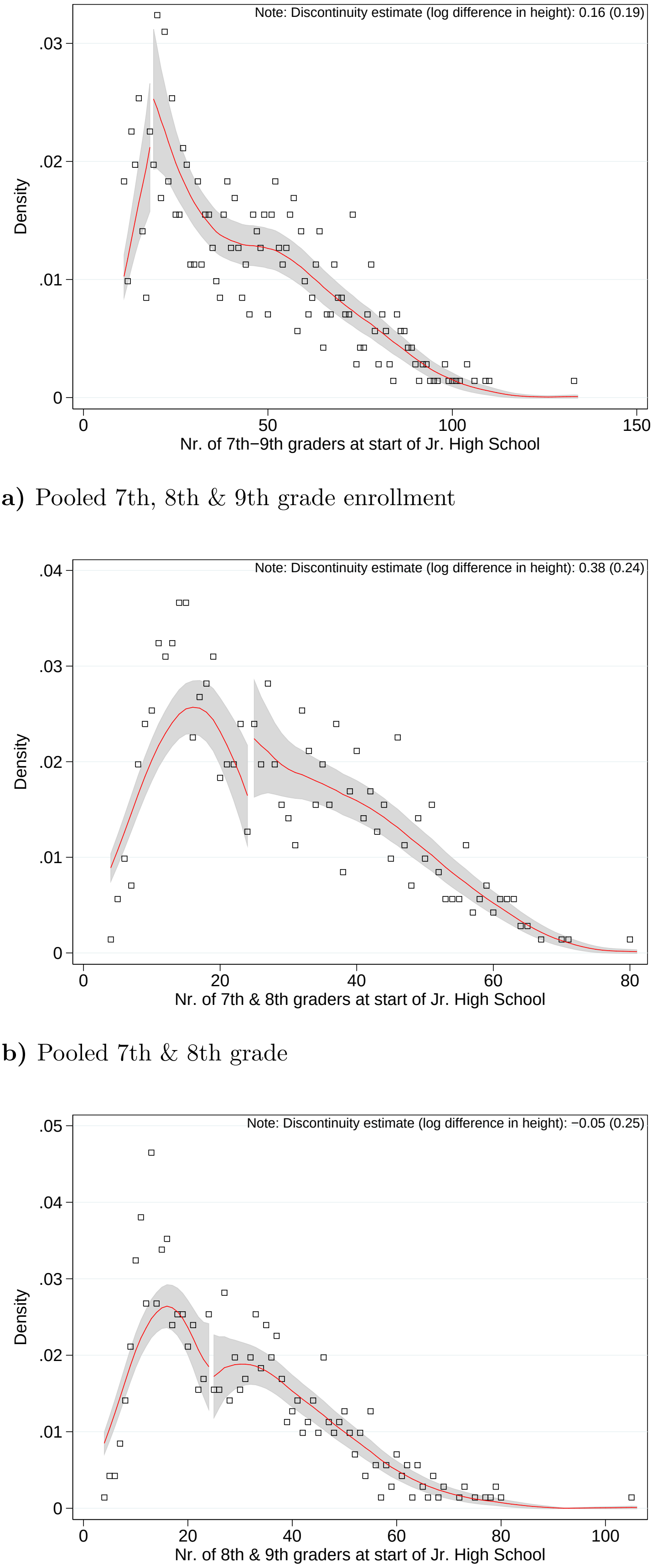
A striking example of sorting was reported in Urquiola and Verhoogen (2009) for Chile. In an earlier class-size study (Leuven et al., 2008) we did not find any similar evidence for Norway. When it comes to institutional sorting this is as expected since catchment areas are fixed.

As mentioned above, there is essentially no grade repeating in Norway. A potentially bigger concern for the endogeneity of classroom composition is the possibility of families moving to different school catchment areas in reaction to or anticipating classroom grade composition. Hægeland et al. (2008), who use the same pupil data as we do, report that 95.3 percent of the pupils lived in their graduation municipality throughout all three years in junior high schools. Since our estimation sample consists of non-urban schools, we expect mobility to be considerably lower. We can implement a check by comparing the administrative headcounts for 7th and 8th grade with the 9th grade headcounts when these 7th and 8th graders are supposed to be in 9th grade (unless they repeat a grade or move to another school). The correlation between these two measures is very high, namely 0.995 for 8th grade and 0.990 for 7th grade. We take this as evidence confirming that endogenous grade repetition and pupil mobility are not a concern in our data.

