OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 0305-9049
doi: 10.1111/obes.12561

Early Years Multi-grade Classes and Pupil Attainment*

DANIEL BORBELY,† MARKUS GEHRSITZ,‡,§ STUART MCINTYRE,‡ GENNARO ROSSI¶ and GRAEME ROY#

†*School of Business, University of Dundee, Dundee, UK (e-mail: DBorbely001@dundee.ac.uk)*
‡*Fraser of Allander Institute, Department of Economics, University of Strathclyde, Glasgow, UK (e-mail: markus.gehrsitz@strath.ac.uk)*
§*Institute for the Study of Labor (IZA), Bonn, Germany*
¶*Department of Economics, University of Sheffield, Sheffield, UK (e-mail: gennaro.rossi@ed.ac.uk)*
#*College of Social Sciences, University of Glasgow, Glasgow, UK (e-mail: Graeme.Roy@glasgow.ac.uk)*

Abstract

We study the effect of exposure to older, more experienced, classroom peers resulting from the widespread use of multi-grade classes in Scottish primary schools. For identification, we exploit that a class-planning algorithm quasi-randomly assigns groups of pupils to multi-grade classes. We find that school-starters benefit from exposure to second-graders in measures of numeracy and literacy. We do not find any evidence that these gains are driven by smaller class sizes or more parental input. While short-lived, these benefits accrue independent of socioeconomic background, to boys and girls alike, and our results provide no evidence that they come at the expense of older peers from the preceding cohort.

I. Introduction

Classroom composition and peer effects have been shown to be important determinants of pupil achievement. Several studies have documented the benefits of classroom exposure

JEL Classification numbers: C36, H52, I21, I26, I28, J24.

*We are grateful to Emma Congreve, Susan Ellis, Gordon McKinlay, Ian Walker and Tanya Wilson as well as to Antonio Acconcia, Marco Alfano, Maria De Paola, Eric Hanushek, David A. Jaeger, Roberto Nisticó, Jonathan Norris, Jens Ruhose, Elia Sartori and Simon Wiederhold for their helpful comments. The paper has also benefited from feedback from the Association for Education Finance and Policy's (AEFP), the European Society for Population Economics' (ESPE), the Royal Economic Society's (RES), the Scottish Economic Society's (SES), and the Society of Labor Economists' (SOLE) 2021 annual conferences, as well as comments at the Centre for Studies in Economics and Finance (CSEF) seminar series. We thank Mick Wilson and his team at Scottish Government for providing the raw data used in this study. We also thank Julian Augley, Fiona James, Suhail Iqbal, David Stobie, Amy Tilbrook and Dionysis Vragkos from the Scottish Centre for Administrative Data Research for their assistance in accessing the data used in this study. This project was supported by the Nuffield Foundation through grant EDO/43743.

to high-ability peers (Hanushek *et al.*, 2003; Lefgren, 2004; Ding and Lehrer, 2007; Neidell and Waldfogel, 2010; Lavy, Paserman, and Schlosser, 2012a; Lavy, Silva, and Weinhardt, 2012b), to female classmates (Hoxby, 2000; Lavy and Schlosser, 2011; Black, Devereux, and Salvanes, 2013; Anelli and Peri, 2019) and to classmates with college-educated mothers (Bifulco, Fletcher, and Ross, 2011; Bifulco *et al.*, 2014) as well as the adverse effects of disruptive peers (Figlio, 2007; Aizer, 2008; Carrell and Hoekstra, 2010; Carrell and Hoekstra, 2012; Carrell, Hoekstra, and Kuka, 2018). The ethnic makeup of classrooms (Angrist and Lang, 2004; Hoxby and Weingarth, 2005; Hanushek, Kain, and Rivkin, 2009; Hanushek and Rivkin, 2009; Fruehwirth, 2013) and the effect of immigrant peers on natives (Gould, Lavy, and Daniele Paserman, 2009; Ballatore, Fort, and Ichino, 2018) have also received attention. However, little is known about a widespread classroom structure that explicitly creates and harnesses peer effects: multi-grade classes. These are classes comprised of pupils from adjacent grades. For instance, first-graders being taught alongside second-graders, and thus being exposed to older, more experienced peers.[1]

Multi-grade classes are widely used. About 28% of schools in the USA use a mixed class setup and more than a third of primary school pupils in France attend multi-grade classes (Leuven and Rønning, 2014). Yet, multi-grade classes have not been widely studied. A notable exception is Sims (2008) who documents that multi-grade classes were an unintended consequence of California's Class Size Reduction Program: to comply with the new policy and thus qualify for additional funding, schools simply pooled pupils from adjacent grades into multi-grade classes. He shows that this had a detrimental impact on the test scores of pupils in multi-grade classes. Recent studies of rural areas of Norway (Leuven and Rønning, 2014) and Italy (Checchi and De Paola, 2018; Barbetta, Sorrenti, and Turati, 2019) have built on this work. They exploit that in these rural settings cohorts are often so small that pooling several year-groups is done out of necessity. With the exception of Checchi and De Paola (2018), they find that pupils in these schools actually benefit from attending multi-grade classes.

In shaping policy, decision-makers need to know whether the benefits documented by this nascent literature translate outside of a rural context or whether – consistent with Sims (2008) – multi-grade groupings may even have a detrimental impact on pupil performance. In our study we are able to examine this issue directly because in Scotland, the subject of this study and a constituent nation of the United Kingdom, multi-grade classes feature in virtually all primary schools. In fact, they are consciously created in both rural and urban schools, which allows our study to investigate their impact on attainment in settings in which the majority of pupils are educated. As such, our study holds important lessons for both policymakers and education practitioners.

In order to identify the causal effect of multi-grade classes, we exploit that in Scottish primary schools, an algorithm ('class planner') determines the most cost-efficient number, size, and composition of classes, subject to nationwide minimum and maximum class size rules. Specifically there are class size limits for single-year classes which vary by grade, and separate caps for multi-grade classes. The class planner is set up to minimize the

[1] A related strand of both the education (Slavin, 1987) and economics literature (Betts, 2011) has explored the effects of ability grouping and academic tracking.

number of classrooms a school needs to create. Combined with fluctuations in enrolment counts across years, this generates variation in the composition of classes within and across schools.[2] In effect, small and random variations in enrolment counts trigger the creation of multi-grade classes in some grades, in some schools and in some years, but not in others.

Enrolment in Scottish primary schools is, in turn and on the whole, determined by random population variation. Every primary school has a catchment area and pupils within a school's catchment area are entitled to attend their catchment area school. Small changes in enrolment in any primary school grade can lead to a re-shuffling of pupils into multi-grade and non-multi-grade classes across all grades of the school. The ramifications of this reshuffling are particularly pronounced in first grade. This renders it all but impossible for parents or school administrators to manipulate the overall school enrolment count to either trigger or prevent the creation of a multi-grade class.

We exploit this natural experiment by instrumenting each pupil's class status (multi-grade or single-year-group) with the class planner's recommendation for whether the pupil's year-group should contribute pupils to a multi-grade class. Note that the class planner only makes a recommendation on how many pupils in a grade should be put into a multi-grade class, but not *which* pupils. We therefore identify a local average treatment effect (LATE). We document that the compliers tend to be older members of cohorts who form the lower-grade part of a multi-grade class. They typically share their multi-grade classroom with the youngest and low-attainment members of the preceding cohort who have an additional year of primary school experience.

We combine our instrumental variable approach with novel, individual-level administrative data collected from successive waves of the Scottish Pupil Census (SPC) from 2007–08 to 2018–19. We link these data with assessment information and observe the exact classroom type and composition in each school and year. However, the predictive power of the class planner is strongest in first grade, whereas analyses of later grades may at times suffer from 'weak instrument' issues (see Bound, Jaeger, and Baker, 1995 and Lee *et al.*, 2022). This article, therefore, focuses its conclusions on the attainment effects of exposure to older, more school-experienced peers in first grade.

We find that exposure to second-graders in the first year of primary school by way of a multi-grade class leads to large improvements in literacy and numeracy. In fact, gains created by multi-grade classes are roughly equivalent to the attainment gap between the average pupil and a pupil in one of the 20% most deprived data zones in Scotland.[3] Boys and pupils from deprived neighbourhoods appear to benefit more from sharing a classroom with more experienced peers, although neither gender nor socio-economic differences are significant in a statistical sense. We also find little in the way of an urban/rural differential. We find no evidence that the achievement gains for school-starters come at the expense of

[2]For instance, the maximum class size for fourth and fifth grade in Scotland is 33, while multi-grade classes are capped at 25. Therefore, for an enrolment count of 45 fourth-graders and 46 fifth-graders, the class planner would recommend the creation of one 33 pupil fourth and fifth-grade class each, and one 25 pupil multi-grade class. Yet with the addition of just one fourth-grade pupil (i.e., 46 pupils in both grades), class size maxima would force the creation of two fourth-grade and two fifth-grade classes.

[3]Data zones are small area statistical geographies constructed by the Scottish Government comprising areas of approximately equal population size.

learning progress of second-graders who shared a multi-grade classroom with first-graders. However, we also document that the benefits for first-graders are short-lived.

Ours is the first study to document the benefits of multi-grade classes in a setting where they are not a niche phenomenon but a staple of the education system. In Scotland, multi-grade classes are used by schools in more affluent and less affluent areas alike, as well as in urban and rural schools. As such, our study pushes a nascent literature on multi-grade groupings forward and adds to its external validity. We also contribute to a growing literature on early years learning, from which we know the disadvantages of early school start and low age rank (Bedard and Dhuey, 2006; Black, Devereux, and Salvanes, 2011; Crawford, Dearden, and Greaves, 2014; Cascio and Schanzenbach, 2016; Ballatore, Paccagnella, and Tonello, 2020). We find that multi-grade classes help the youngest pupils in these classes at least as far as attainment is concerned – this underlines a distinction between absolute and relative age.

