THE ECONOMIC JOURNAL

The Economic Journal, **126** (*June*), 1164–1192. Doi: 10.1111/ecoj.12177 © 2014 Royal Economic Society. Published by John Wiley & Sons, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

CLASSROOM GRADE COMPOSITION AND PUPIL ACHIEVEMENT*

Edwin Leuven and Marte Rønning

This article 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. This finding 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.

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 industrialised countries (Little, 2004).[1] In 2007, about 28% of schools in the US report 'using multi-age grouping to organise most classes or most pupils'.[2] Similarly, in 2001 about 25% 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% 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 enrolment, 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 areas. For example, 32% of American public schools located in cities report using multi-age grouping, compared to 26% 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

* Corresponding author: Edwin Leuven, Department of Economics, University of Oslo, PO Box 1095 Blindern, 0317 Oslo, Norway. Email: edwin.leuven@econ.uio.no.

We thank Adam Booij, Eric Bettinger, Julie Cullen, Monique De Haan, Pascaline Dupas, Tarjei Havnes, Magne Mogstad, Hessel Oosterbeek, Holger Sieg, David Sims, seminar participants, an editor and two anonymous referees for valuable comments. A special thanks to Maria Fitzpatrick for providing us with descriptive statistics from SASS 2007. Leuven is also affiliated with the CEPR, CESifo, ESOP, IZA and Statistics Norway. The usual disclaimer applies.

[1] Multigrade and multi-age can correspond to different educational practices when age and grade do not coincide. In most industrialised countries there is a close correspondence between age and grade, in which case the distinction bears little practical meaning. This seems to be the common interpretation of multi-age grouping in education circles, see e.g. Mariano and Kirby (2009) and the references therein.

[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'Èducation Nationale.

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 his age and length of schooling, and therefore with cognitive development and achievement (Bedard and Dhuey, 2006; Leuven *et al.*, 2010; Fredriksson and Öckert, 2013). In addition, peers from higher grades can serve as role models in terms of non-academic behaviour, which can feed back to school achievement. Finally, classrooms' grade composition can also significantly affect teacher inputs and teaching methods.

There is surprisingly little 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 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 (Hoxby, 2000; Ammermueller and Pischke, 2009; Lavy *et al.*, 2012*b*; Black *et al.*, 2013). 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, 2013). 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.*, 2011). Social experiments are however rare and have their own limitations (Heckman and Smith, 1995); quasi-experiments are an interesting alternative (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 (2012) 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.

This study sets out to estimate how classroom grade composition in Norway affects pupil achievement and presents a number of significant contributions to the literature. First, we use a novel identification approach that exploits institutional features in Norway that significantly change the grade composition of classrooms. Norwegian junior high schools are bound by national regulation that uses enrolment by grade level to determine classroom grade composition. These rules determine predicted grade mixing which we use as instruments for actual grade mixing. Second, the

institutional features allow us to both instrument for grade composition and class size. The third contribution of this study is that we separate the average effect of grade mixing into that of sharing the class room with lower grades *versus* higher grades.

To summarise our results briefly, we find evidence that a one-year exposure to a classroom that combines two grade levels increases examination performance by about 4% 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 we find evidence that 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 spent 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 than 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. We argue below that these results can go a long way towards explaining the contradictory findings in the literature.

In what follows we start by describing the institutional context and our data sources. After outlining our empirical approach in Section 3, we present our estimation results in Section 4 and discuss how classroom age composition affects pupil achievement on the short and long term. Section 5 concludes.

1. Institutional Settings and Data

1.1. *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 enrolment – and there are no school fees. Schools are governed at the local school district level by the municipality and have catchment areas, implying that, given residence, there is no parental school choice.[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. In general, grade retention is strongly related to maturity (Cahan and Cohen, 1989) and if schools would practice grade retention then this could introduce an extra endogenous margin of classrooms' ability composition. However, as shown in Strøm (2004) and Bedard and Dhuey (2006), there is no grade retention in Norway. As a consequence, nearly everybody starts junior high school the year they turn fourteen.

[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 2% did not start primary school the 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.

Our analysis focuses on small comprehensive 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 comprehensive, most of which are located outside the four major cities.[6] Since most of these schools are relatively small, it is common practice to combine multiple grades in a single classroom. All junior high schools in Norway – including the comprehensive 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.

1.2. *Data*

We use administrative enrolment data (provided by Statistics Norway) on all pupils who graduated from junior high school in the school years 2001/02 and 2002/03. 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, socio-economic 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 teacher set and graded tests in the final year, as well as centralised exit examinations (from Statistics Norway). At the end of the final year in junior high school, all pupils in Norway are required to take an exit examination. Although the curriculum includes many subjects, a written exit examination is only undertaken in one of three subjects: mathematics, Norwegian and English. The examinations are centrally assigned and it is not known in advance what the examination topic will be and are, therefore, beyond the control of schools, teachers and pupils. In the analysis, we pool these three subjects and standardise them with zero mean and standard deviation one. The teacher tests as well as the examination scores are used to construct pupils' junior high school exit test scores which are important for secondary school choice. For students, both the examination and teachers scores are, therefore, important because they are used for tracking decisions.

The correlation between the teacher score and the examination score is 0.8. Although both the examination 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 examinations 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

[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.

than the examination scores which are based on a single test. One 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. Contrasting results based on teacher tests and examination test can, therefore, tell us something about the importance of relative grading. We discuss these issues in more detail in the context of our results below.

Grade mixing mostly occurs outside the major cities in comprehensive schools. We, therefore, restrict our population of interest to these comprehensive schools. 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 predicted 7th grade class room when pupils start junior high school.

Our analysis data set consists of *circa* 10,000 pupils and 400 schools. This amounts to about 10% of the pupil population and two out of five schools in Norway. About 200 schools, one out five 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 them to the total population of junior high school pupils. Relative age – which equals 0 for the youngest pupil (born 31st December) and 1 for the relatively oldest one (born 1st January) – is on average 0.5. This implies that pupils in their final year of junior high school in Norway are on average 15.5 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 considerably larger in the whole population compared to the comprehensive 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 ($p = 0.295$). Again, and – as we will show below – by virtue of the institutional rules, the mixing schools are smaller with smaller classes.

1.3. *Grade-mixing in Norway*

Grade mixing in Norway is the practice of combining pupils from different grade levels in one classroom. In these classrooms, the curriculum is graded (8th graders are taught 8th grade curriculum etc.). Contrary to many other countries, middle school students in Norway receive their instruction in all subjects in this classroom.

Table 1

Descriptive Statistics

	All schools		Small schools		Mixing schools	
	Mean	SD	Mean	SD	Mean	SD
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	0.48	0.50
Parental characteristics						
Mother's education						
Junior high school or less (≤ 10)	0.11	0.31	0.12	0.32	0.14	0.34
High schools (11–13)	0.56	0.50	0.64	0.48	0.63	0.48
College (14+)	0.30	0.46	0.22	0.41	0.21	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	0.61	0.49
College (14+)	0.28	0.45	0.18	0.39	0.16	0.37
Cohabiting	0.67	0.47	0.71	0.45	0.70	0.46
N observations	98,254		9,821		2,262	
School characteristics						
Comprehensive school	0.53	0.50	1		1	
School size	152.9	119.1	40.9	24.1	20.5	9.4
Class size	21.0	5.9	16.0	5.6	12.9	3.4
Teacher hours per pupil	98.2	38.6	128.7	44.3	156.2	47.1
N schools	1,040		414		195	

Each class has a main (contact) teacher but pupils are often taught by specialised teachers in the main subjects, such as mathematics, Norwegian and English, but sometimes also in history and other subjects. The small schools that combine grades studied in this article have typically up to three teachers, and almost never more than four. Given the size of these schools, these specialised course teachers will most likely teach all students in the school. This greatly reduces the scope for teacher sorting.