Although within-school sorting is not a concern, we have not ruled out between-school sorting. This would arise when parents sort prior to the start of junior high school or when school districts try to manipulate the discontinuities. For this reason we check whether we can detect discontinuities in the enrollment densities. We follow McCrary (2008) and calculate these discontinuities using local linear regression techniques.

Figure 4 pools the different years in our data, and shows density plots for the three discontinuities that we exploit in the analysis. The top figure shows total junior high school enrollment where the discontinuity is at 18. As can be seen from the graph, the density peaks around enrollment of 19, but we cannot reject that there is no discontinuous jump at 18. The estimated log difference in the height of the density is 0.27, but not statistically significant. The middle figure shows a similar graph for combined enrollment of 7th and 8th graders where the discontinuity lies at 24. Here the estimated density is also higher to the right of the discontinuity, but again not statistically significant. Finally the lower figure shows the estimated discontinuity for the combined enrollment of 8th and 9th graders for the pooled years. Now the estimated density is somewhat lower at the right side of the kink and also not significant.

These results are probably not too surprising since school districts in our current application are rural, have typically one school, with the next school often a few



(a) Pooled 7th, 8th & 9th grade enrollment

(b) Pooled 7th & 8th grade

(c) Pooled 8th & 9th grade enrollment

Figure 4. Density checks

hours away by car. Since Norway has catchment areas, parents would often need to move to another municipality in order to enroll their child in another school. Social cost of sorting is therefore high, and parents would also typically need to find new employment.