Finally, we show that multi-grade classes save classrooms – and thus costs – while at the same time accruing net benefits in terms of pupil performance. Indeed, our results suggest that multi-grade classes are a viable way to better reconcile policymakers' goals of promoting higher-achieving pupils and pursuing value-for-money in education spending.

II. Data and background

Pupils in Scotland typically start school in August of the year in which they turn five. They attend primary school from first grade (P1) to seventh grade (P7) before transferring into secondary schools. Government-funded public schools are free for the approximately 700,000 pupils aged 5–19. There is only a small private school sector, accounting for about 4% of pupils, which is mostly clustered in the populous *Central Belt* of the country. The Scottish education system has always been separate from that of the rest of the UK, education is devolved to the Scottish Government. In contrast to England where parental school rankings are solicited and pupils then matched to schools with open slots, school choice in Scotland resembles the system that is in place in most of the USA. That is, school choice is largely contingent on non-overlapping catchment areas which are drawn up by local authorities (roughly equivalent to school districts), and rarely ever change. Each primary school has a catchment area and any pupil whose main residence is within this boundary is entitled to a place in that school. Parents may also ask for their children to attend a school other than their catchment area school via so-called placing requests. These are applications to the local council to transfer a child to a specified school. However, these requests are not automatically approved and, overall, only about 5% of pupils in our sample attend a school different from the one of their catchment area.[4] Therefore, sorting into catchment areas of schools that are perceived to be desirable is a strictly dominant strategy for parents. Rossi (2021), for instance, documents that housing prices on two sides of catchment border areas in Scotland differ on average by as much as 4%.

[4]Councils are under no obligation to grant these requests and will not do so if a school is at capacity. Places are allocated based on criteria decided by each Local Authority, typically children with additional support needs and/or with siblings in the specified schools get priority.

TABLE 1
Maximum class size rules

Grade	*Max. size*
Primary 1 (P1)	25
Primary 2 and 3 (P2, P3)	30
Primary 4 to 7 (P4–P7)	33
Composites (all grades)	25

Notes: Maximum class size rules in Scotland as of 2019. P1 cutoff was 30 prior to 2011.

The Scottish Government centrally sets maximum class size rules in primary school which apply to the entire nation: class size in P1 must not exceed 25 pupils, the maximum for P2 and P3 is 30, and classes in P4–P7 are formed as multiples of, at most, 33 (see Table 1). A widespread feature of Scottish primary education are multi-grade classes, known as 'composite classes' in Scotland (we use the two terms interchangeably throughout this paper). These are classes comprised of pupils from adjacent grades. The maximum class size for multi-grade classes is 25 and each grade needs to contribute a minimum of five pupils.

Figure 1 provides an illustration of the distribution of pupils across single-year and multi-grade classes in 2018. Composite classes typically stretch across two grades and more than one in six Scottish primary school pupils attend a multi-grade class. In contrast to most of the examples in the literature to date, multi-grade classes are by no means a rural phenomenon in Scotland. For example, in 2018, 84% of primary schools in the City of Glasgow – the fourth largest city in the UK – featured at least one composite class.

Our data are drawn from the Scottish Pupil Census (SPC) for school years 2007–08 to 2018–19. The SPC takes place every year in September and collects information on every individual pupil and the schools they attend. Upon entering the Scottish school system, every pupil is assigned a unique ID, the so-called Scottish Candidate Number (SCN). We use the SCN to link pupils' records across years and to assessment data. Since 2015–16, every pupil's progress is assessed in both numeracy and literacy as either 'Below Early Level', 'Early Level', and at '1st/2nd/3rd/4th' level. These assessments are teacher based but informed by standardized test scores to ensure consistency.

Not least because test scores from one-off standardized tests, in contrast to assessments by professional teachers, lack the ability to perform a summative judgement of pupils' abilities, teacher assessments might be considered a preferable measure of attainment. Nonetheless, one might be concerned about teachers being subjective or even biased towards pupils in multi-grade classes. Figure 2 suggests that this is not the case. It shows that fourth graders' standardized reading test scores, which we obtained for a subsample from the Scottish Survey of Literacy and Numeracy (SSLN), are highly correlated with their teacher assessments.[5] Crucially, there is no evidence that teachers compensate students who form the bottom part of composite classes with better marks relative to those in non-composite classes. We are thus confident that our outcome measure accurately captures learning progress.

[5]The SSLN last took place in 2015–16, the first year for which we have teacher assessments. It was only administered to a subset of fourth and seventh grader and only reading ability was tested. We can, therefore, not include figures similar to Figure 2 for first graders or numeracy outcomes.

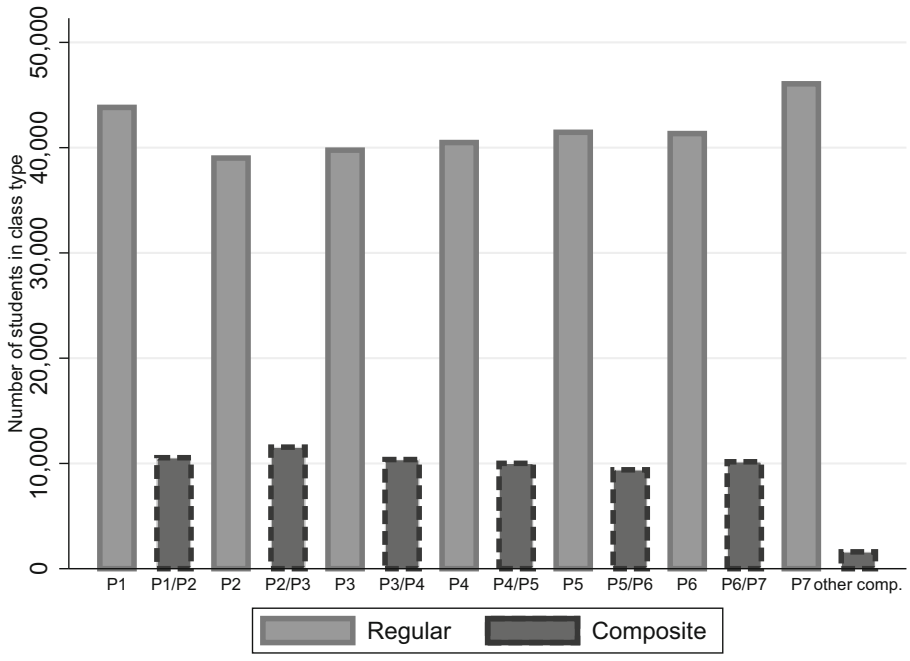

Figure 1. Pupils by grade and class type (2018).
Notes: This bar chart shows the distribution by class type (single-year vs multi-grade) of pupils in Scottish primary schools in 2018

Assessments are made at the end of P1 when pupils are expected to perform at early level, and at the end of P4 and P7 when students are expected to perform at the first and second level, respectively. We use the SCN to link each pupil to their assessments and create indicators for whether a pupil performs at least at the expected level in a given stage.

The SPC also documents the school and name of the class that each pupil attends as well as each pupil's grade or cohort. Since ours is individual level data, we can easily identify multi-grade classes and calculate class sizes which we cross-checked with official aggregates published by the Scottish Government. Table 2 presents summary statistics for about 190,000 first-graders who between 2015–16 and 2018–19 attended one of the 1,437 primary schools in our sample. Eighty-five and seventy-six percent of first-graders perform at level in numeracy and literacy respectively. The average class size is 21.8, about half the sample is female and the average school starting age is 5.2 years. We use the so-called Scottish Index of Multiple Deprivations (SIMD) as a proxy for socio-economic background. The SIMD ranks 6,976 'data zones' from most to least deprived in terms of income, employment, education, health, access to services, crime and housing. Unsurprisingly, about 20% of pupils come from households located in areas ranking in the bottom quintile.[6]

[6]Our sample also differs marginally from the original population data. We excluded about 1% of pupils who are either in special education classes, receive a Gaelic Medium education, or are in classes in which non-English speakers (e.g. refugees) were grouped together regardless of age/grade.