Grade mixing thus affects (by definition) the student composition of the classroom but can also change instruction. Like the class size literature, the effects we will estimate are reduced form in the sense that they capture the sum of these potential mechanisms.

To provide some more context to our estimates, we contacted about 400 small comprehensive middle schools by email and sent them a link to a small online survey. The questionnaire asked about classroom organisation in mixed grade classes, and explicitly referred to the time period covered by our sample (the early 2000s). The response rate was quite low, as we received only complete responses from 62 schools (about one out of six contacted schools).[7] To investigate non-response we estimated a probit that included student characteristics (gender, age, parental education, parental cohabiting) and school characteristics (school size, class size and grade mixing). The student characteristics are highly insignificant ($p = 0.445$). The school characteristics are marginally significant at the 10% level ($p = 0.097$). However, a joint test on all explanatory variables fails to reject the null ($p = 0.160$) which suggests that, even though the sample is small, it is fairly representative.

The results from the survey are reported in Table 2. There can be one, two or three grades in a single classroom and we asked for each of these contingencies what the typical classroom practice would have been in the surveyed school.

Table 2

Grade-mixing and Classroom Organisation

	No. of grades in classroom		
	One	Two	Three
How many teachers in the classroom?			
One	91	83	47
Two	9	17	53
How do pupils typically work?			
Individual or small groups	52	74	75
Whole class	48	26	25
How are pupils grouped?			
No systematic grouping	65	39	39
Friends sit together	2	0	3
Students of similar level or grade	33	61	58

Note. Based on survey responses from 62 school administrators.

[7] This is probably explained in part because we asked about conditions that took place about ten years ago (the grade mixing rules that we are using ceased after the school year 2002/03).

The first question concerned the number of teachers. Single grade classrooms have typically one teacher. In classrooms that mix grades there is more often a second teacher present. This is in particular true for classrooms that combine three grades. There, 53% of the schools report that two teachers would be present. We also gave schools the possibility to expand on their answer in an open question. Here schools often reported that even though more than a single teacher could be present in the classroom, this would depend on the need of the students and could vary across courses.

We also asked, for each grade mixing contingency, how it affected instruction. More specifically, we asked how students would work: individually, in small groups, or as a whole class. Compared to single grade classrooms, schools report that students are more likely to work individually or in small groups than together with the whole class. In the open question, many schools noted that these practices could vary substantially depending on the task at hand. Nevertheless, the responses suggest that there is substantially less common instruction in mixed grade classrooms, which would be expected since the curriculum is graded.

Finally, we asked how students would typically be seated in class: without a system, with friends or with students of the same level or grade. We see that for single grade classrooms the large majority of schools state that there is no systematic grouping. In mixed grade classrooms on the other hand, schools are more likely to report that they would physically seat students from a similar level or grade together. This would of course facilitate the more individual or group mode of working that school reports in the previous question.

To summarise, we see that grade mixing affects not only classroom composition but also instruction. Since the curriculum remains graded, there is more individual or group level work compared to single grade classes where instruction is more often at the whole class level. These changes in instruction are also reflected in the physical organisation of the classroom, where students are more likely to be seated in proximity of similar peers.

2. Maximum Class Size Rules

Junior high schools in Norway were subject to maximum class size rules (Leuven *et al.*, 2008). What makes these rules unique is that they sometimes 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

(*a*) 30 pupils when there is one cohort in the class
(*b*) 24 pupils when there are two cohorts in the class
(*c*) 18 pupils when there are three cohorts in the class

2 When there are multiple cohorts in a class, they need to be adjacent if possible
3 The school cannot simultaneously have mixed age and age-homogeneous classes within the same grade level, or parallel mixed age classes

Schools are supposed to follow the Education Act. Rule 1(*a*) requires schools to open an extra classroom if enrolment in a single grade classroom would exceed 30. This rule is similar to the familiar Maimonides rule, exploited first by Angrist and Lavy (1999), and for Norway by Leuven *et al.* (2008).

Where these rules are different is that they affect not only class size but also class grade composition. A school with no more than 18 pupils is expected to have a single classroom that combines students from three grades. If enrolment is greater than 18 a school will need to have two classrooms, where one will combine two grades as long as their combined enrolment does not exceed 24. Beyond 24 schools are supposed to have only single graded classrooms. Figure 1(*a*) illustrates this predicted grade mixing as a function of school size. The decision whether to combine three grades in a classroom or not, depends only on their combined enrolment not exceeding 18. Figure 1(*a*) applies to each school year.

Figure 1(*b*) shows the contemporaneous relationship between school size and multiple grade classrooms that we observe in our data. The *x*-axis in Figure 1(*b*) 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 no longer supposed to combine all three

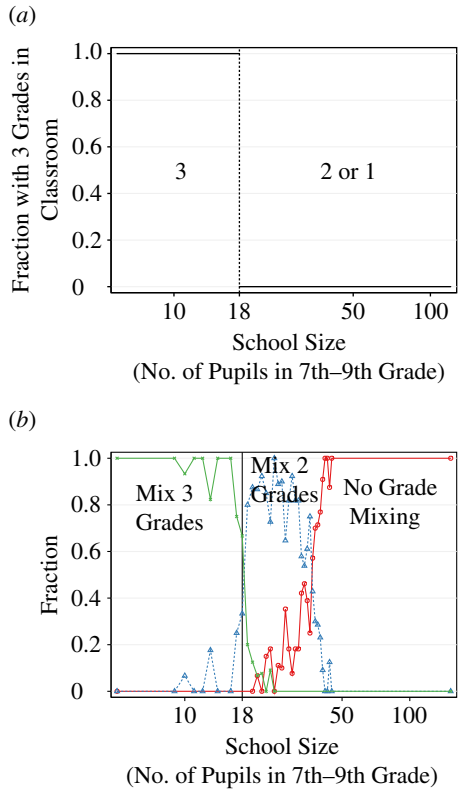


Fig. 1. *Classroom Grade Composition by School Size*
Notes. (*a*) Predicted. (*b*) Actual, 7th grade.

grades in a single classroom. We see that there is a close relationship between actual grade mixing and the grade mixing rule. The propensity to combine three grades drops sharply by about 0.5 after school size 18. Where to the left of this threshold schools essentially mix all three grades, for schools larger than 18 pupils the picture is, as expected, 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, namely, whether their combined enrolment exceeds 24 or not. For schools larger than 50 pupils there is no longer any grade mixing taking place.

How schools exactly combine two grades in a single classroom is somewhat more involved because of the requirement that these grades need to be adjacent. This means that schools can either combine 7th and 8th graders in a single classroom, or 8th and 9th graders. On top of this there is the constraint that schools are not supposed to have two (or more) identical combined 7th/8th (or 8th/9th) grade classrooms. Finally, schools are not supposed to have, e.g. a single graded classroom, say 7th grade, when they have a mixed grade classroom that contains students from grade 7. As a consequence, similar enrolment patterns will have different implications for the classroom composition depending on the grade a student is in. Figure 2 illustrates in detail how schools are predicted to apply these rules.

When the combined enrolment of both 7th/8th grade and 8th/9th grade exceeds 24, schools are predicted to have only single grade classrooms. This corresponds to the top-right quadrants in Figures 2(*a*)–(*c*). Here grade mixing does not depend on the grade a student is in. Next, consider the top-left quadrant where 7th/8th grade enrolment does not exceed 24 but the combined enrolment of 8th/9th grade is greater than 24. In this case, we predict a single 9th grade classroom and a combined 7th/8th grade classroom. Seventh and 8th graders are, therefore, expected to find themselves in a classroom with two grades, as indicated in Figure 2(*a*)–(*b*), and 9th graders in a single grade classroom, as in Figure 2(*c*). Similar reasoning follows for the bottom-right quadrant when the combined enrolment of 8th/9th grade does not exceed 24, but 7th/8th grade enrolment does.