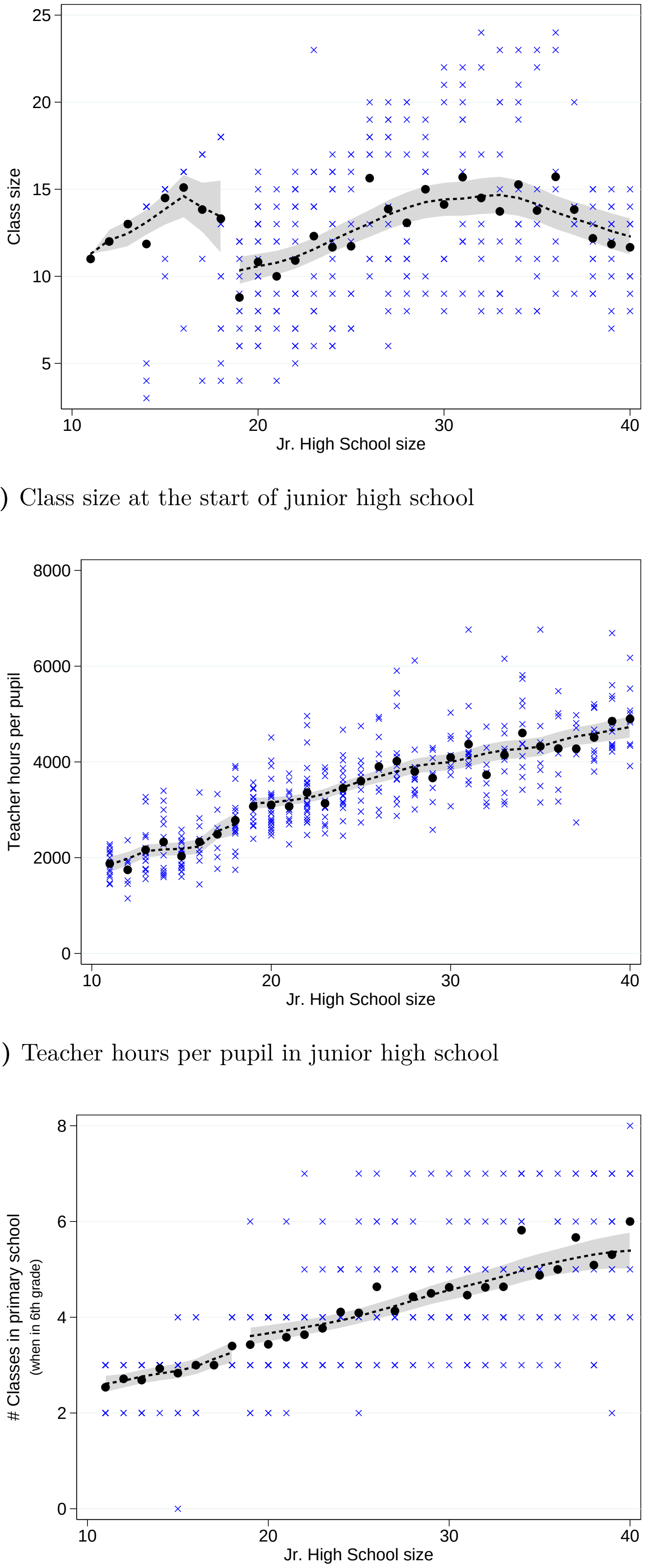
4.2 Confounding discontinuities

Although we do not find any evidence of sorting, we know that class size discontinuously changes when combining grades. The reason is of course that, keeping enrollment fixed, combining grades involves less classrooms and therefore mechanically larger classes. This is illustrated in Figure 5a which plots the data points corresponding to the schools in our sample and a smoothed regression line and bootstrapped confidence interval at both sides of the discontinuity.

We see that combining grades not only changes the classroom's grade composition, but also class-size. Although Leuven et al. (2008) did not find evidence that class size affects achievement in Norwegian junior high schools and rule out small effects, the population of schools in the current paper is different and also the variation in class-size is at smaller class size levels than commonly estimated. We therefore control for class-size when estimating how grade mixing affects achievement. As discussed above, in our estimations class-size is instrumented with predicted class-size at the start of junior high school as in Angrist and Lavy (1999).

Even though class size discontinuously drops, for the interpretation of our results it is important to understand the input substitution around the kinks. In the classic IV approach that exploits maximum class size rules (Angrist and Lavy, 1999) teacher input (at least in terms of time) is kept constant. From our administrative data we know the ratio of teacher hours per pupil at the junior high school level. Figure 5b shows that the drop in class-size does not seem to be accompanied by a drop in teacher hours per pupil. This suggests that when schools combine grades and have larger classes, input in terms of teacher time remains constant if these mixed grade classrooms are taught by multiple teachers. Unfortunately we cannot verify this because our data does not allow us to link teachers to classrooms.

In primary school, pupils from different grades can also be combined in a single classroom. These rules are however different from those in junior high school (both in terms of thresholds, but also in that they rely on more cohorts simultaneously). One might nevertheless be concerned that grade mixing in junior high school correlates with grade mixing in primary school. Since combining grades changes the number of classes we verify whether we observe a discontinuous change in the number of classes when the pupil was in 6th grade (the final grade of primary school). Figure 5c shows



(a) Class size at the start of junior high school

(b) Teacher hours per pupil in junior high school

(c) Number of classes at the end of primary School

Figure 5. Confounding discontinuities

that there is no evidence of such a confounding discontinuity.

5 The effect of class room grade composition on achievement

This section presents the outcomes of our analysis. We start out by considering average effects of classroom grade composition on exam scores at the end of junior high school. After these overall results we present separate effect estimates for boys and girls.

5.1 Exam scores

The results from estimating equation (1) by OLS are shown in the first two columns of Table 4. The first column is a simple regression of standardized exam scores on average classroom grade composition during junior high school. This shows that pupils who have been in classes with one more grade level in their class during junior high school perform approximately 6 percent of a standard deviation better on the exam. The second column adds class size, school size, and our family background characteristics. The effect of number of grades increases somewhat to an effect size of 7 percent and remains significant at the 10 percent level.

The second and third column in Table 4 present the estimates after instrumenting number of grade levels in the classroom using 2SLS. Table 3 reports the first-stage results. When we test the joint significance of our instruments, the predicted grade level dummies, we obtain an F-statistic equal to 232. The first set of 2SLS estimates assumes that class size is exogenous. The point estimate increases somewhat to 0.095 to the OLS estimates, and is unaffected when we also instrument class size in the final column. The estimated effect of class size is small and positive, 0.002, insignificant yet precisely estimated. Not only are there no confounding effects of class size on the number of grade levels in class, but this also confirms the earlier finding of Leuven et al. (2008) that class size effects in Norwegian junior high schools are negligible.