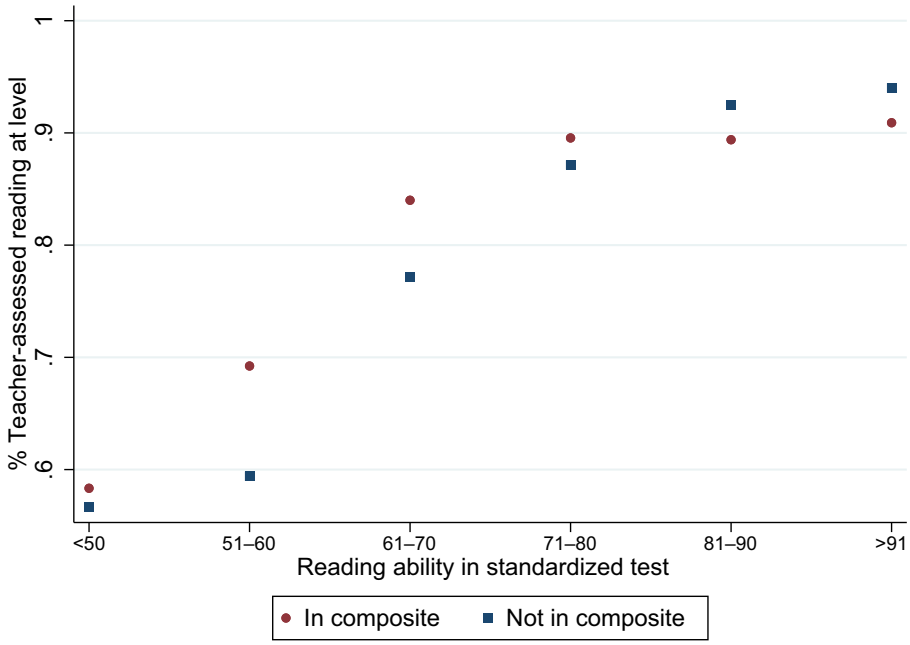

Figure 2. Comparison of standardized test-scores and teacher assessments.
Notes: This chart compares teacher assessment in reading with standardized reading test scores from the Scottish Survey of Literacy and Numeracy (SSLN) for 1,921 P4 pupils for which both assessment types were available. Pupils are grouped into six bins according to their performance on a standardized reading test during the school year. The position of the dots (pupils in P4/P5 composite classes) and squares (pupils who are not in P4/P5 classes) then indicates the percentage of each bin that were assessed by their teachers at the end of the school year to perform at least at the expected level in reading [Colour figure can be viewed at *wileyonlinelibrary.com*]

III. Empirical design

Our aim is to compare attainment between pupils who attend multi-grade classes and those in single-year classes, ceteris paribus. We model attainment of pupil i in classroom c and grade g of school s in year t as a function of class type, observable student and school socio-economic characteristics as well as unobservable attributes. The following equation describes this education production function in its simplest form:

$$A_{icgst} = \beta_0 + \beta_1 \text{Comp}_{cgst} + \gamma X_{igst} + \delta_s + \tau_t + \varepsilon_{icgst}, \quad (1)$$

where A_{icgst} is achievement, in particular student competency in numeracy and/or literacy; Comp_{cgst} is either a dummy that is equal to one for a multi-grade class and zero for a single-grade class, or a continuous variable equal to the number of older (younger) peers from preceding (succeeding) cohorts; X_{igst} is a vector of observed student characteristics such as age, gender, ethnicity, and socio-economic background, an indicator for whether a pupil attends a school outside their catchment area, school-level fractions of the same characteristics, as well as a control for grade enrolment and class-size. δ_s and τ_t are sets of school fixed effects and year fixed-effects, respectively. Finally, ε_{icgst} is any other determinant of achievement.

TABLE 2
Summary statistics

	First-graders (P1)		*Fourth-graders (P4)*		*Seventh-graders (P7)*	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Numeracy – performing at level	0.851	0.356	0.759	0.428	0.731	0.444
Literacy – performing at level	0.759	0.428	0.690	0.463	0.679	0.467
Reading – performing at level	0.819	0.385	0.777	0.416	0.775	0.417
Writing – performing at level	0.791	0.406	0.721	0.449	0.708	0.455
Listening and talking at level	0.871	0.335	0.844	0.363	0.829	0.377
Class size	21.813	3.265	26.635	3.955	26.413	4.323
Grade enrolment	46.168	19.381	46.650	18.788	44.333	17.801
Female	0.491	0.500	0.493	0.500	0.491	0.500
White	0.828	0.377	0.855	0.352	0.878	0.327
Native English speaker	0.926	0.262	0.924	0.265	0.937	0.243
Bottom 20% SIMD	0.226	0.418	0.217	0.412	0.216	0.411
Age (in years)	5.210	0.307	8.205	0.308	11.209	0.313
From outside catchment area	0.048	0.212	0.058	0.234	0.068	0.252
% Female in school	0.490	0.032	0.490	0.032	0.490	0.032
% White British	0.848	0.123	0.852	0.116	0.853	0.119
% Native English speakers	0.922	0.098	0.925	0.092	0.925	0.095
% in bottom 20% SIMD	0.223	0.265	0.217	0.262	0.217	0.261
% Placing request	0.048	0.213	0.058	0.234	0.068	0.252
Number of Students in School	317.454	126.694	319.516	127.876	317.994	128.655
Observations	190,704		194,804		186,082	
Number of schools	1,437		1,428		1,435	

Notes: All data stem from Scottish Pupil Census (SPC) 2015–16 to 2018–19, with assessment data added by matching via Scottish Candidate Number (SCN).

Our main empirical concern is the potential correlation between Comp_{cgst} and ε_{icgst}. Pupils who are placed in multi-grade classes are not randomly selected. In fact, both unobservable and observable pupil characteristics determine multi-grade status. For instance, head teachers might be inclined to select high ability students as the bottom part of a multi-grade class who are then pooled with low attainment pupils from the stage above. They are also encouraged to take social bonds into account, so as to keep groups of friends together. Maturity and age are also important considerations. Table 3 shows that older first graders are more likely to be placed in a P1/P2 multi-grade class whereas the opposite is true for second graders.

While age and other demographic characteristics are observable, ability and social networks are not. Since ε_{icgst} and Comp_{cgst} are likely to be correlated even after accounting for X_{igst}, school and year fixed effects, estimating equation (1) by OLS will not give us a consistent estimate of β_1, since not all relevant characteristics would have been held constant.

In order to be able to identify a causal effect of Comp_{cgst} on A_{icgst}, we use potentially exogenous variation in Comp_{cgst} created by a class planning algorithm. Local authorities use this tool to calculate the cost-minimizing number and type of classes, using a school's enrolment counts for each grade as inputs. In particular, the class planner takes into account that multi-grade classes can be used as means of reducing the number of classes that a school needs to create, considering maximum class-size rules and ensuring that

TABLE 3
Self-Selection of composite class pupils

	Prob(CompP1/P2) – First graders			*Prob(CompP1/P2) – Second graders*			
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
Female	0.004***	0.004***	0.003***	−0.004**	−0.004***	−0.003*	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
White	0.006**	−0.004*	−0.004	0.010***	−0.002	−0.002	0.000
	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)
Native English speaker	0.015***	0.016***	0.015***	−0.010**	−0.009***	−0.008**	−0.002
	(0.004)	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.004)
Bottom 20% SIMD	−0.001	−0.003	−0.003	0.004	0.006***	0.006***	0.002
	(0.004)	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.003)
Outside catchment area	0.014***	0.017***	0.018***	-0.004	0.003	0.003	0.000
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)
Age (in years)	0.132***	0.135***		−0.103***	−0.105***		−0.104***
	(0.005)	(0.005)		(0.005)	(0.005)		(0.006)
First age quartile			−0.013***			0.051***	
			(0.002)			(0.003)	
Third age quartile			0.027***			−0.020***	
			(0.002)			(0.002)	
Fourth age quartile			0.098***			−0.028***	
			(0.004)			(0.003)	
Low literacy							0.029***
							(0.004)
Low numeracy							0.036***
							(0.005)
Observations	190,704	190,704	190,704	203,139	203,139	203,139	139,198
R-squared	0.018	0.179	0.181	0.010	0.162	0.163	0.175
School FE	No	Yes	Yes	No	Yes	Yes	Yes

Notes: Heteroscedasticity-robust SEs adjusted for clustering at the school-by-year level are reported in parentheses. This table regresses a dummy indicator for whether a pupil is part of a P1/P2 composite class on pupil characteristics. The first three columns show the results for first-graders who form the bottom component of a P1/P2 composite class. Columns (4) through (7) show our results for second graders who form the top component of a P1/P2 composite class. Note that only P1 pupils from our main sample (with valid assessment data) are used. In column (7) only P2 pupils for whom P1 assessments (from previous year) were available, are part of the sample. Low literacy and low numeracy, respectively, indicate that P2 pupils scored below early level when in first grade.
***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

each grade contributes at least five pupils to a multi-grade class (if it is optimal to create one).

To illustrate our source of identifying variation, Figure 3a shows the optimal allocation – as predicted by the class planner – for one of the schools in our sample. Enrolment counts for all seven grades are in the high 40s or low 50s, as is typical in the average school. For illustrative purposes, we zoom in on the bottom three grades. The class planner here determines that the optimal allocation is to create two single-year classes for each grade. Figure 3b, on the other hand, shows the optimal allocation, as calculated by the class planner, for a case which is identical to the one in Figure 3a except that there are now 44 instead of 45 pupils enrolled in first grade. This marginal change triggers several multi-grade classes across different stages, and the suggested reallocation ultimately saves one classroom in a higher grade. This example illustrates that marginal