When both 7th/8th and 8th/9th grade enrolment is not larger than 24, schools are in principle confronted with the choice of either forming a combined 7th/8th grade classroom together with a separate 9th grade classroom, or a 8th/9th grade classroom with a separate 7th grade classroom. We predict that schools try to keep the combined classroom as small as possible, and therefore choose a separate 7th grade classroom (rather than a 9th grade classroom) if there are more 7th than 9th graders. This means that above the diagonal in the bottom-left quadrant – where there are as many 7th as 9th graders – we predict seeing a combined 7th/8th grade classroom and below the diagonal a combined 8th/9th grade classroom.

Figure 3 shows actual grade mixing as a function of the relevant cohort sizes and how schools go from a double to a single grade classroom. Most schools find themselves in either the top-right or bottom-left quadrant. In the top-right quadrant the rules stipulate that grades are not to be combined, which is indeed what we observe in the data, with a few exceptions these are all regular single grade classrooms. In the bottom-left quadrant schools are predicted to combine two grades. Again, schools do not always follow the rules but, as predicted, we see that 7th graders are more likely to be in a mixed classroom above the diagonal when 9th grade enrolment is larger than that of

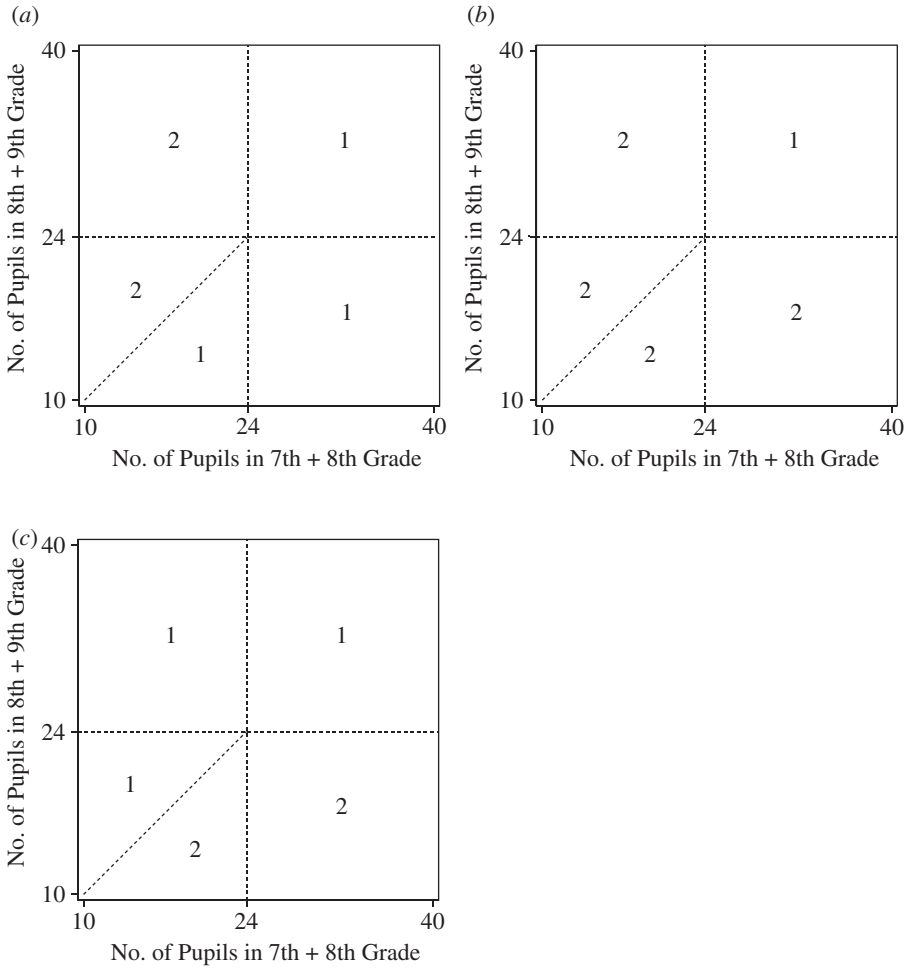


Fig. 2. *Predicted Grade Mixing by Grade Level When School Size > 18*
Notes. (*a*) For 7th graders. (*b*) For 8th graders. (*c*) For 9th graders.

7th grade enrolment. Similarly, we see 9th graders more often in a combined classroom below the diagonal. Lastly, and consistent with the rest, 8th graders are typically in a mixed grade classroom when enrolment puts a school in the bottom-left quadrant.

In the top-left and bottom-right quadrants, we predicted different grade mixing depending on the grade students are in. Seventh graders are predicted to be in a mixed grade classroom in the top-left quadrant and in a single grade classroom in the bottom-right quadrant. For 9th graders we predicted the reverse. Eighth graders are in both cases predicted to be in a mixed grade classroom. This is indeed what we observe, although again some schools deviate from the rules.

We will now outline how we exploit these institutional rules that govern grade mixing decision in an instrumental variables framework.

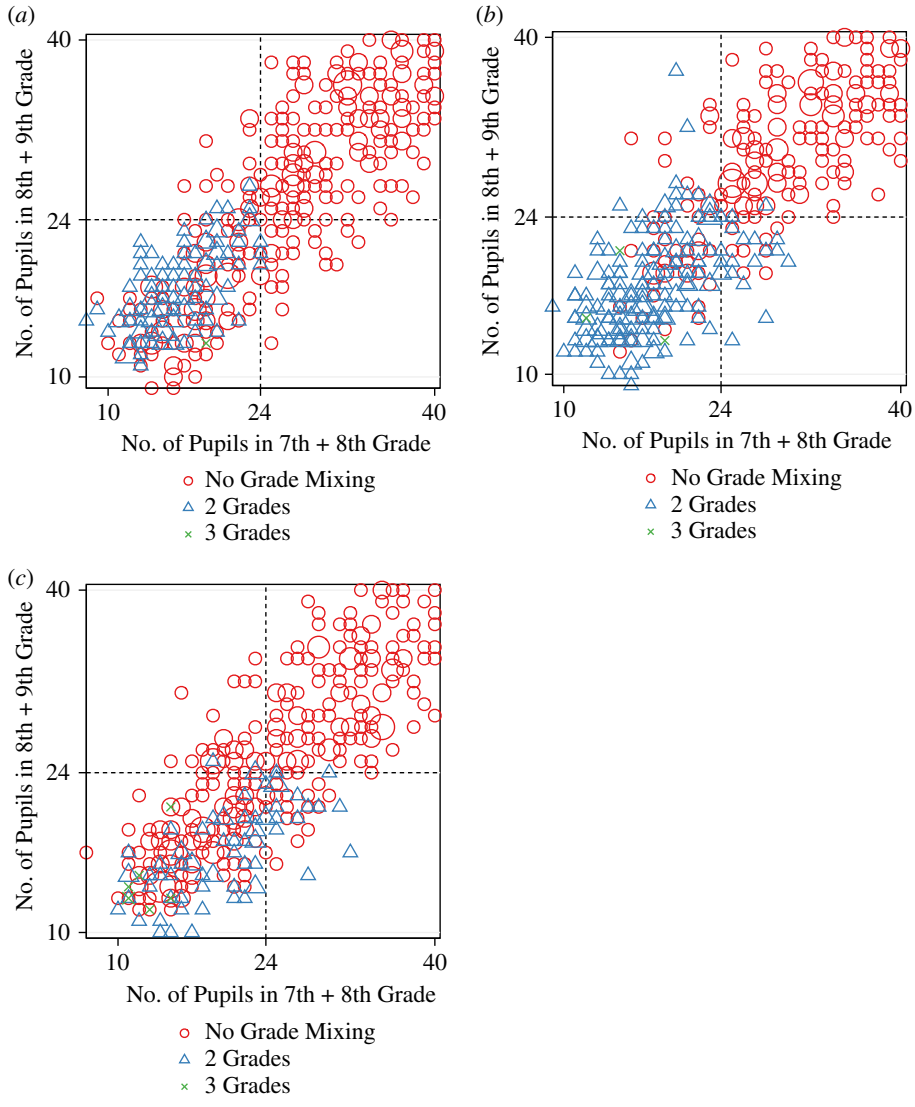


Fig. 3. *Actual Grade Mixing by Grade Level When School Size > 18*
Notes. (*a*) For 7th graders. (*b*) For 8th graders. (*c*) For 9th graders.