Turning to the control variables we see that the oldest pupils in the cohort, born in January, score about 16 percent of a standard deviation higher than the youngest in the cohort born in December. Girls also score significantly higher than boys, and exam scores are also better for children of higher educated and cohabiting parents. Finally, we see that there is no statistically significant relation between the running variable, school size, and exam scores (dropping school size from our regressions does not affect the results).

Table 3. First stage regressions

	# Grades		# Lower grades		# Higher grades		Class size	
Predicted # grades:								
- 3 in 7th grade	0.582	$(0.087)^{***}$	0.071	(0.057)	0.510	$(0.057)^{***}$	-5.362	$(0.760)^{***}$
- 2 in 7th grade	0.068	(0.043)	-0.018	(0.029)	0.086	$(0.028)^{***}$	0.309	(0.381)
- 3 in 8th grade	0.710	$(0.078)^{***}$	0.386	$(0.074)^{***}$	0.324	$(0.054)^{***}$	0.724	(0.575)
- 2 in 8th grade	0.282	$(0.038)^{***}$	0.147	$(0.026)^{***}$	0.135	$(0.022)^{***}$	0.348	(0.418)
- 3 in 9th grade	0.493	$(0.082)^{***}$	0.462	$(0.073)^{***}$	0.031	(0.042)	1.420	$(0.523)^{***}$
- 2 in 9th grade	0.099	$(0.046)^{**}$	0.106	$(0.031)^{***}$	-0.007	(0.028)	0.241	(0.434)
Predicted class size	-0.003	$(0.001)^{**}$	-0.002	$(0.001)^{**}$	-0.001	(0.001)	0.738	$(0.058)^{***}$
Relative Age	0.000	(0.006)	0.004	(0.004)	-0.004	(0.004)	0.012	(0.094)
Girl	0.001	(0.004)	-0.002	(0.003)	0.003	(0.002)	-0.134	$(0.065)^{**}$
M - High school	-0.009	(0.008)	-0.005	(0.005)	-0.003	(0.004)	-0.049	(0.112)
M - College	-0.009	(0.009)	-0.005	(0.006)	-0.005	(0.005)	-0.034	(0.142)
F - High school	-0.008	(0.006)	-0.008	$(0.004)^{**}$	0.000	(0.004)	-0.179	$(0.086)^{**}$
F - College	-0.005	(0.007)	-0.008	$(0.005)^{*}$	0.003	(0.005)	-0.137	(0.102)
Parents cohabit	0.008	(0.006)	0.002	(0.004)	0.007	$(0.003)^{*}$	0.233	$(0.082)^{***}$
School size / 100	-0.135	$(0.051)^{***}$	-0.038	(0.031)	-0.097	$(0.030)^{***}$	-1.517	(1.742)
	1.168	$(0.033)^{***}$	0.087	$(0.020)^{***}$	0.081	$(0.019)^{***}$	6.188	$(0.799)^{***}$
First stage F-statistics:								
- all instruments	200.0		144.1		145.2		29.5	
- predicted grade mixing dummies	231.9		168.0		167.9		9.5	
- predicted class size	4.0		5.8		0.6		162.0	
Joint F test ind. char. (p-value)	0.516		0.306		0.296		0.039	

Note: Standard errors are heteroscedasticity robust and corrected for school-level clustering. */**/*** statistically significant at the 10/5/1 percent level. The College and High school dummies refer to (M)other' and (F)ather' education. All regressions include a constant term.