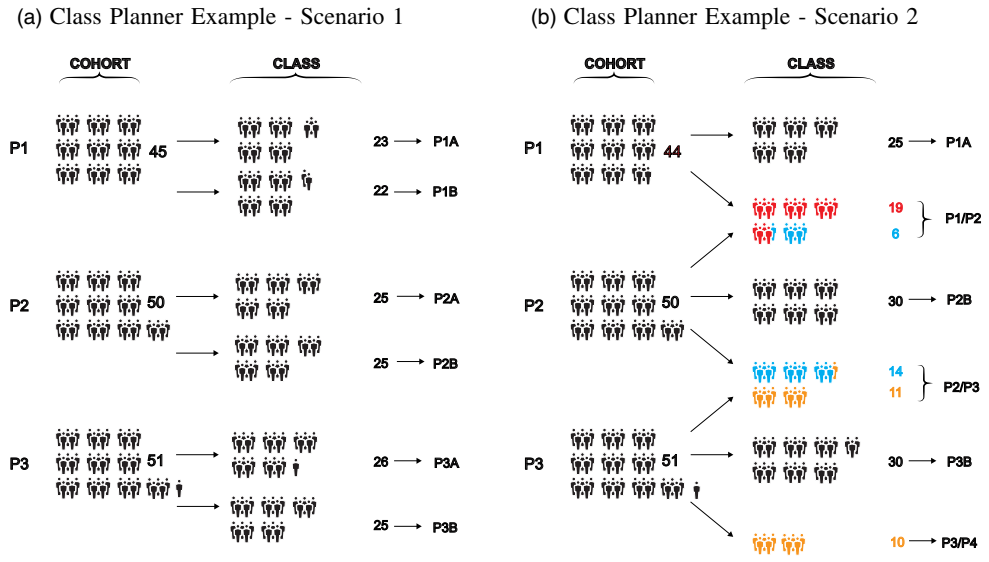

Figure 3. Class planner examples.
Notes: This is an illustration of the allocations suggested by the class planner. In reality, enrolment counts for all seven primary school grades are fed into the class planner, for ease of interpretation we focus here on the bottom three grades of an anonymized primary school. We show two scenarios. The only difference between both scenarios is that in scenario 1 (on the left) this school has an enrolment count of 45 first graders, whereas in scenario 2 (on the right), there are 44 first graders enrolled. As is apparent from the figure, this marginal difference leads to fundamentally different class planner predictions. In scenario 1, none of the pupils is assigned to a composite class (i.e. $\text{Comp}_{gst}^{\text{pred}} = 0$), in scenario 2 all grades are assigned to treatment. (a) Class planner example – scenario 1; (b) Class planner example – scenario 2 [Colour figure can be viewed at *wileyonlinelibrary.com*]

changes in enrolment counts in any grade may trigger multi-grade classes and reshuffle pupils into different class types across all grades. As a result, pupils are quasi-randomly exposed to peers from either the same or older/younger age groups. We use the predictions of the algorithm as an instrument for the class status of each pupil. In its simplest form, we instrument Comp_{cgst} with an indicator for whether the class planner suggests that grade g should contribute to a multi-grade class.

One key identifying assumption in our empirical setup is that of a strong first stage. Local authorities use the class planner tool to allocate teaching resources to schools based on enrollment counts. Head teachers are not obliged to exactly follow the class allocation suggested by the class planner. However, given that they only receive the resourcing commensurate to the number of classes predicted by the class planner, their ability to deviate from class planner suggestions is limited. We analytically assess compliance and thus the strength of our instrument by running a standard first-stage regression corresponding to the following equation:

$$\text{Comp}_{icgst} = \alpha_0 + \alpha_1 \text{Comp}_{gst}^{\text{pred}} + \pi X_{igst} + \theta_s + \mu_t + r_{icgst}, \tag{2}$$

where r_{icgst} is a regression residual. Comp_{icgst} is a dummy indicator for whether class c in grade g which contains pupil i, is a multi-grade class whereas $\text{Comp}_{gst}^{\text{pred}}$ is an indicator for

TABLE 4
First-stage results

	First graders (P1)		*Fourth graders (P4)*				*Seventh graders (P7)*	
	(1) Bottom Comp.	*(2) Older peers*	*(3) Bottom Comp.*	*(4) Top Comp.*	*(5) Younger peers*	*(6) Older peers*	*(7) Top Comp.*	*(8) Younger peers*
$CompLow^{pred}_{gst}$	0.087***	1.039***	0.011***	−0.006	0.006	0.231***		
	(0.005)	(0.044)	(0.004)	(0.005)	(0.039)	(0.040)		
$CompUp^{pred}_{gst}$			0.005	0.021***	0.265***	−0.013	0.018***	0.212***
			(0.005)	(0.005)	(0.041)	(0.041)	(0.005)	(0.050)
Observations	190,704	190,704	194,804	194,804	194,804	194,804	186,082	186,082
R-squared	0.192	0.151	0.243	0.246	0.139	0.136	0.290	0.217
Number of schools	1,437	1,437	1,428	1,428	1,428	1,428	1,435	1,435
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat	366.2	552.7	4.554	9.971	12.55	12.81	12.81	17.84

Notes: Heteroscedasticity-robust SEs adjusted for clustering at the school-by-year level are reported in parentheses. This table shows the results for our estimation of a first-stage equation (2) in which we regress our endogeneous measures of class composition on our instruments which indicate whether a grade should contribute to a composite class. Covariates include pupil age, sex and ethnicity an indicator for whether pupil is from a neighbourhood in bottom 20% of deprivation (SIMD), whether a pupil attends a school outside their catchment area, classize and grade enrolment counts (and its square), the size of the school, and the percentage of pupils in a school that are female, white British, native English speakers, and in the bottom 20% of deprivation respectively. All specifications contain a set of school fixed effects and a set of year fixed effects.
The reported F-statistic is HAC and was calculated using the method developed by Kleibergen and Paap (2006).
***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

whether, according to the class planner, grade g should contribute to a multi-grade class, thus exogenously boosting the probability that pupils in this grade end up in a multi-grade class.

Our analysis of first graders allows us to isolate the effects of exposure to more experienced P2 peers. In our main specification, we therefore redefine our treatment dummy variable, $Comp_{icgst}$, as a continuous variable that measures the number of peers from the preceding cohort of second graders, $P2Peers_{icgst}$, who share multi-grade classroom c with pupil i.

Instrumental variable regressions, while consistent, always yield biased estimates, even if all identifying assumptions are met. Bound *et al.* (1995) show that weak instruments may massively exacerbate this finite sample bias that is inherent to Two-Stage-Least-Squares (2SLS) instrumental variable estimation. A common indicator of instrument strength is the first-stage F-statistic which is typically assessed against a cut-off (Stock and Yogo, 2002).

Furthermore, recent work by Lee *et al.* (2022) suggests that valid inference requires an F-statistic larger than 104.7 when using the standard critical values to construct confidence intervals. Our estimation of equation (2) is shown in Table 4. It indicates a strong first stage for our sample of first-graders with F-statistics of 366 and 552, respectively. These are also displayed at the bottom of our second stage results Table 6 in section IV. All F-statistics are heteroscedasticity and autocorrelation consistent (HAC) and were obtained using the method developed by Kleibergen and Paap (2006). Given these large F-statistics, for our analysis of first graders we can use standard inference procedure that constructs 95% confidence intervals as $\hat{\beta} \pm 1.96 \times \hat{se}(\hat{\beta})$.

By contrast, our first-stage results for P4 and P7 (the only other two stages with outcome data) are substantially smaller. This suggests that finite sample bias may not be negligible. In addition, and because the F-statistics are well below any $F > 104.7$ threshold, conventional confidence intervals are too narrow. Given that any findings for fourth and seventh graders will suffer from both bias and inference issues, we therefore focus our analysis on first-graders. We nonetheless report our results for fourth and seventh-graders for completeness in Table A2 in Appendix S1. Following the guidance in Lee *et al.* (2022) and the example of von Hinke (2022), in this Table we also report confidence intervals that are adjusted for the weak first stage.

There are several reasons why we might expect that our instrument would be stronger for lower grades compared to later grades. The main driver is the way the class planner is set up. The most cost-efficient pupil allocation provided by the algorithm is not always a unique solution. The class planner is coded to work sequentially through enrolment counts in each grade from P1 to P7 in calculating class allocations. It is thus more likely to suggest composite classes in earlier grades. This is also consistent with head teacher preference who may find pooling 5- and 6-year old pupils into a single classroom more appealing than pooling 11- and 12-year olds. After all, the former is just a continuation of the nursery/kindergarten setup, whereas the latter is a more discrete classroom composition break in pupils' primary school trajectory. In addition, head teachers may have concerns about the suitability of particular groups of pupils to learn effectively in mixed grade classrooms, and thus decide to not stick with a class planner suggestion. These suitability assessments are, however, much harder to make for school starters whose abilities schools have little information on.

In addition to being relevant (i.e. a strong first stage), a valid instrument must meet three further assumptions: (1) it needs to be as good as randomly assigned (so called independence assumption); (2) it needs to satisfy the exclusion restriction; and (3) it needs to be monotonically related to the treatment of interest. We have shown above that our instrument meets the relevance assumption and will outline below why it is likely to also satisfy the other three assumptions.

The independence assumption is credibly met because class planner predictions are ultimately generated by random population variation. A useful check for this assumption is to compare the observable characteristics of pupils who are in cohorts which – by virtue of class planner predictions – were assigned to contribute to a composite class, with observables of students who were not assigned to a composite class. If assignment to treatment is indeed as good as random, then both groups should closely resemble one another. The balancing tests in Table 5 suggest that this is indeed the case. For example, 49.0% of pupils whose cohort is assigned to contribute to a multi-grade class are girls, 83.0% are white, and the average school starting age is 5.21 years. The corresponding averages for pupils whose cohort is *not* assigned to contribute to a multi-grade class are 49.1%, 82.5% and 5.21 years and thus all but identical. Due to our large sample size, simple t-tests in column (3) in some cases flag small differences as statistically significant, but these differences disappear once school fixed effects and year fixed effects are taken into account. Therefore, none of the coefficients in column (4) are statistically significant at the 5% level. In addition, if we run a linear regression of $\text{Comp}_{gst}^{\text{pred}}$ on the 10 control variables in Table 5 plus school and year fixed effects, the F-stat for the null hypothesis

TABLE 5
Balancing test

	Not assigned	*Assigned*	*Raw differences*	*Adj. differences*
White	0.825	0.830	0.005***	−0.004*
	[0.380]	[0.376]	(0.002)	(0.002)
Native English Speaker	0.921	0.928	0.007***	0.002
	[0.270]	[0.259]	(0.001)	(0.002)
Female	0.491	0.490	−0.001	−0.004
	[0.500]	[0.500]	(0.003)	(0.003)
Bottom 20% SIMD	0.228	0.224	−0.004*	0.001
	[0.420]	[0.417]	(0.002)	(0.002)
Age (in years)	5.210	5.210	−0.000	0.002
	[0.307]	[0.307]	(0.002)	(0.002)
% Female in school	0.490	0.490	−0.000**	−0.000
	[0.032]	[0.032]	(0.000)	(0.001)
% in bottom 20% SIMD	0.227	0.222	0.009***	−0.000
	[0.272]	[0.263]	(0.001)	(0.001)
% White British	0.842	0.851	0.008***	0.001
	[0.131]	[0.120]	(0.000)	(0.001)
% Native English speakers	0.916	0.924	−0.005***	−0.000
	[0.105]	[0.095]	(0.001)	(0.001)
Number of students in school	309.3	320.6	11.289***	1.076
	[124.8]	[127.3]	(0.646)	(0.711)
Observations	53,219	137,485	190,704	190,704

Notes: This table compares pupils whose cohort was assigned to contribute to a multi-grade class (i.e. $\text{Comp}_{gst}^{\text{pred}} = 1$) with those whose cohort was not assigned (i.e. $\text{Comp}_{gst}^{\text{pred}} = 0$). Column (1) and (2) show the means and SDs (in brackets), and column (3) computes the raw differences by regressing pupil characteristics on assignment status. In column (4), we also control for school fixed effects and year fixed-effects.

that the coefficients on the 10 controls variables are zero is 1.86. Hence, we cannot reject the null that none of the control variables predicts assignment to a multi-grade class.