3. Empirical Strategy

3.1. *Estimation Approach*

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 study 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:

$$y_i = \alpha \times g_i + \gamma \times ssize_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad (1)$$

where y_i is pupil i's achievement at the end of junior high school. 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 $g = 1$. For example, if they were mixed in 7th grade with 8th graders, and not mixed in grades 8 and 9, then $g = (2 + 1 + 1)/3 = 4/3$, etc. This is a natural parametrisation of the policy that we study, which acts on the raw number of grades in the classroom as illustrated by the graphs above. Since the policy changes the raw number of grades, estimating the effect of the raw number of grades, therefore, delivers policy relevant average effects. We also add school and family control variables in $\boldsymbol{x}_i$, which include parental education, whether parents are living together, pupils' gender and relative age.[8]

As documented above, grade mixing is governed by the rules set by the Ministry of Education but 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.

We follow an instrumental variable approach in the spirit of Angrist and Lavy (1999) and use the predicted grade mixing documented in Figures 1(*a*) and 2 to construct instruments for actual grade mixing to take any remaining endogeneity into account. More in particular, using the enrolment of 7th, 8th and 9th graders in a given school year we can determine the predicted grade mixing according to the rules. For each student i we calculate the predicted grade mixing separately for each grade level j when she was in junior high school. Predicted grade mixing for student i in grade j, $\mathrm{E}(g_{ij})$, is defined in Table 3 where n_{it}^j is the number of j-th graders in student i's school in year t.

Table 3

Predicted Grade Mixing $\mathrm{E}(g_{ij})$ *as a Function of Grade Specific Enrolment*

	Student is in		
	7th grade	8th grade	9th grade
$n_{ij}^7 + n_{ij}^8 + n_{ij}^9 \leq 18$	3	3	3
$n_{ij}^7 + n_{ij}^8 + n_{ij}^9 > 18$ and:			
$n_{ij}^7 + n_{ij}^8 \leq 24, n_{ij}^8 + n_{ij}^9 > n_{ij}^7 + n_{ij}^8$	2	2	1
$n_{ij}^7 + n_{ij}^8 \leq 24, n_{ij}^8 + n_{ij}^9 \leq n_{ij}^7 + n_{ij}^8$	1	2	2
$n_{ij}^8 + n_{ij}^9 \leq 24, n_{ij}^7 + n_{ij}^8 > n_{ij}^8 + n_{ij}^9$	1	2	2
$n_{ij}^8 + n_{ij}^9 \leq 24, n_{ij}^7 + n_{ij}^8 \leq n_{ij}^8 + n_{ij}^9$	2	2	1
$n_{ij}^7 + n_{ij}^8, n_{ij}^8 + n_{ij}^9 > 24$	1	1	1

[8] Relative age = (1 − day of birth) / 364, so that the relatively oldest pupil has age 1 and the youngest age 0. We also estimated specifications where we instrument actual age using relative age as in Bedard and Dhuey (2006) and Black *et al.* (2013). This does not affect our results. We report estimation results from reduced form models with respect to age for simplicity.

In our 2SLS estimation, we use six predicted grade mixing dummies, one for each grade and value of $E(g_{ij})$, leaving out the reference group of no grade mixing. The first stage for average grade mixing in junior high school thus becomes

$$g_i = \sum_{j=7}^{9} \sum_{n=2}^{3} \delta_{jn} \mathbb{1}_{[E(g_{ij})=n]} + \delta_s \times ssize_i + \boldsymbol{x}_i' \boldsymbol{\delta}_x + u_i. \quad (2)$$

We control throughout for school size ($ssize_i$), the combined enrolment of 7th, 8th and 9th grade, when the pupil started junior high school. School size can be thought of as a running variable and potential confounder.[9]

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 he was in junior high school. For example, when mixed with 8th and 9th when in 7th grade, and not mixed afterwards 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

$$y_i = \alpha_+ g_i^+ + \alpha_- g_i^- + \lambda \times ssize_i + \boldsymbol{x}_{ij}' \boldsymbol{\beta} + \varepsilon_1, \quad (3)$$

where we instrument both g_i^+ and g_i^- with the same set of instruments as in (2).[10,11]

The variation that allows us to estimate α_+ and α_- separately is illustrated in Table 4, 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').[12]

[9] The relationship between school size and test scores in the total population is linear with a slope coefficient close to zero. Moreover, our results change little with a quadratic or cubic specification in school size, with the exception that the first stages become weaker. We, therefore, report results based on a linear specification.

[10] We also estimated specifications based on binary variables for being mixed, and being mixed with pupils from lower *versus* 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 the text are scaled in terms of number of grades and are therefore more straightforward to interpret quantitatively.

[11] Grade-specific effects are qualitatively similar to the estimates based on (3). They lack precision, however, and we cannot reject that the 2SLS is under-identified. To estimate grade-specific effects we would require substantially larger samples than we have and, preferably, test scores measured at the end of every grade (as opposed to only at the end of middle school).

[12] Although in theory the individual sequences could be used as instruments, the small cells lead to over-fitting in the first-stage, and therefore biases the 2SLS towards OLS. For this reason our instrument set aggregates these individual sequences.

Table 4

Classroom Count of Observed Grade Mixing Sequences

Sequence	*N*	Sequence	*N*	Sequence	*N*
111	412	221	71	311	1
112	4	222	25	313	4
121	49	223	7	321	5
122	40	231	1	322	5
123	9	232	2	323	6
132	2	233	10	331	3
133	3			332	5
				333	100
				Total	764

Notes. 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.

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 article but excluding the comprehensive 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 in mixed 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 average class size when in junior high school – the same class size measure as in Leuven *et al.* (2008).

Predicted class size, our class size instrument, is the analogue of the standard instrument that is used in the class size literature and is defined as follows:

$$\mathrm{E}(csize_i) = n_{i7}^{7} + (n_{i7}^{8} + n_{i7}^{9}) \times \mathbb{1}_{[\mathrm{E}(g_{i7})=3]} + 0.5 n_{i7}^{8} \times \mathbb{1}_{[\mathrm{E}(g_{i7})=2]}. \quad (4)$$

Equation (4) implies that in a single grade class the expected class size on junior high school start is simply the number of 7th grades: n_{i7}^{7}. When all three grades are predicted to be mixed ($\mathrm{E}(g_{i7}) = 3$), expected class size is the number of 7th–10th grades: $n_{i7}^{7} + n_{i7}^{8} + n_{i7}^{9}$. When two grades are predicted to be combined ($\mathrm{E}(g_{i7}) = 2$), this can either be 7th and 8th grade or 8th and 9th grade. In the first case, the expected class size in 7th grade is the number of 7th and 8th grades, $n_{i7}^{7} + n_{i7}^{8}$. The second case expected class size is the number of 7th grades, n_{i7}^{7}, since the 8th and 9th graders are in a separate mixed classroom. We assume that these two cases have equal probability (0.5). Combining the different scenarios 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.

With a single discontinuity it would not be possible to separately estimate grade mixing and class size effects. However, the rules generate many discontinuities. We can separate the grade mixing and class size effects because – for a given drop in class size – these discontinuities differ in the way they affect classrooms' grade composition. The grade mixing and class size effects are therefore identified by pooling the discontinuities and relying on homogeneity of the class size effect across discontinuities. In our setup, this is essentially achieved by the separable specification (which one can also think of as a first order Taylor expansion of the underlying structural function).