Table 4. The relation between grade mixing and pupil performance, dependent variable is the exam scores - OLS

	OLS	OLS	2SLS	2SLS
# Grades	0.059	0.068	0.095	0.091
	(0.031)*	(0.038)*	(0.044)**	(0.043)**
Class size		0.002	0.002	0.001
		(0.004)	(0.004)	(0.006)
School size / 100		-0.079	-0.038	-0.022
		(0.130)	(0.137)	(0.134)
Relative Age		0.158	0.159	0.158
		(0.033)***	(0.033)***	(0.033)***
Girl		0.358	0.358	0.358
		(0.021)***	(0.021)***	(0.021)***
M - High school		0.216	0.217	0.217
		(0.030)***	(0.030)***	(0.030)***
M - College		0.636	0.637	0.637
		(0.037)***	(0.037)***	(0.037)***
F - High school		0.150	0.151	0.151
		(0.024)***	(0.024)***	(0.024)***
F - College		0.439	0.440	0.439
		(0.033)***	(0.033)***	(0.033)***
Parents cohabit		0.224	0.224	0.224
		(0.022)***	(0.022)***	(0.022)***
Instrument class size				✓
R-squared	0.001	0.206		

Note: Standard errors are heteroscedasticity robust and corrected for school-level clustering. */**/*** statistically significant at the 10/5/1 percent level. The College and High school dummies refer to (M)other' and (F)ather' education. All regressions include a constant term. Estimation sample contains 388 schools and 9,636 pupils.

These results might be surprising, in the sense that the heterogeneity of the classroom increases when combining grades. The results of Duflo et al. (2008) for Kenya for example suggest that this should have deteriorated pupils' achievement. To gain more insight into what is driving this result, Table 5 reports estimation results using equation (3). The top panel of the table present estimates for exam scores and the second panel presents the results for the teacher set and graded tests. For both outcomes we present OLS and 2SLS estimates of the effects of g^- and g^+ , and also the effect of class size. To take away any remaining concerns about omitted variables, such as endogenous sorting to schools we also report estimation results from specifications that include school fixed effects.

In the first OLS specification the point estimate of the effect of exposure to the number of lower grades (g^- in equation 3) on exam scores is -0.11. This suggests that sharing the classroom with a lower grade is detrimental for the exam scores, the point estimate however lacks statistical significance at conventional levels. Pupils in classes where a higher grade level is added score significantly higher on the exam. Adding school fixed effects to the equation does not significantly change the estimates but comes at the cost of a substantial loss in the precision of the estimates.

When we instrument both grade composition variables the point estimates increase. For the number of lower grades we now obtain a point estimate of about -0.22 which is close to being significant at the ten percent level. The point estimate for the number of higher grades in the class room is 0.42 and significant at the 1 percent level. Recall from Table 2 that if pupils are mixed, then they typically spend time with both lower and higher grades. This explains the effects in Table 4: grade mixing is on average beneficial because pupils benefit more from being with higher grades than they loose from being with lower ones. The final column reports the 2SLS estimates from the specification with school fixed effects. The effect for the number of lower grades drops but remains negative even though it is no longer statistically significant. The effect for the number of higher grader increases. We cannot reject equality of the 2SLS estimates with and without fixed effects: when we bootstrap these estimates to perform a Wald test we obtain a test statistic of 0.69 with a p-value of 0.708.

The second panel of Table 5 adds estimates for the teacher set and graded test scores. These results confirm the conclusion based on the exam scores, namely that students benefit from sharing the classroom with higher grades, and are harmed if the other grade level in the classroom is lower. Note that we have more precision on the teacher scores than on the exam scores. This is what we expected because the teacher scores are based on multiple evaluations and therefore probably less noisy

Table 5. The effect on pupil achievement of being mixed with higher/lower grades

	OLS		2SLS	
A. Exam Score				
# Lower grades	-0.108	-0.023	-0.224	-0.140
	(0.074)	(0.174)	(0.134)*	(0.234)
# Higher grades	0.261	0.271	0.425	0.714
	(0.076)***	(0.150)*	(0.142)***	(0.261)***
Class size	0.002	-0.007	-0.000	-0.005
	(0.004)	(0.007)	(0.006)	(0.006)
B. Teacher Score				
# Lower grades	-0.085	-0.143	-0.262	-0.470
	(0.052)	(0.093)	(0.099)***	(0.188)**
# Higher grades	0.164	0.297	0.384	0.631
	(0.058)***	(0.091)***	(0.112)***	(0.216)***
Class size	0.002	-0.005	0.001	-0.009
	(0.003)	(0.004)	(0.004)	(0.005)*
School FE's		✓		✓