The exclusion restriction requires planner predictions (i.e. our instrument) to only affect learning outcomes through class-type. While this assumption is not formally testable, planner predictions are in practice indeed only used to determine the number and types of classes. Moreover, random fluctuations in the enrolment counts for *any* grade may change planner predictions across all grades. It is, thus, not conceivable that head teachers or parents can manipulate enrolment counts in order to consciously trigger or prevent multi-grade classrooms in a specific grade. In addition, as composite class formation is impossible for parents to foresee, we confidently rule out sorting into schools on the basis of an anticipated placement in a composite class. Our instrument is, therefore, unlikely to be correlated with parent or school characteristics that have an independent effect on our outcome of interest.

Furthermore, we want to rule out an alternative channel through which class planner predictions might affect our outcome: class size. If multi-grade classes were deployed to save resources, one might worry that class sizes may be larger, on average, whenever there is an incentive to create a multi-grade class. While this might be plausible at the school level, it is not necessarily the case at the grade level. As we discuss in section IV, class size in first grade is the same as for composite classes as it is for single-year classes. Nevertheless, we will be controlling for class size in all specifications.

We will also re-run all analyses using Conley, Hansen, and Rossi's (2012) 'plausibly exogenous' instrumental variable approach.[7] The method estimates confidence intervals for our main effect of interest while relaxing the exclusion restriction. Figure A1 shows that our estimates are robust to mild violations of the exclusion restriction. More severe violations, on the other hand, lead to very wide confidence intervals. That is, our identification strategy relies on the exclusion restriction to be quite strictly met. However, given the discussion on potential violations of the exclusion restriction above, we believe this is plausible in our context.

Lastly, the monotonicity assumption requires that the very assignment of cohorts to contribute to a multi-grade class does not in fact lower the probability of attending a composite class for affected pupils. That is we assume that there are no 'defiers' (to use the terminology of Angrist, Imbens, and Rubin, 1996) who press to become part of a composite class precisely because their cohort was not assigned to contribute to a composite class. Defier behaviour seems unlikely in our research setting, indicating that the monotonicity assumption is also met.[8]

Hence, as long as our instrument is valid, we can consistently estimate a LATE. That is different from a population average treatment effect (ATE) for two reasons. First, head teachers may not always follow the suggestions of the class planner. Even though head teachers who do not stick with algorithmic suggestions face clear budgetary issues, we have outlined above that compliance, while strong, is not perfect. Second, while it is as good as randomly determined whether a *grade* contributes to a multi-grade class, the specific subset of *pupils* who, in turn, are assigned to such a multi-grade class is not a randomly selected sample.

The interpretation of our LATE hinges on who these 'compliers' are. Table 3, for instance, shows that age is a strong positive predictor of attending a multi-grade class. The oldest pupils of a cohort are more likely to become the lower-grade component of a multi-grade class whereas the youngest members of a cohort are more likely to become the higher-grade part of a multi-grade class. The coefficients in Table 3 lack causal interpretation, but this pattern is consistent with insights from school officials and teachers who we consulted as part of our research. Other socio-economic characteristics are only weak predictors. For instance, girls are a mere 0.4 pp more likely to attend a P1/P2 than boys. Hence, the compliers in our study tend to be comparatively mature school-starters, but do not otherwise differ substantially from fellow school-starters in terms of observable background characteristics.

While age has an independent effect on attainment (Black *et al.*, 2011), it is important to note that non-random selection of pupils who are taught in multi-grade classes does not induce bias into our estimated LATEs. Indeed, our instrumental variable technique addresses exactly this selection issue. Put differently, an instrumental variable approach requires that whether a cohort is assigned to contribute to a composite class is as good as random; but which members of such a cohort comply with the assignment does not have to random. Indeed, this selective compliance is what renders a LATE 'local'. Intuitively,

[7]Clarke and Matta (2018) recently developed the corresponding software package and the method has been used by, among others, Barban *et al.* (2021).
[8]Our setup allows, of course, for 'never-takers' who refuse the treatment when assigned to it.

our identification strategy compares pupils who – by virtue of random variations in enrolment counts – end up in a multi-grade class with older peers, against pupils who would have ended up in a multi-grade class, had the enrolment count in their school-year just marginally differed from their actual enrolment count. While our LATE might thus not yield a universal average treatment peer effect, it is arguably more policy-relevant than the ATE. After all, we identify peer effects for those school starters who are, in practice, most likely to be exposed to second-graders by way of multi-grade classes.

IV. Results

In this section we present our estimates for the effect of exposure to older, more (school) experienced peers by way of multi-grade classes. For comparison, we report OLS estimates alongside 2SLS coefficients corresponding to equation (1). All specifications control for individual pupil characteristics, time-variant school characteristics, school fixed effects and year fixed effects. SEs are adjusted for clustering at the school-by-year level throughout.

Second-stage results

The 'naïve' OLS estimates in column (1) of Table 6 indicate a marginally significant positive ''effect" on numeracy which is nonetheless very close to zero. This stands in contrast to the sizeable, positive and statistically significant 2SLS estimates in columns (2) and (3) of Panel A. These results show that for first-graders, exposure to an additional older peer raises the probability of performing at level or better in numeracy by 0.8 to 1.1 percentage points. On average, P1/P2 classes contain about 10 P2 pupils, so this translates into an average increase of 9–11 percentage points for pupils attending a typical composite class (see columns (5) and (6)). Panel B shows that our effects are slightly larger for literacy. Each P2 peer increases performance by 1.3 to 1.5 percentage points. The coefficients in both columns (8) and (9) are statistically significant at the 5% level. This translates into a 15–16 percentage point increase in the probability of performing at least at the expected level in literacy for pupils in a multi-grade class.[9]

While our 2SLS estimates are large, they are in line with the previous literature. For instance, Leuven and Rønning (2014) find that multi-grade classes in Norway increase younger pupils' performance by 0.4 SDs, and Barbetta *et al.* (2019) document educational gains for sharing a class with older peers of as much as 0.33 SDs. Our point estimates suggest improvements of 0.28 SDs for numeracy and 0.35 SDs for literacy. By way of comparison, these gains are large enough to close the attainment gap between the average pupil and a pupil in one of the 20% most deprived data zones in Scotland.

We also find no evidence that gains for first graders come at the expense of lower attainment among their second grade peers. The second stage results in columns (2) and (4) of Panel A in Table 7 indicates a small negative effect on maths assessments of second graders who shared a multi-grade classrooms with first-graders. However, the

[9]Columns (1) and (2) of Table A1 in Appendix S1 report the reduced form estimates. The ratio of our reduced-form and first-stage effects is approximately equal to our 2SLS coefficients.

TABLE 6
Second-stage results – first graders (P1)

Panel A: Numeracy – Performing at least at level						
	(1) *OLS*	*(2)* *2SLS*	*(3)* *2SLS*	*(4)* *OLS*	*(5)* *2SLS*	*(6)* *2SLS*
P2 peers	0.001*	0.008**	0.011**			
	(0.000)	(0.003)	(0.005)			
Composite				−0.002	0.090**	0.108**
				(0.004)	(0.037)	(0.054)
Class size	0.002***	0.001	0.006*	0.002***	0.001**	0.005
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)
Observations	190,704	190,704	190,704	190,704	190,704	190,704
Number of schools	1,437	1,437	1,437	1,437	1,437	1,437
Class-size instrumented	No	No	Yes	No	No	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		552.7	212.7		366.2	190.5
Panel B: Literacy – Performing at least at level						
	(7) *OLS*	*(8)* *2SLS*	*(9)* *2SLS*	*(10)* *OLS*	*(11)* *2SLS*	*(12)* *2SLS*
P2 peers	0.001**	0.013***	0.015**			
	(0.000)	(0.004)	(0.007)			
Composite				0.003	0.158***	0.152**
				(0.004)	(0.046)	(0.067)
Class size	0.002***	0.001	0.004	0.002***	0.001**	0.002
	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)
Observations	190,704	190,704	190,704	190,704	190,704	190,704
Number of schools	1,437	1,437	1,437	1,437	1,437	1,437
Class-size instrumented	No	No	Yes	No	No	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		552.7	212.7		366.2	190.5