The first stages we report below show that given this specification, the grade mixing and class size effects are well identified in the data. The class size instrument almost exclusively loads on class size and the mixing instruments almost exclusively on grade mixing. That our specification is reasonable is supported by our results which are extremely robust: the estimated grade mixing effects are essentially unchanged

(*i*) with and without controlling for class size; and
(*ii*) with and without instrumenting for class size.

Moreover, the coefficient on class size is insignificant and extremely small. It is hard to think of a scenario where class size is an important omitted variable that biases our grade mixing estimates but that gives us zero class size effects and unchanged grade mixing effects in the wide range of specifications that we report.

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 is that there are alternative confounding changes of related school inputs. We discuss each in turn.

3.2. *Sorting*

We can distinguish between two main sources of sorting. The first is supply side sorting which arises when schools or local education authorities manipulate enrolment 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 redraw school catchment areas sometimes 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 cannot be changed in response to annual enrolment fluctuations in Norway.

The second potential source of sorting comes from the demand side. When parents prefer mixing or non-mixing classrooms they might decide to enrol their children in a different school. If, for example, more advantaged families sort in different ways from disadvantaged families, the underlying pupil population at both sides of the discontinuities is 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. One may however be concerned by the possibility of families moving to different school catchment areas in reaction to or anticipating classroom grade composition during high school. Hægeland *et al.* (2012), who use the same pupil data as we do, report that in Norway as a whole 95.3% of the pupils lived in their graduation municipality throughout all three years in junior high schools. We can implement a check by comparing the administrative head counts for 7th and 8th grade with the 9th grade head counts when these 7th and 8th graders are supposed to be in 9th grade (unless they 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 during high school are not a concern in our data.

To see whether there is any indication of parents sorting prior to the start of junior high school, we also check whether we can detect discontinuities in the enrolment densities. We follow McCrary (2008) and calculate these discontinuities using local linear regression techniques using optimal bandwidths. Figure 4 pools the different years in our data and shows density plots for the three discontinuities that we exploit in the analysis. The Figure 4(*a*) shows total junior high school enrolment where the discontinuity is at 18. As can be seen from the graph, the density peaks around enrolment 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 0.28 and not statistically significant. Figure 4(*b*) shows a similar graph for combined enrolment of 7th and 8th graders where the discontinuity lies at 24. Here the estimated density is also higher to the right of the discontinuity and again not statistically significant. Finally, the Figure 4(*c*) shows the estimated discontinuity for the combined enrolment 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.

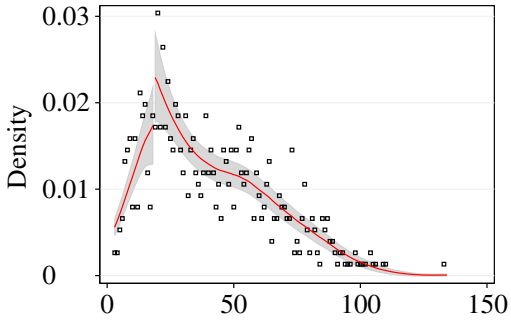
These results are probably not too surprising since the school districts in our data are rural and have typically one school, with the next school often a long car drive away. Since Norway has catchment areas, parents would often need to move to another municipality in order to enrol their child in another school. They would need to find new employment or face a long commute and the economic and social cost of sorting is, therefore, probably very high.

3.3. *Confounding Discontinuities*

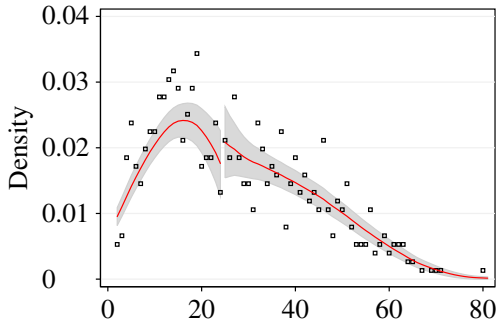
3.3.1. *Class size/pupil–teacher ratio*

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 enrolment fixed, combining grades involves less classrooms and therefore mechanically larger classes. This is illustrated in Figure 5(*a*) which plots the data points corresponding to the schools in our sample and a smoothed regression line and confidence interval at both sides of the discontinuity. Since we are interested in estimating the causal effect of changing the classroom grade composition, we need to

(*a*) Note: Discontinuity Estimate (Log Difference in Height): 0.28 (0.18)



(*b*) Note: Discontinuity Estimate (Log Difference in Height): 0.23 (0.21)



(*c*) Note: Discontinuity Estimate (Log Difference in Height): –0.15 (0.23)

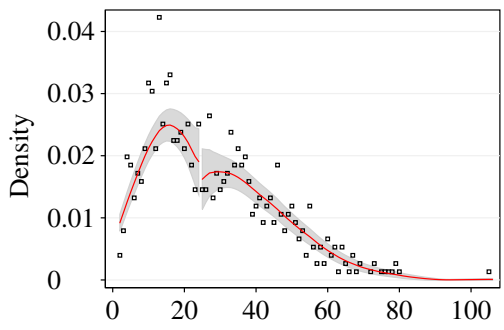


Fig. 4. *Density Checks*

Notes. (*a*) Pooled 7th, 8th & 9th grader enrolment. (*b*) Pooled 7th & 8th grader enrolment. (*c*) Pooled 8th & 9th grader enrolment.

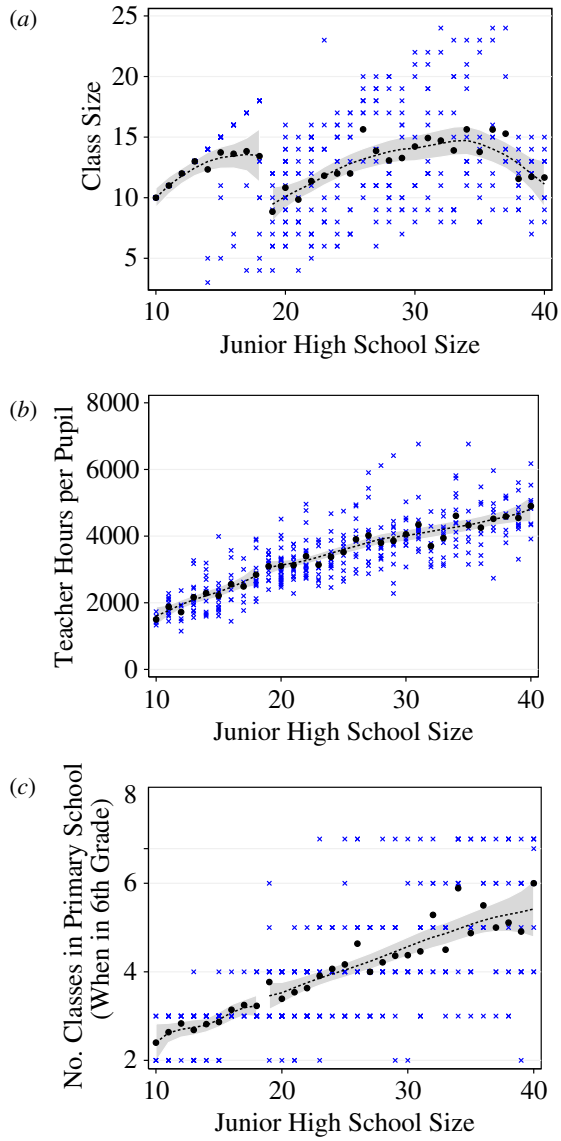


Fig. 5. *Confounding Discontinuities*

Notes. (*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.

keep the pupil–teacher ratio constant. This implies that we need to control for class size in our specifications.