Note: All regressions include a constant term and the full set of controls in Table 4. Standard errors are heteroscedasticity robust and corrected for school-level clustering. */**/*** statistically significant at the 10/5/1 percent level. Estimation sample contains 388 schools and 9,636 pupils.

than the exam scores. Contrary to the exam scores, which are externally set and graded, teacher grades may however have a relative component. If teachers grade on a reference curve that depends on classroom composition then the presence of higher grades would lower relative scores, and the presence of lower grades would increase relative scores. Relative grading will thus cause a bias towards zero. The effects on teacher grades are however of the same order of magnitude as those on the exam score, suggesting that the relative grading component in teacher grades is minor. The table also reports effect estimates based on fixed effects estimation for the teacher grades. The bootstrapped Wald test for the 2SLS results equals 1.85 with a p-value of 0.397. Like for the exam grades we do not reject equality of the estimates of the grade composition effects on teacher grades, increasing confidence in the validity of the IV results.

To summarize, we thus find that the effect of grade mixing starkly depends on the exact grade composition of the classroom. In our study students benefit on average from grade mixing. It is however important to point out that this not only depends on the positive effects outweighing the negative ones, but also on the specific grade mixing sequences students are exposed to. Interestingly, once we allow for this possibility we can reconcile some of the apparently contradictory findings in the literature. A recent example is Sims (2008), who finds a negative effect of the fraction of students in mixed-grade classrooms on the (average) achievement of 2nd and 3rd graders. His instrument – the number of classrooms that are saved by combining the current grade with lower grade pupils – suggests that the complier group consists of schools who combine 2nd or 3rd graders with pupils from lower grades to economize on the number of classrooms. In this case the estimate will be the local average treatment effect of being mixed with lower grade pupils which we expect to be negative. The positive effect of Thomas (2011) on the other hand can be explained because it is the effect for first graders of sharing the classroom with higher grade pupils, namely from 2nd grader.

5.2 Gender differences

To further investigate heterogeneity in these effects we also perform our estimations separately by gender. The first two columns of Table 6 show the effects of classroom grade composition first for girls and then for boys. We find in the first column large and significant effects for girls' exam scores. The positive effects again dominate the negative ones. The point estimates go in the same direction for boys, although the point estimate on the negative effect for lower grades is close to zero, and the positive effect for higher grades is not significant. When we test for equality of the effects across gender we can, however, not reject the null hypothesis that they are equal (p=0.22).

The last two columns show the results for the teacher test scores. Here we have positive and statistically significant effects of higher grades for both girls and boys. The estimates are also of the same order of magnitude. We again find negative effects, this time for both genders even though we lack precision for boys. We again do not reject equality of the effects across gender (p=0.43).

One interesting aspect of the results for teacher test scores is their size relative to those for the centralized exams. For girls the estimated effects are smaller, whereas for boys they are larger. Although the average results above gave no indication that relative grading mattered, the results for girls are consistent with this explanation. The results for boys are however more difficult to reconcile with relative grading because there we see the converse. In the end we cannot reject equality between the effects on the exam scores and the teacher scores for both sexes. Interpretation in terms of relative grading should done with caution, and a conservative take on our findings is that we find similar results for boys and girls.

We also investigated heterogeneity of the effects with respect to pupils' relative age. There is some indication that effects are larger in absolute size for the relatively

Table 6. Gender differences, 2SLS estimates

	Exam score		Teacher score	
	Girls	Boys	Girls	Boys
# Lower grades	-0.510	-0.019	-0.356	-0.199
	(0.167)***	(0.209)	(0.151)**	(0.158)
# Higher grades	0.644	0.278	0.362	0.431
	(0.187)***	(0.209)	(0.168)**	(0.170)**
Class size	0.004	-0.005	0.007	-0.006
	(0.008)	(0.007)	(0.006)	(0.006)
N schools	386	382	386	382
N	4642	4994	4642	4994

Note: All regressions include a constant term and the full set of controls in Table 4. Standard errors are heteroscedasticity robust and corrected for school-level clustering. */**/*** statistically significant at the 10/5/1 percent level.

younger students in the cohort. Because these interactions are imprecisely estimated and too inconclusive we do not report them.