Notes: Heteroscedasticity-robust SEs adjusted for clustering at the school-by-year level are reported in parentheses. This table shows the results for our estimation of equation (1) by Ordinary Least Squares (OLS) and 2-Stage-Least-Squares (2SLS) regression. Our outcomes of interest are dummy indicators for whether a pupil performs at least at the expected level in numeracy or literacy, respectively. All results refer to our sample of first graders (P1). Covariates include pupil age, sex, and ethnicity, an indicator for whether pupil is from a neighbourhood in bottom 20% of deprivation (SIMD), whether a pupil attends a school outside their catchment area, grade enrolment counts and its square, the size of the school, and the percentage of pupils in a school that are female, white British, native English speakers, and in the bottom 20% of deprivation, respectively. All specifications contain a set of school fixed effects and a set of year fixed effects. The reported first-stage F-statistic is HAC and was calculated using the method developed by Kleibergen and Paap (2006).
***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

point estimate is not statistically significant at any reasonable level of significance. The standard errors are also small enough to rule out effects that are large enough to offset the gains to first graders. In the same vein, we find small statistically insignificant negative effects on second graders' literacy (see columns (6) and (8)). This is in contrast to OLS estimates (columns (5) and (7)) which suggest statistically significant detrimental effects, but which are biased due to negative selection of P2 pupils into P1/P2 multi-grade classes. One caveat here is that second-graders are not assessed in the same year that they share a classroom with first graders, but only once they get to fourth grade. Hence, the main takeaway from panel A of Table 7 is that there is no evidence for medium-term adverse

TABLE 7
Second-stage results – performance in fourth grade (P4)

	Numeracy				*Literacy*			
	(1) OLS	*(2) 2SLS*	*(3) OLS*	*(4) 2SLS*	*(5) OLS*	*(6) 2SLS*	*(7) OLS*	*(8) 2SLS*
Panel A: Performance of second graders (P2) in fourth grade (P4)								
P1 peers	−0.006***	−0.002			−0.006***	−0.003		
	(0.001)	(0.004)			(0.001)	(0.004)		
Composite			−0.054***	−0.020			−0.052***	−0.024
			(0.006)	(0.039)			(0.006)	(0.042)
Class size	0.002***	0.003**	0.002***	0.003**	0.002***	0.003*	0.002***	0.003*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	194,666	194,666	194,666	194,666	194,666	194,666	194,666	194,666
Number of schools	1,449	1,449	1,449	1,449	1,449	1,449	1,449	1,449
F-Stat		346.6		282.5		346.6		282.5
Panel B: Performance of first graders (P1) in fourth grade (P4)								
P2 peers	0.004***	−0.005			0.004***	−0.001		
	(0.000)	(0.004)			(0.000)	(0.004)		
Composite			0.032***	−0.055			0.036***	−0.007
			(0.004)	(0.046)			(0.004)	(0.049)
Class size	0.001	0.001**	0.001*	0.001**	0.001**	0.001**	0.001**	0.001***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	192,428	192,428	192,428	192,428	192,428	192,428	192,428	192,428
Number of schools	1,443	1,443	1,443	1,443	1,443	1,443	1,443	1,443
F-Stat		442.8		305.9		442.8		305.9

Notes: Heteroscedasticity-robust SEs adjusted for clustering at the school-by-year level are reported in parentheses. This table shows the results for our estimation by Ordinary Least Squares (OLS) and 2-Stage-Least-Squares (2SLS) regression. In Panel A, our outcomes of interest are dummy (0/1) indicators for whether a second grader (P2) performs at least at the expected level in numeracy or literacy two years later in fourth grade (P4). In Panel B it is the same measure but for first-graders (P1) when assessed in P4. The explanatory variable measures the number of younger P1 peers or older P2 peers a pupil was exposed to by way of a P1/P2 composite class. All specifications include covariates for pupil age, sex, and ethnicity, an indicator for whether pupil is from a neighbourhood in bottom 20% of deprivation (SIMD), whether a pupil attends a school outside their catchment area, grade enrolment counts and their squared values, the size of the school, and the percentage of pupils in a school that are female, White British, native English speakers, and in the bottom 20% of deprivation, respectively. All specifications also contain a set of school fixed effects and a set of year fixed effects. The reported first-stage F-statistic is HAC and was calculated using the method developed by Kleibergen and Paap (2006).
***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

effects of P1/P2 multi-grade classes on those second graders who are placed in these classes.

Panel B of Table 7 shows our results for first-graders' performance once they have progressed to fourth grade (P4). Of course, pupils are subject to a variety of other influences as they progress from P1 to P4, all of which may amplify or mitigate the effects of starting school in a multi-grade class. This makes it challenging to identify a clean and precisely estimated effect of P1/P2 attendance on attainment in P4. The OLS estimates for multi-grade status in first grade are all positive, reflecting the positive selection of P1s into P1/P2 composite classes. However, our 2SLS estimates document that once they have progressed to fourth grade, there is no statistically significant difference in attainment between pupils who shared a classroom with second graders when they were in first grade and those who were in single-year groupings. In other words, the attainment

gains shown in Table 6 appear to fade out over time. This pattern can partly be explained by the transitory nature of composite classes. Only about 22% of pupils who were in a multi-grade class in the previous school year remain in a multi-grade class the year after. Indeed only 2.2% of pupils who start primary school in a multi-grade setting remain in such a setting throughout primary school whereas almost half the pupils who start primary school in a single-year class room experience a multi-grade setting at a later point. This is because the class planning algorithm is sensitive to small changes in enrolment which can trigger a reshuffling of pupils into single-year and composite classes every year. After first-grade, pupils may be grouped with either older or younger peers, which makes these dynamics hard to model. Panel B of Table 7 suggests that this lack of persistence in peer effects could drive a medium-run regression towards the mean.

Mechanisms and heterogeneity

So far, we have said little about the mechanisms that might underpin the large, statistically significant short-run effects of multi-grade classes that we set out in the previous section. Here we explore six potential explanations and describe what our analysis tells us about each: the role of class size, breaks in peer groups and social stigma, whether the type of activity assessed reveals anything about the mechanism, potential socioeconomic channels, whether there might be additional staffing support and resources, and gender composition effects.

Throughout our analysis, we control for class size and report the corresponding regression output. In all tables it has been noticeable that the effect of class size tends to be economically insignificant and in most specifications it is also statistically insignificant. Similar to Leuven, Oosterbeek, and Rønning's (2008) study of Norwegian middle schools, the class size coefficients in Table 6 are very small and positive, range from 0.001 to 0.006 depending on the specification, and none of them are statistically significant at the 5% level. This is not surprising as both single-year P1 and multi-grade P1/P2 classes are capped at 25 pupils and consequently have virtually identical average class sizes (21.8 for single-year, 22.0 for composite classes). It is thus unlikely that class size is driving these positive effects.[10]

Note that our analysis focuses on school starters. That makes it unlikely that *breaks* in peer groups are driving our results. Scottish primary schools do not typically have kindergarten grades but take in first-graders from a variety of smaller day-cares. While school and social networks are clearly important (see Crosnoe, Cavanagh, and Elder Jr, 2003; Bramoullé, Djebbari, and Fortin, 2009; De Giorgi, Pellizzari, and Redaelli, 2010; De Giorgi and Pellizzari, 2014; Patacchini, Rainone, and Zenou, 2017; Lavy and Sand, 2019), they are only beginning to form in first grade. In the same vein, it is unlikely that stigma or feelings of inferiority (or superiority) are driving our results.

[10]We also deployed an instrumental variable strategy in which class size is instrumented by class size predictions that are obtained exploiting maximum class size cutoffs (see columns (3) and (6) of Table 6). This identification is in the mould of Angrist and Lavy's (1999) seminal work and their recent follow-up study (Angrist *et al.*, 2019). Appendix S2 elaborates on this approach.

TABLE 8
Second-stage results (P1) for literacy subcategories

	Reading		*Writing*		*Listening and talking*	
	(1) *OLS*	*(2)* *2SLS*	*(3)* *OLS*	*(4)* *2SLS*	*(5)* *OLS*	*(6)* *2SLS*
P2 peers	0.001*	0.008**	0.001	0.012***	0.000	0.004
	(0.000)	(0.003)	(0.000)	(0.004)	(0.000)	(0.003)
Class size	0.002***	0.001*	0.002***	0.001	0.002***	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	190,704	190,704	190,704	190,704	190,704	190,704
Number of schools	1,437	1,437	1,437	1,437	1,437	1,437
School FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		552.7		552.7		552.7

Notes: Heteroscedasticity-robust SEs adjusted for clustering at the school-by-year level are reported in parentheses. This table shows the results for our estimation of equation 1 by Ordinary Least Squares (OLS) and 2-Stage-Least-Squares (2SLS) regression. Our outcomes of interest are dummy indicators for whether a pupil performs at least at the expected level in three subcategories of literacy. All results refer to our sample of first graders (P1). Covariates include pupil age, sex, and ethnicity an indicator for whether pupil is from a neighbourhood in bottom 20% of deprivation (SIMD), whether a pupil attends a school outside their catchment area, grade enrolment counts and its square, the size of the school, and the percentage of pupils in a school that are female, White British, native English speakers, and in the bottom 20% of deprivation respectively. All specifications contain a set of school fixed effects and a set of year fixed effects. The reported first-stage F-statistic is HAC and was calculated using the method developed by Kleibergen and Paap (2006).
***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

Five-year old school starters will have no reference point for their experienced class structure.[11]

Our finding that there are larger and more pronounced gains for literacy compared to numeracy suggests that the type of activity being assessed might shed some light on potential mechanisms. Table 8 breaks down our literacy assessment into its three components: reading, writing, and listening and talking. These subcategories may offer pointers on the channel through which exposure to more mature peers improves literacy. While listening and talking are – by definition – interactive activities, reading and writing can be improved by working on one's own. Columns (2) and (4) show that the gains appear to be concentrated in improvements in reading and writing ability respectively, whereas the effect for listening and talking (column (6)) are smaller and not statistically significant at the 5% level. While this breakdown does not allow us to fully disentangle these mechanisms, it suggests that it is not the direct interaction with older peers that is driving these improvements. Instead younger pupils may be motivated and spurred on by observing peers who are have already acquired reading and writing proficiency.