From our administrative data, we know the ratio of teacher hours per pupil at the junior high school level. Figure 5(*b*) 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. This would remove the need to control for class size in order to estimate the

ceteris paribus effect of changing classroom grade composition. The results of Leuven *et al.* (2008) also suggest that there is no need to control for class size – although for a different reason – since they did not find evidence that class size affects achievement in Norwegian junior high schools and can rule out small effects.

The population of schools in the current study is however different and, also, the variation in class size is at smaller class size levels than in Leuven *et al.* (2008). Furthermore, we are also not certain that teacher hours are indeed balanced in the classrooms that we compare because our data do not allow us to link teachers to classrooms. To address these concerns, we control for class size when estimating how grade mixing affects achievement. As discussed above, when we control for class size, it is instrumented with predicted class size at the start of junior high school as in Angrist and Lavy (1999). It turns out that the estimated effects of grade composition are insensitive to whether or not we control for class size. Moreover, we do not find evidence of class size effects. This is consistent with the balancing of teacher hours shown in Figure 5(*b*).

Finally, one may be concerned that there is an independent effect of the number of teachers present in the classroom *conditional* on teacher hours per pupil. Although we cannot rule this out, previous research from the Tennessee STAR experiment did not find that having a teaching aide in the classroom improved student outcomes (Krueger, 1999).

3.3.2. *Classroom composition in primary school*

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 different and 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 5(*c*) shows that there is no evidence of such a confounding discontinuity.[13]

4. 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 examination and teacher test scores at the end of junior high school. After these overall results we present separate effect estimates for boys and girls.

4.1. *Examination and Teacher Scores*

The results from estimating (1) by OLS are shown in the first and fourth columns of Table 6. The first column is a simple regression of standardised examination scores

[13] An analysis of grade mixing in primary is not possible because we cannot reconstruct the complete grade mixing histories for our cohorts, and there are no test scores available at the primary level.

Table 5

First Stage Regressions

	No. of grades	No. of lower grades	No. of higher grades	Class size
Predicted no. of grades				
Three in 7th grade	0.528***	0.031	0.497***	−5.491***
	(0.074)	(0.045)	(0.055)	(0.657)
Two in 7th grade	0.067*	−0.001	0.068***	0.578
	(0.039)	(0.028)	(0.025)	(0.366)
Three in 8th grade	0.789***	0.417***	0.372***	0.817
	(0.070)	(0.066)	(0.049)	(0.586)
Two in 8th grade	0.348***	0.179***	0.169***	0.469
	(0.035)	(0.024)	(0.023)	(0.460)
Three in 9th grade	0.531***	0.498***	0.033	1.588***
	(0.068)	(0.062)	(0.041)	(0.504)
Two in 9th grade	0.041	0.073**	−0.033	0.029
	(0.043)	(0.030)	(0.027)	(0.416)
Predicted class size	−0.001	−0.001	−0.001	0.751***
	(0.001)	(0.001)	(0.001)	(0.055)
Relative age	0.003	0.006*	−0.004	0.006
	(0.006)	(0.004)	(0.004)	(0.091)
Girl	0.001	−0.003	0.004	−0.130**
	(0.004)	(0.003)	(0.002)	(0.064)
M – high school	−0.005	−0.004	−0.001	−0.058
	(0.007)	(0.005)	(0.004)	(0.106)
M – college	−0.009	−0.006	−0.003	−0.092
	(0.008)	(0.005)	(0.004)	(0.133)
F – high school	−0.005	−0.007*	0.002	−0.160*
	(0.005)	(0.004)	(0.004)	(0.084)
F – college	−0.002	−0.006	0.004	−0.113
	(0.006)	(0.004)	(0.004)	(0.101)
Parents cohabit	0.008	0.002	0.006*	0.208***
	(0.005)	(0.004)	(0.003)	(0.079)
School size/100	−0.062	−0.014	−0.048*	−1.359
	(0.038)	(0.024)	(0.026)	(1.686)
First stage F-*statistics*				
All instruments	394.6	283.5	245.3	32.8
Predicted grade mixing dummies	446.2	330.6	268.9	14.8
Predicted class size	1.2	1.0	0.4	187.9
Joint F test ind. char. (p)	0.579	0.371	0.430	0.025

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

on average classroom grade composition during junior high school, while controlling for class size, school size and our family background characteristics. This shows that pupils who have been in classes with one more grade level in their class during junior high school perform approximately 7% of a standard deviation better on the examination. The corresponding estimate for teacher score in the fourth column is 5% of a standard deviation. We, therefore, find a modest but statistically significant gradient between grade mixing and student achievement at the end of junior high school.

Although the OLS estimates are suggestive, we address potential endogeneity of grade mixing by estimating these effects using 2SLS as outlined above. Table 5

Table 6

The Relationship Between Grade Mixing and Pupil Performance, Dependent Variable is the Examination Scores – OLS & 2SLS

	Examination score			Teacher score		
	OLS	2SLS	2SLS	OLS	2SLS	2SLS
No. of grades	0.074**	0.091**	0.088**	0.047*	0.069**	0.070**
	(0.036)	(0.041)	(0.041)	(0.025)	(0.028)	(0.028)
Class size	0.002	0.002	0.001	0.000	−0.000	−0.001
	(0.004)	(0.004)	(0.005)	(0.003)	(0.003)	(0.004)
School size/100	−0.064	−0.038	−0.026	−0.202**	−0.168*	−0.153*
	(0.126)	(0.130)	(0.128)	(0.088)	(0.091)	(0.089)
Relative age	0.151***	0.151***	0.151***	0.171***	0.172***	0.172***
	(0.033)	(0.033)	(0.033)	(0.029)	(0.029)	(0.029)
Girl	0.356***	0.356***	0.356***	0.530***	0.530***	0.530***
	(0.022)	(0.022)	(0.022)	(0.018)	(0.018)	(0.018)
M – high school	0.203***	0.203***	0.203***	0.198***	0.198***	0.198***
	(0.030)	(0.030)	(0.030)	(0.026)	(0.026)	(0.026)
M – college	0.621***	0.621***	0.621***	0.641***	0.642***	0.642***
	(0.036)	(0.036)	(0.036)	(0.030)	(0.030)	(0.030)
F – high school	0.162***	0.163***	0.163***	0.126***	0.126***	0.126***
	(0.024)	(0.024)	(0.024)	(0.022)	(0.022)	(0.022)
F – college	0.455***	0.455***	0.455***	0.459***	0.459***	0.459***
	(0.033)	(0.033)	(0.033)	(0.027)	(0.027)	(0.027)
Parents cohabit	0.229***	0.229***	0.229***	0.286***	0.286***	0.286***
	(0.022)	(0.022)	(0.022)	(0.017)	(0.017)	(0.017)
Instrument class size						
R^2	0.139		✓	0.239		✓
N (Schools)		9,821 (412)			10,161(414)	

Notes. Standard errors are heteroscedasticity robust and corrected for school-level clustering. */ **/ ***, denote statistically significant at the 10%, 5% and 1% level, respectively. The college and high school dummies refer to (M)other's and (F)ather's education. All regressions include a constant term and a cohort dummy.

reports the corresponding first-stage results. Although compliance is not perfect, the instruments are strong predictors of number of grades in the classroom, with the coefficients on 3 predicted grades larger than 2 predicted grades. When we test the joint significance of our instruments, the predicted grade level dummies, we obtain an F-statistic equal to 446.[14] The individual characteristics are jointly insignificant, which means that after controlling for school size and the instruments there is no correlation with grade mixing and classrooms' socio-economic student composition.