6 Conclusion

To estimate the impact of classroom grade composition on pupil achievement we exploit discontinuous grade mixing rules in Norwegian junior high schools in an instrumental variables setup. Using high stake exit tests and teacher set and scored tests we find that pupils in combination classes perform slightly better that homogeneous single grade classrooms. This effect is driven by pupils benefiting from sharing the classroom with more mature peers from higher grades. We also find that the presence of lower grade peers decreases achievement. Further analysis shows some indication that effects are larger for girls.

Our results contribute to two strands of work. The first literature to which we contribute studies the nature and consequences of peer effects. A classroom becomes more heterogeneous when two or more grades are mixed. This opens the scope for direct negative or positive spillovers due to the presence of more or less able peers. Our results are consistent with such externalities. Classroom composition can also significantly affect teacher inputs. This is what Duflo et al. (2008) found in their experimental study of tracking in Kenya. In the end our estimates will capture all these externalities of classroom grade composition. We therefore consider our estimates to be policy relevant effects.

As emphasized by Duflo et al. (2008), the effects of tracking are likely to be context specific and our study is no exception. Classrooms in Norway are, for example, from the outset smaller and more homogenous than in Kenya. Teachers are also more qualified and experienced in handling heterogeneous classrooms. In this context we find positive (and not negative) impacts of de-tracking grades. Our results thus highlight that the dynamics of classrooms are complicated and less straightforward than often thought. This shows that extrapolation without common support can give misleading results, a point forcefully made by Carrell et al. (2011).

The second, and main, contribution of our paper concerns combination classes which, as we documented in the introduction, are an important mode of classroom organization around the world. We know however little about how time in such classes affects pupils's learning outcomes. Our results show that pupils can on average benefit from them, but we also find that this depends crucially on how the classroom is balanced in terms of lower and higher grades. We take this also as a cautionary tale. Pupils can be worse off if negative effects of lower grades cannot be countered with positive effects channeled by the presence of higher grades.

References

Ammermueller, A. and Pischke, J. (2009). Peer effects in european primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, 27(3):315–348.

Angrist, J. and Lang, K. (2004). Does school integration generate peer effects? evidence from boston's metco program. *American Economic Review*, 94(5):1613–1634.

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

Bedard, K. and Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics*, 121(4):1437–1472.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2010). Under pressure? the effect of peers on outcomes of young adults. Working Paper No. 16004, NBER.

Boozer, M. and Cacciola, S. (2001). Inside the'black box'of project STAR: Estimation of peer effects using experimental data. Working paper, Economic Growth Center, Yale University.

Cahan, S. and Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development*, 60(5):1239–1249.

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2011). From natural variation to optimal policy? the Lucas critique meets peer effects. Unpublished working paper, Department of Economics, UC Davis.

Duflo, E., Dupas, P., and Kremer, M. (2008). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. Working Paper No. 14475, NBER.

Fradette, A. and Lataille-Démoré, D. (2003). Les classes à niveaux multiples: point mort ou tremplin pour l'innovation pédagogique. *Revue des Sciences de l'Éducation*, 29(3):589–607.

Fredriksson, P. and Öckert, B. (2005). Is early learning really more productive? the effect of school starting age on school and labor market performance.

Hægeland, T., Raaum, O., and Salvanes, K. (2008). Pennies from heaven? using exogeneous tax variation to identify effects of school resources on pupil achievements. Discussion Paper 3561, Institute for the Study of Labor (IZA).

Heckman, J. and Smith, J. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, 9(2):85–110.

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation. Working Paper No. 7867, NBER.

Lavy, V., Paserman, M., and Schlosser, A. (2008). Inside the black box of ability peer effects: Evidence from variation in low achievers in the classroom. Working Paper No. 14415, NBER.

Leuven, E., Lindahl, M., Oosterbeek, H., and Webbink, H. (2010). Expanding schooling opportunities for 4-year-olds. *Economics of Education Review*, 29:319–328.

Leuven, E., Oosterbeek, H., and Rønning, M. (2008). Quasi-experimental estimates of the effect of class size on achievement in norway. *Scandinavian Journal of Economics*, 110(4):663–693.