Another mechanism that might help to explain our results is if parents invest more effort into supporting their children if they end up in a multi-grade class. More generally, there is growing evidence suggesting that parents from lower socioeconomic strata may provide less educational input to their offspring (Francesconi and Heckman, 2016; Fredriksson, Öckert, and Oosterbeek, 2016). Multi-grade classes may exacerbate these inequalities if gains for first graders are driven by greater investment by affluent parents. While we

[11]Stigma and loss of social networks may, of course, play a more important role in the case of second-graders who are chosen to attend P1/P2 classes.

TABLE 9
Second-stage results (P1): effect heterogeneity

	Numeracy				*Literacy*			
	(1) OLS	*(2) 2SLS*	*(3) OLS*	*(4) 2SLS*	*(5) OLS*	*(6) 2SLS*	*(7) OLS*	*(8) 2SLS*
Panel A: Heterogeneous effects by level of deprivation								
	Top 60% SIMD		Bottom 40% SIMD		Top 60% SIMD		Bottom 40% SIMD	
P2 peers	0.001**	0.006*	0.000	0.009*	0.001**	0.011**	0.001	0.016***
	(0.000)	(0.003)	(0.000)	(0.005)	(0.000)	(0.004)	(0.001)	(0.006)
Observations	106,653	106,653	84,051	84,051	106,653	106,653	84,051	84,051
Number of schools	1,411	1,411	1,269	1,269	1,411	1,411	1,269	1,269
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		376.9		339		376.9		339
Panel B: Heterogeneous effects by school location								
	Urban		*Rural*		*Urban*		*Rural*	
P2 peers	0.001**	0.008**	−0.000	0.005	0.001**	0.012***	0.001	0.017**
	(0.000)	(0.004)	(0.001)	(0.006)	(0.000)	(0.005)	(0.001)	(0.008)
Observations	143,834	143,834	46,870	46,870	143,834	143,834	46,870	46,870
Number of schools	972	972	486	486	972	972	486	486
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		425.1		131.2		425.1		131.2
Panel C: Heterogeneous effects by pupil sex								
	Boys		*Girls*		*Boys*		*Girls*	
P2 peers	0.000	0.009**	0.001**	0.006	0.001	0.016***	0.001***	0.011**
	(0.000)	(0.004)	(0.000)	(0.004)	(0.001)	(0.005)	(0.000)	(0.004)
Observations	97,125	97,125	93,579	93,579	97,125	97,125	93,579	93,579
Number of schools	1,435	1,435	1,435	1,435	1,435	1,435	1,435	1,435
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		480.4		489.1		480.4		489.1

Notes: Heteroscedasticity-robust SEs adjusted for clustering at the school-by-year level are reported in parentheses. This table shows the results for our estimation of equation (1) by Ordinary Least Squares (OLS) and 2-Stage-Least-Squares (2SLS) regression. Our outcomes of interest are dummy indicators for whether a pupil performs at least at the expected level in numeracy or literacy, respectively. All results refer to our sample of first graders (P1). Unless they are the category of interest, covariates include pupil age, sex, and ethnicity, an indicator for whether pupil is from a neighbourhood in bottom 20% of deprivation (SIMD), whether a pupil attends a school outside their catchment area, grade enrolment counts and its square, the size of the school, and the percentage of pupils in a school that are female, white British, native English speakers, and in the bottom 20% of deprivation respectively. All specifications contain a set of school fixed effects and a set of year fixed effects. The reported first-stage F-statistic is HAC and was calculated using the method developed by Kleibergen and Paap (2006).
***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

cannot directly measure parental effort, we can explore whether there are differences in our results across socioeconomic status. Panel A of Table 9 indicates few such differences. In fact, our point estimates suggest that pupils from postcodes which are ranked in the two bottom quintiles in terms of deprivation tend to benefit slightly more from exposure to more experienced peers than pupils in the top three quintiles. However, these differences are not significant at any reasonable level of statistical significance.

In discussions with educational decision makers in Scotland it became clear that there is neither special training, nor additional support for teachers who are in charge of multi-grade classes. That is because teaching approaches in P1 do not differ substantially

between single-year and composite classes. In both setups, group-based teaching is the norm. Teachers periodically evaluate students' learning progress and split them into groups, often seated separately within the same class room. So while the delivery of teaching is all but identical across multi-grade and single year classes, a key difference is that in multi-grade classes P1 pupils may share a table with P2 pupils. First-graders in composite and multi-year classes are also taught the same curriculum. This not only means that P1 pupils are not exposed to more advanced material in composite classes than they would be in a single grade class. It also means that their older class mates should not experience any deviations from the curriculum. This is a potential concern raised by Checchi and De Paola (2018) in motivating the negative effects they find for pupils in the final stages of primary school.

Nevertheless, it might be the case that more teaching resources are provided to the teaching of multi-grade classes and that this helps explain our findings. We explore this dimension as far as we can given the data available. While individual teachers cannot be identified in our data, Table A3 in Appendix S1 shows that there are no differences in terms of staffing (e.g. presence of teaching assistants or additional teachers).

Furthermore, urban schools tend to find teacher recruitment easier and are on average larger which may make them more likely to develop teachers who specialize in the instruction of multi-grade classes. But again, panel B of Table 9 reveals little in the way of effect differences between urban and rural schools.

Finally, we stratify our sample by pupil gender. Heterogeneous effects across gender is a common finding in the peer effects literature. For example, Lavy *et al.* (2012b) show that girls benefit more from exposure to high-ability peers than boys. There is also some evidence that educational inputs have a stronger influence on girls compared to boys (Anderson, 2008; Angrist, Lang, and Oreopoulos, 2009; Lavy *et al.*, 2012b). However, some of this evidence is from studies looking at secondary schools, where social interaction and gender identity effects are likely to be stronger than in the early years primary school setting we investigate. In our case, panel C of Table 9 shows that we cannot reject the null hypothesis of no gender differences, even though our point estimates for boys tend to be larger than those for girls in both literacy and numeracy.

V. Conclusion

This study explores the impact of sharing a multi-grade classroom with more experienced peers in early primary school. We combine population-level pupil data with an instrumental variables estimation strategy that exploits exogenous variation in the creation of multi-grade classes generated by a class planning algorithm. We find that the presence of second graders improves first-graders' reading, writing and maths performance, as measured by teacher assessments that are informed by standardized test scores. It is important to note that we estimate a LATE. That is, these benefits may not accrue to the average school-starter but only to the oldest cohort members who – if assigned to multi-grade classes – are typically exposed to second-graders by way of a multi-grade classes. While these effects wash out over time, we also find no evidence of a detrimental impact of the classroom presence of younger first-graders on those second-graders who make up the older component of multi-grade classes.

Our paper adds to two strands of literature. First, our findings are consistent with, and generalize beyond, recent research on multi-grade classes that exploited that small population variations in sparsely populated areas of Norway (Leuven and Rønning, 2014) and Italy (Checchi and De Paola, 2018; Barbetta *et al.*, 2019) lead to the lumping together of grades in rural middle and elementary school, respectively. We show that the benefits of exposure to older pupils by way of a multi-grade class, also accrue in urban settings where multi-grade classes are created by design and where school-starters are placed in multi-grade classes often for only one year at a time. While further research in this area is certainly warranted, the overall body of evidence suggests that multi-grade classes, especially in the early years of primary education, have the potential to be a useful tool to stimulate the learning of academically strong and relatively mature pupils by exposing them to older, more experienced peers.

Second, we contribute to an important literature on peer effects. We demonstrate that first graders benefit from exposure to more mature peers with an additional year of primary schooling under their belt. Our research thus re-enforces the common finding that externalities from peers are important determinants of pupil attainment. In fact, our study suggests that these spillovers are more important than conventional education production inputs, such as class size. As such, our findings also have important implications for policymakers and education practitioners. Our study suggests that multi-grade classes deliver better learning outcomes for first-graders while simultaneously acting as a way for policymakers to allocate resources more efficiently.

Finally, our paper suggests that an interesting avenue for future research may be a further investigation of gender differences. In particular the question whether multi-grade classes affect boys and girls differentially – depending on the respective number or fraction of boys and girls among the more senior students in a multi-grade class – constitutes an important avenue for further studies.

Final Manuscript Received: January 2022

References

Aizer, A. (2008). *Peer Effects and Human Capital Accumulation: The Externalities of ADD*, National Bureau of Economic Research (NBER) Working Paper No. 14354.

Anderson, M. L. (2008). 'Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry preschool, and early training projects', *Journal of the American Statistical Association*, Vol. 103, pp. 1481–1495.

Anelli, M. and Peri, G. (2019). 'The effects of high school peers' gender on college major, college performance and income', *The Economic Journal*, Vol. 129, pp. 553–602.

Angrist, J. D. and Lang, K. (2004). 'Does school integration generate peer effects? Evidence from Boston's Metco Program', *American Economic Review*, Vol. 94, pp. 1613–1634.

Angrist, J. D. and Lavy, V. (1999). 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement', *The Quarterly Journal of Economics*, Vol. 114, pp. 533–575.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). 'Identification of causal effects using instrumental variables', *Journal of the American statistical Association*, Vol. 91, pp. 444–455.