The remaining columns in Table 6 present the estimates after instrumenting number of grade levels in the classroom using 2SLS for various class size specifications. The second column reports a statistically significant 2SLS estimate of 0.091 of the number of grades in a classroom on examination achievement while controlling for class size, assuming it is exogenous. This is somewhat higher than the comparable OLS

[14] To check that weak instruments are not an issue we also estimated our models using LIML which is unbiased in over-identified models. The effects we obtain are always nearly identical to the 2SLS estimates reported in the text.

estimate in the first column. The estimate barely changes when we also instrument class size in the third column. From Table 5, which reports first-stage for class size, we can see that class size is well identified and mostly loads on predicted class size with an F statistic of 188. The final two columns report the overall grade mixing effects for the teacher score. Again, we see an increase in the estimated effect to 0.07 once we instrument grade mixing. Also instrumenting class size in the final column does not change this.

The estimated effects of class size are very small, 0.001 and −0.001, insignificant yet precisely estimated. The results in Table 6 show that not only are there no confounding effects of class size on the number of grade levels in class but also confirm 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% of a standard deviation higher than the youngest in the cohort born in December. Girls also score significantly higher than boys and 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 examination scores. There is a small negative relationship between school size and teacher scores but dropping school size from our regressions does not change any of the results.

We, therefore, find that pupils in mixed grade classrooms outperform pupils in single grade classrooms. This might be surprising, in the sense that the heterogeneity of the classroom increases when combining grades. The results of Duflo *et al.* (2011) for Kenya, for example, suggest that this should have deteriorated pupils' achievement. To gain more insight into what is driving this result, Table 7 reports estimation results using (3). The first two columns of the Table present estimates for examination scores and the last two columns present 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.

Table 7

The Effect on Pupil Achievement of Being Mixed with Higher/Lower Grades

	Examination score		Teacher score	
	OLS	OLS	2SLS	2SLS
No. of lower grades	−0.105	−0.194	−0.050	−0.238**
	(0.075)	(0.136)	(0.052)	(0.095)
No. of higher grades	0.265***	0.388***	0.150***	0.399***
	(0.076)	(0.143)	(0.056)	(0.107)
Class size	0.002	0.000	0.000	−0.001
	(0.004)	(0.005)	(0.003)	(0.004)
N (Schools)	9,821 (412)		10,161 (414)	

Notes. All regressions include a constant term, cohort dummy 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% level.

In the first OLS specification the point estimate of the effect of exposure to the number of lower grades (g^- in (3)) on examination scores is −0.105. This suggests that sharing the classroom with a lower grade is detrimental for the examination 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 examination. 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 −0.194 which is still insignificant. The point estimate for the number of higher grades in the class room is 0.388 and remains significant at the 1% level. The final two columns of Table 7 adds estimates for the teacher set and graded test scores. These results confirm the conclusion based on the examination 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 examination scores. This is what we expected because the teacher scores are based on multiple evaluations and are, therefore, probably less noisy than the examination scores. Contrary to the examination 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 examination score, suggesting that the relative grading component in teacher grades is minor. Recall from Table 4 that if pupils are mixed, then they typically spend time with both lower and higher grades. This explains the small positive effects in Table 6: grade mixing is on average beneficial because pupils benefit more from being with higher grades than they lose from being with lower ones.

To summarise, 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 economise 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 (2012) 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 grade.

4.2. *Gender Differences*

Girls and boys experience different school outcomes. For example, not only do girls outperform boys in reading in all countries in the PISA study but in most countries this

gap is increasing (OECD, 2010). Boys score higher in mathematics but there is no gender gap in science performance. In nearly all OECD countries, upper secondary graduation rates for young women exceed those for young men (OECD, 2012). This good performance of girls relative to boys in compulsory schooling is also reflected by increased enrolment rates of women in colleges with in some countries the gender gap even reversing in favour of women (Goldin *et al.*, 2006).

Although gender differences in educational performance are well documented, much less is known about what underlies these differences. There is a growing literature that documents systematic gender differences in risk preferences, social preferences, and competitive preferences (Croson and Gneezy, 2009). On average, women are found to be more risk averse, socially malleable and averse to competition. There is also a large psychological literature that documents gender differences in social behaviour (Eagly and Wood, 1991). This suggests that peer groups may influence boys and girls differently.

There are studies that have compared peer effects for boys and girls in other contexts. Lavy *et al.* (2012*a*) find that in English secondary schools girls benefit more from better peers than boys. Using data from Trinidad, Jackson (2010) also finds that the benefits of attending schools with better performing pupils are larger for girls than for boys. Duflo *et al.* (2011) also find larger effects of tracking on mathematics performance for girls than for boys in Kenya. Lavy and Schlosser (2011) on the other hand find that the classroom's gender composition affects boys and girls similarly in Israel, while Black *et al.* (2013) find positive effects for girls and negative effects for boys in Norwegian lower secondary schools.

To investigate heterogeneity in the effect of a classroom' grade composition, we therefore perform our estimations separately by gender. The first two columns of Table 8 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' examination 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.

Table 8
Gender Differences, 2SLS Estimates

	Examination score		Teacher score	
	Girls	Boys	Girls	Boys
No. of lower grades	−0.413**	−0.054	−0.291*	−0.220
	(0.163)	(0.206)	(0.150)	(0.147)
No. of higher grades	0.565***	0.298	0.361**	0.478***
	(0.181)	(0.209)	(0.163)	(0.158)
Class size	0.008	−0.008	0.006	−0.008
	(0.008)	(0.007)	(0.006)	(0.006)
N (Schools)	4,744 (408)	5,077 (402)	4,883 (410)	5,278 (405)

Notes. 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. */ **/ and *** denote statistically significant at the 10%, 5% and 1% level, respectively.

When we test for equality of the effects across gender we can, however, not reject the null hypothesis that they are equal ($p = 0.393$).

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 ($p = 0.625$).

One interesting aspect of the results for teacher test scores is their size relative to those for the centralised examinations. 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 examination scores and the teacher scores for both sexes. Interpretation in terms of relative grading should be 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 younger students in the cohort. Because these interactions are imprecisely estimated and too inconclusive we do not report them.

5. 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 than in 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 gives 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. The second, and main, contribution of our article concerns combination classes which, as we documented in the introduction, are an important mode of classroom organisation 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 channelled by the presence of higher grades.

Appendix A

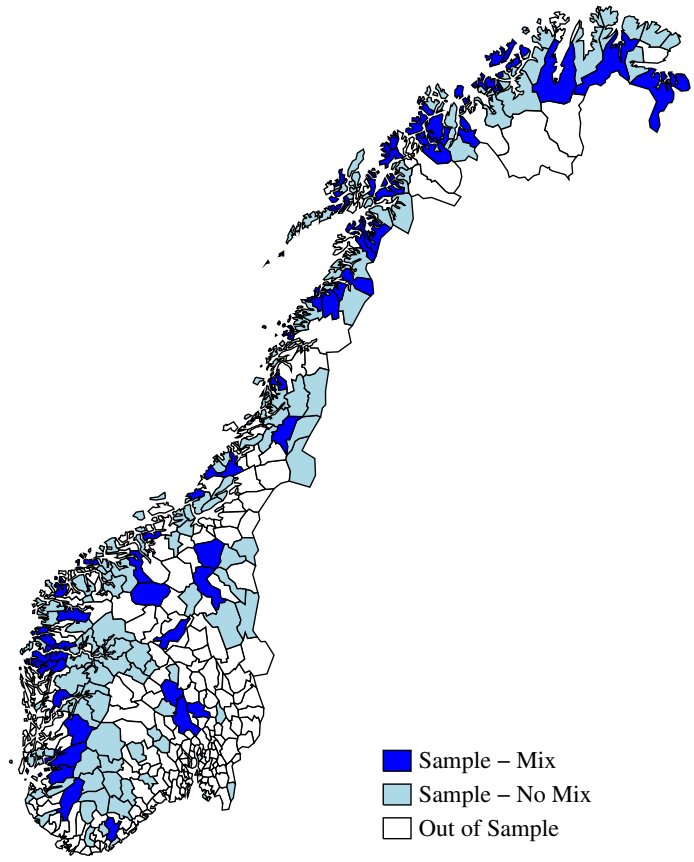


Fig. A1. *Regional Coverage*

Table A1

Reduced Form Regressions

	Examination score	Teacher score
Predicted no. of grades		
Three in 7th grade	0.204**	0.203***
	(0.091)	(0.063)
Two in 7th grade	0.010	0.053
	(0.052)	(0.035)
Three in 8th grade	0.048	0.039
	(0.105)	(0.072)
Two in 8th grade	0.047	−0.007
	(0.057)	(0.043)
Three in 9th grade	−0.089	−0.108*
	(0.090)	(0.062)