Little, A. W. (2004). Learning and teaching in multigrade settings. Paper prepared for the UNESCO 2005 EFA Monitoring Report.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

Mason, D. and Burns, R. (1997). Reassessing the effects of combination classes. *Educational Research and Evaluation.*

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Mulryan-Kyne, C. (2005). The grouping practices of teachers in small two-teacher primary schools in the Republic of Ireland. *Journal of Research in Rural Education*, 20(17):20–17.

Rockoff, J. and Lockwood, B. (2010). Stuck in the middle: Impacts of grade configuration in public schools. *Journal of Public Economics*, 94:1051–1061.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1):175–214.

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.

Sojourner, A. (2008). Inference on peer effects with missing peer data: Evidence from project STAR. *Unpublished manuscript, Department of Economics, Northwestern University.*

Strøm, B. (2004). Student achievement and birthday effects. *Unpublished manuscript, Norwegian University of Science and Technology.*

Thomas, J. L. (2011). Combination classes and educational achievement. Unpublished working paper, Department of Economics, UC San Diego.

Urquiola, M. and Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, 99(1):179–215.

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.

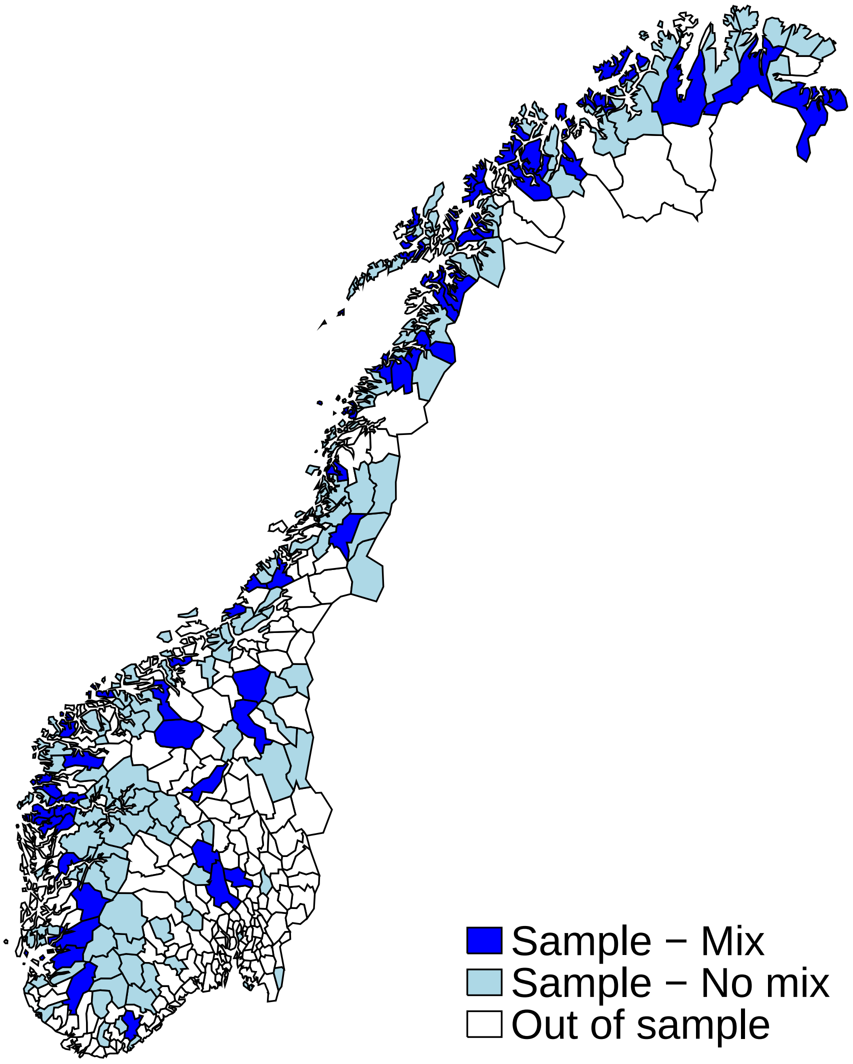


Figure A1. Regional coverage