Angrist, J., Lang, D. and Oreopoulos, P. (2009). 'Incentives and services for college achievement: evidence from a randomized trial', *American Economic Journal: Applied Economics*, Vol. 1, pp. 136–163.

Angrist, J. D., Lavy, V., Leder-Luis, J. and Shany, A. (2019). 'Maimonides' rule redux', *American Economic Review: Insights*, Vol. 1, pp. 309–324.

Ballatore, R. M., Fort, M. and Ichino, A. (2018). 'Tower of Babel in the classroom: immigrants and natives in Italian schools', *Journal of Labor Economics*, Vol. 36, pp. 885–921.
Ballatore, R. M., Paccagnella, M. and Tonello, M. (2020). 'Bullied because younger than my mates? The effect of age rank on victimisation at school', *Labour Economics*, Vol. 62, 101772.
Barban, N., De Cao, E., Oreffice, S. and Quintana-Domeque, C. (2021). 'The effect of education on spousal education: a genetic approach', *Labour Economics*, Vol. 71, 102023.
Barbetta, G. P., Sorrenti, G. and Turati, G. (2019). 'Multigrading and child achievement', *Journal of Human Resources*, Vol. 56, pp. 940–968.
Bedard, K. and Dhuey, E. (2006). 'The persistence of early childhood maturity: international evidence of long-run age effects', *The Quarterly Journal of Economics*, Vol. 121, pp. 1437–1472.
Betts, J. R. (2011). 'The economics of tracking in education', in Hanushek E. A., Machin S. and Woessmann L. (eds), *Handbook of the Economics of Education*, Elsevier, Amsterdam, Vol. 3, pp. 341–381.
Bifulco, R., Fletcher, J. M. and Ross, S. L. (2011). 'The effect of classmate characteristics on post-secondary outcomes: evidence from the Add Health', *American Economic Journal: Economic Policy*, Vol. 3, pp. 25–53.
Bifulco, R., Fletcher, J. M., Oh, S. J. and Ross, S. L. (2014). 'Do high school peers have persistent effects on college attainment and other life outcomes?', *Labour Economics*, Vol. 29, pp. 83–90.
Black, S. E., Devereux, P. J. and Salvanes, K. G. (2011). 'Too young to leave the nest? The effects of school starting age', *The Review of Economics and Statistics*, Vol. 93, pp. 455–467.
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, pp. 119–153.
Bound, J., Jaeger, D. A. and Baker, R. M. (1995). 'Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak', *Journal of the American Statistical Association*, Vol. 90, pp. 443–450.
Bramoullé, Y., Djebbari, H. and Fortin, B. (2009). 'Identification of peer effects through social networks', *Journal of Econometrics*, Vol. 150, pp. 41–55.
Carrell, S. E. and Hoekstra, M. L. (2010). 'Externalities in the classroom: how children exposed to domestic violence affect everyone's kids', *American Economic Journal: Applied Economics*, Vol. 2, pp. 211–228.
Carrell, S. E. and Hoekstra, M. (2012). 'Family business or social problem? The cost of unreported domestic violence', *Journal of Policy Analysis and Management*, Vol. 31, pp. 861–875.
Carrell, S. E., Hoekstra, M. and Kuka, E. (2018). 'The long-run effects of disruptive peers', *American Economic Review*, Vol. 108, pp. 3377–3415.
Cascio, E. U. and Schanzenbach, D. W. (2016). 'First in the class? Age and the education production function', *Education Finance and Policy*, Vol. 11, pp. 225–250.
Checchi, D. and De Paola, M. (2018). 'The effect of multigrade classes on cognitive and non-cognitive skills. Causal evidence exploiting minimum class size rules in Italy', *Economics of Education Review*, Vol. 67, pp. 235–253.
Clarke, D. and Matta, B. (2018). 'Practical considerations for questionable IVs', *The Stata Journal*, Vol. 18, pp. 663–691.
Conley, T. G., Hansen, C. B. and Rossi, P. E. (2012). 'Plausibly exogenous', *Review of Economics and Statistics*, Vol. 94, pp. 260–272.
Crawford, C., Dearden, L. and Greaves, E. (2014). 'The drivers of month-of-birth differences in children's cognitive and non-cognitive skills', *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 177, p. 829.
Crosnoe, R., Cavanagh, S. and Elder, G. H., Jr. (2003). 'Adolescent friendships as academic resources: the intersection of friendship, race, and school disadvantage', *Sociological Perspectives*, Vol. 46, pp. 331–352.
De Giorgi, G. and Pellizzari, M. (2014). 'Understanding social interactions: Evidence from the classroom', *The Economic Journal*, Vol. 124, pp. 917–953.
De Giorgi, G., Pellizzari, M. and Redaelli, S. (2010). 'Identification of social interactions through partially overlapping peer groups', *American Economic Journal: Applied Economics*, Vol. 2, pp. 241–275.

Ding, W. and Lehrer, S. F. (2007). 'Do peers affect student achievement in China's secondary schools?', *The Review of Economics and Statistics*, Vol. 89, pp. 300–312.
Figlio, D. N. (2007). 'Boys named Sue: disruptive children and their peers', *Education Finance and Policy*, Vol. 2, pp. 376–394.
Francesconi, M. and Heckman, J. J. (2016). 'Child development and parental investment: introduction', *The Economic Journal*, Vol. 126, pp. F1–F27.
Fredriksson, P., Öckert, B. and Oosterbeek, H. (2016). 'Parental responses to public investments in children: evidence from a maximum class size rule', *Journal of Human Resources*, Vol. 51, pp. 832–868.
Fruehwirth, J. C. (2013). 'Identifying peer achievement spillovers: implications for desegregation and the achievement gap', *Quantitative Economics*, Vol. 4, pp. 85–124.
Gould, E. D., Lavy, V. and Daniele Paserman, M. (2009). 'Does immigration affect the long-term educational outcomes of natives? Quasi-experimental evidence', *The Economic Journal*, Vol. 119, pp. 1243–1269.
Hanushek, E. A. and Rivkin, S. G. (2009). 'Harming the best: how schools affect the black-white achievement gap', *Journal of Policy Analysis and Management*, Vol. 28, pp. 366–393.
Hanushek, E. A., Kain, J. F., Markman, J. M. and Rivkin, S. G. (2003). 'Does peer ability affect student achievement?', *Journal of Applied Econometrics*, Vol. 18, pp. 527–544.
Hanushek, E. A., Kain, J. F. and Rivkin, S. G. (2009). 'New evidence about Brown v. Board of Education: the complex effects of school racial composition on achievement', *Journal of Labor Economics*, Vol. 27, pp. 349–383.
von Hinke, S. (2022). 'Education, dietary intakes and exercise', *Oxford Bulletin of Economics and Statistics*, Vol. 84, pp. 214–240.
Hoxby, C. (2000). *Peer Effects in the Classroom: Learning from Gender and Race Variation*, National Bureau of Economic Research (NBER) Working Paper No. 7867.
Hoxby, C. M and Weingarth, G. (2005). *Taking Race Out of the Equation: School reassignment and the Structure of Peer Effects*, Unpublished.
Kleibergen, F. and Paap, R. (2006). 'Generalized reduced rank tests using the singular value decomposition', *Journal of Econometrics*, Vol. 133, pp. 97–126.
Lavy, V. and Sand, E. (2019). 'The effect of social networks on students' academic and non-cognitive behavioural outcomes: evidence from conditional random assignment of friends in school', *The Economic Journal*, Vol. 129, pp. 439–480.
Lavy, V. and Schlosser, A. (2011). 'Mechanisms and impacts of gender peer effects at school', *American Economic Journal: Applied Economics*, Vol. 3, pp. 1–33.
Lavy, V., Paserman, M. D. and Schlosser, A. (2012a). 'Inside the black box of ability peer effects: evidence from variation in the proportion of low achievers in the classroom', *The Economic Journal*, Vol. 122, pp. 208–237.
Lavy, V., Silva, O. and Weinhardt, F. (2012b). 'The good, the bad, and the average: evidence on ability peer effects in schools', *Journal of Labor Economics*, Vol. 30, pp. 367–414.
Lee, D. S., McCrary, J., Moreira, M. J. and Porter, J. (2022). 'Valid t-ratio inference for IV', *American Economic Review*, Vol. 112, pp. 3260–3290.
Lefgren, L. (2004). 'Educational peer effects and the Chicago public schools', *Journal of Urban Economics*, Vol. 56, pp. 169–191.
Leuven, E. and Rønning, M. (2014). 'Classroom grade composition and pupil achievement', *The Economic Journal*, Vol. 126, pp. 1164–1192.
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, pp. 663–693.
Neidell, M. and Waldfogel, J. (2010). 'Cognitive and noncognitive peer effects in early education', *The Review of Economics and Statistics*, Vol. 92, pp. 562–576.
Patacchini, E., Rainone, E. and Zenou, Y. (2017). 'Heterogeneous peer effects in education', *Journal of Economic Behavior & Organization*, Vol. 134, pp. 190–227.
Rossi, G. (2021). *School Performance, Non-Cognitive Skills and House Prices*, Strathclyde Discussion Papers in Economics No. 21 - 2.

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, pp. 457–478.
Slavin, R. E. (1987). 'Ability grouping and student achievement in elementary schools: a best-evidence synthesis', *Review of Educational Research*, Vol. 57, pp. 293–336.
Stock, J. H. and Yogo, M. (2002). *Testing for Weak Instruments in Linear IV Regression*, National Bureau of Economic Research (NBER) Working Paper No. 0284.

Supporting Information

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

Appendix S1. Supporting information
Appendix S2. Supporting information