Table A1
(*Continued*)

	Examination score	Teacher score
Two in 9th grade	0.096*	−0.046
	(0.051)	(0.036)
Predicted class size	0.003	−0.002
	(0.004)	(0.003)
Relative age	0.150***	0.172***
	(0.033)	(0.029)
Girl	0.356***	0.531***
	(0.022)	(0.018)
M – high school	0.200***	0.199***
	(0.030)	(0.026)
M – college	0.618***	0.642***
	(0.036)	(0.030)
F – high school	0.162***	0.127***
	(0.024)	(0.022)
F – college	0.455***	0.460***
	(0.033)	(0.027)
Parents cohabit	0.232***	0.287***
	(0.022)	(0.017)
School size/100	−0.025	−0.161*
	(0.136)	(0.093)

Notes. Standard errors are heteroscedasticity robust and corrected for school-level clustering. */ **/ and *** denote statistically significant at the 10%, 5%, and 1% level, respectively. The college and high school dummies refer to (M)other's and (F)ather's education. All regressions include a constant term.

University of Oslo
Statistics Norway (SSB)

Accepted: 20 May 2014

Additional Supporting Information may be found in the online version of this article:

Data S1.

References

Ammermueller, A. and Pischke, J.S. (2009). 'Peer effects in European primary schools: evidence from the progress in international reading literacy study', *Journal of Labor Economics*, vol. 27(3), pp. 315–48.
Angrist, J.D. and Lang, K. (2004). 'Does school integration generate peer effects? Evidence from Boston's Metco program', *American Economic Review*, vol. 94(5), pp. 1613–34.
Angrist, J.D. and Lavy, V. (1999). 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement', *Quarterly journal of economics*, vol. 114(2), pp. 533–75.
Bedard, K. and Dhuey, E. (2006). 'The persistence of early childhood maturity: international evidence of long-run age effects', *Quarterly Journal of Economics*, vol. 121(4), pp. 1437–72.
Black, S.E., Devereux, P.J. and Salvanes, K.G. (2013). 'Under pressure? The effect of peers on outcomes of young adults', *Journal of Labor Economics*, vol. 31(1), pp. 119–53.
Boozer, M.A. and Cacciola, S.E. (2001). 'Inside the 'black box' of project STAR: estimation of peer effects using experimental data', Economic Growth, Yale University, Center Discussion Paper No. 832.
Cahan, S. and Cohen, N. (1989). 'Age versus schooling effects on intelligence development', *Child Development*, vol. 60(5), pp. 1239–49.
Croson, R. and Gneezy, U. (2009). 'Gender differences in preferences', *Journal of Economic Literature*, vol. 47(2), pp. 448–74.

Duflo, E., Dupas, P. and Kremer, M. (2011). 'Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya', *American Economic Review*, vol. 101(5), pp. 1739–74.
Eagly, A. and Wood, W. (1991). 'Explaining sex differences in social behavior: a metaanalytic perspective', *Personality and Social Psychology Bulletin*, vol. 17(3), pp. 306–15.
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*, vol. 29(3), pp. 589–607.
Fredriksson, P. and Öckert, B. (2013). 'Life-cycle effects of age at school start', Economic Journal, vol. 124 (571), pp. 977–1004.
Goldin, C., Katz, L.F. and Kuziemko, I. (2006). 'The homecoming of American college women: the reversal of the college gender gap', *Journal of Economic Perspectives*, vol. 20(4), pp. 133–56.
Hægeland, T., Raaum, O. and Salvanes, K.G. (2012). 'Pennies from heaven? Using exogeneous tax variation to identify effects of school resources on pupil achievements', *Economics of Education Review*, vol. 31(5), pp. 601–14.
Heckman, J.J. and Smith, J.A. (1995). 'Assessing the case for social experiments', *Journal of Economic Perspectives*, vol. 9(2), pp. 85–110.
Hoxby, C.M. (2000). 'Peer effects in the classroom: learning from gender and race variation', Working Paper 7867, NBER.
Jackson, C. (2010). 'Do students benefit from attending better schools? Evidence from rule-based student assignments in Trinidad and Tobago', Economic Journal, vol. 120(549), pp. 1399–429.
Krueger, A. (1999). 'Experimental estimates of education production functions', *Quarterly Journal of Economics*, vol. 114(2), pp. 497–532.
Lavy, V., Olmo Silva, O. and Weinhardt, F. (2012*a*). 'The good, the bad, and the average: evidence on ability peer effects in schools', *Journal of Labor Economics*, vol. 30(2), pp. 367–414.
Lavy, V., Paserman, M.D. and Schlosser, A. (2012*b*). 'Inside the black box of ability peer effects: evidence from variation in low achievers in the classroom', Economic Journal, vol. 122(559), pp. 208–37.
Lavy, V. and Schlosser, A. (2011). 'Mechanisms and impacts of gender peer effects at school', *American Economic Journal: Applied Economics*, vol. 3(2), pp. 1–33.
Leuven, E., Lindahl, M., Oosterbeek, H. and Webbink, H.D. (2010). 'Expanding schooling opportunities for 4-year-olds', *Economics of Education Review*, vol. 29(3), pp. 319–28.
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*, vol. 110(4), pp. 663–93.
Little, A.W. (2004). 'Learning and teaching in multigrade settings', Background paper for *UNESCO EFA Global Monitoring Report 2005*.
Manski, C.F. (1993). 'Identification of endogenous social effects: the reflection problem', *Review of Economic Studies*, vol. 60(3), pp. 531–42.
Mariano, L.T. and Kirby, S.N. (2009). 'Achievement of students in multigrade classrooms: evidence from the Los Angeles unified school district', *RAND* Working Paper WR-685-IES.
Mason, D.A. and Burns, R.B. (1997). 'Reassessing the effects of combination classes', *Educational Research and Evaluation*, vol. 3(1), pp. 1–53.
McCrary, J. (2008). 'Manipulation of the running variable in the regression discontinuity design: a density test', *Journal of Econometrics*, vol. 142(2), pp. 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*, vol. 20(17), pp. 1–14.
OECD. (2010). *PISA 2009 Results: Learning Trends*, Paris: OECD.
OECD. (2012). *Education At A Glance 2012*, Paris: OECD.
Rothstein, J. (2010). 'Teacher quality in educational production: tracking, decay, and student achievement', *Quarterly Journal of Economics*, vol. 125(1), pp. 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*, vol. 27(3), pp. 457–78.
Sojourner, A. (2013). 'Inference on peer effects with missing peer data: evidence from project STAR', Economic Journal, vol. 123(569), pp. 574–605.
Strøm, B. (2004). 'Student achievement and birthday effects', Unpublished Manuscript, Norwegian University of Science and Technology.
Thomas, J.L. (2012). 'Combination classes and educational achievement', *Economics of Education Review*, vol. 31(6), pp. 1058–66.
Urquiola, M. and Verhoogen, E. (2009). 'Class-size caps, sorting, and the regressiondiscontinuity design', *American Economic Review*, vol. 99(1), pp. 179–215.
Veenman, S. (1995). 'Cognitive and noncognitive effects of multigrade and multi-age classes: a best-evidence synthesis', *Review of Educational Research*, vol. 65(4), pp. 319–81.