I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14678

Early-Years Multi-Grade Classes and Pupil Attainment

Daniel Borbely
Markus Gehrsitz
Stuart McIntyre
Gennaro Rossi
Graeme Roy

AUGUST 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14678

Early-Years Multi-Grade Classes and Pupil Attainment

Daniel Borbely
University of Dundee

Markus Gehrsitz
University of Strathclyde and IZA

Stuart McIntyre
University of Strathclyde

Gennaro Rossi
University of Strathclyde

Graeme Roy
University of Glasgow

AUGUST 2021

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Early-Years Multi-Grade Classes and Pupil Attainment*

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 find no 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 do not come at the expense of older peers from the preceding cohort.

JEL Classification: C36, H52, I21, I26, I28, J24
Keywords: multi-grade classes, peer effects, class-size, cognitive skills

Corresponding author:
Markus Gehrsitz
Department of Economics and Fraser of Allander Institute
University of Strathclyde
Glasgow
United Kingdom
E-mail: markus.gehrsitz@strath.ac.uk

* We are grateful to Antonio Acconcia, Emma Congreve, Susan Ellis, Gordon McKinlay, Ian Walker and Tanya Wilson as well as to 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.

1 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 to high-ability peers (Hanushek et al. 2003; Lefgren 2004; Ding and Lehrer 2007; Neidell and Waldfogel 2010; Lavy et al. 2012a,b), to female classmates (Hoxby 2000; Lavy and Schlosser 2011; Black et al. 2013; Anelli and Peri 2019) and to classmates with college-educated mothers (Bifulco et al. 2011, 2014) as well as the adverse effects of disruptive peers (Figlio 2007; Aizer 2008; Carrell and Hoekstra 2010, 2012; Carrell et al. 2018). The ethnic makeup of classrooms (Angrist and Lang 2004; Hoxby and Weingarth 2005; Hanushek et al. 2009; Hanushek and Rivkin 2009; Fruehwirth 2013) and the effect of immigrant peers on natives (Gould et al. 2009; Ballatore et al. 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.

Multi-grade classes are widely used. About 28% of schools in the US 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, with the notable exceptions of studies of rural areas of Norway (Leuven and Rønning 2014) and Italy (Checchi and De Paola 2018; Barbetta et al. 2019) where cohorts are often so small that pooling several year-groups is done out of necessity. By contrast in Scotland, a constituent nation of the United Kingdom and the subject of our study, multi-grade classes are consciously created in virtually all primary schools.

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 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.[1] 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

[1]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.

to 2018/19. We link these data with assessment information and observe the exact classroom type and composition in each school-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 et al. (1995) and Lee et al. (2020)). This paper, 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. Boys and pupils from deprived neighbourhoods appear to benefit more from sharing a classroom with more experienced peers, although neither gender nor socioeconomic 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 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 et al. 2011; Crawford et al. 2014; Cascio and Schanzenbach 2016; Ballatore et al. 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.

2 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 United States. 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 5% of pupils in our sample attend a school different from the one of their catchment area.[2] 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%.

[2]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.

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. 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. Composite classes typically stretch across two grades and more than one in six Scottish primary school pupils attend a multi-grade class.[3] In contrast to 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. 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 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

[3]Figure A1 in the Appendix provides an illustration of the distribution of pupils across single-year and multi-grade classes in 2018.

multi-grade classes and calculate class sizes which we cross-checked with official aggregates published by the Scottish Government. Appendix Table A1 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 'datazones' (small area statistical geographies) 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.[4]

3 Empirical Design

Our aim is to compare attainment between pupils who attend multi-grade classes and those in single-year classes. 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:

$$A_{icgst} = \beta_0 + \beta_1 Comp_{cgst} + \gamma X_{igst} + \delta_s + \tau_t + \varepsilon_{icgst} \tag{1}$$

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

[4]Our sample also differs marginally from the original population data. We excluded the 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.

such as age, gender, ethnicity, and socio-economic background, 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 and school-year fixed-effects, respectively.

Our main empirical concern relates to the endogeneity of $Comp_{cgst}$. 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 1 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. As a result a simple OLS estimation of equation (1) is likely to be severely biased.

To overcome this endogeneity problem, we use exogenous variation 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 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 1a 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 1b, 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 1a 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 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 regressions corresponding to the following equation:

$$Comp_{icgst} = \alpha_0 + \alpha_1 Comp_{gst}^{pred} + \gamma X_{igst} + \delta_s + \tau_t + \varepsilon_{icgst} \tag{2}$$

Where $Comp_{icgst}$ is a dummy indicator for whether class c in grade g which contains pupil i, is a multi-grade class whereas $Comp_{gst}^{pred}$ is an indicator for 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). Recent work by Lee et al. (2020) suggests that in order to achieve valid estimation parameters, an F-statistic of larger than 104 is required. Our estimation of equation (2) indicates a strong first stage for our sample of first-graders with F-statistics of 368 and 556, respectively. These are displayed at the bottom of our second stage results Table 2 in Section 4.[5] All F-statistics are heteroscedasticity and autocorrelation consistent (HAC) and were obtained using the method developed by Kleibergen and Paap (2006). Note that by contrast, our first-stage results for P4 and P7 (the only other two stages with outcome data) are well below any $F > 104$ threshold. We therefore focus our analysis on first-graders but report our results for fourth and seventh-graders for completeness in Appendix Table A4.

There are several reasons for greater compliance with class planner predictions in early years compared to later years. 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.

Our identification strategy also requires our instrument to be exogenous and the exclusion restriction to hold.[6] Class planner predictions are ultimately generated by random population variation, making the exogeneity assumption credible. The exclusion restriction

[5]Our full first stage results are reported in Table A2 in the Appendix.

[6]It is unlikely that our research design features "defiers" which would violate the monotonicity assumption.

requires planner predictions 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. Our instrument is, therefore, unlikely to be correlated with parent or school characteristics that have an independent effect on our outcome of interest.

Hence, β_1 of equation (1) will yield a local average treatment effect (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 1, 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 1 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 0.4% 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. It is what makes these effects "local". Indeed, our instrumental variable technique addresses exactly this selection issue. Intuitively, 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.

4 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 school-year fixed effects. Standard errors are adjusted for clustering at the school and year level throughout.

4.1 Second Stage Results

Columns (2) and (3) of Panel A in Table 2 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)). These sizable effects stand in contrast to naïve OLS estimates in column (1) which indicate a precisely estimated zero effect. 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 (2) and (3) 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.[7]

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 standard deviations. Our point estimates suggest improvements of 0.28 standard deviations for numeracy and 0.35 standard deviations 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 3 indicates a small negative effect on maths assessments of second graders who shared a multi-grade classrooms with first-graders. However, the 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 3 is that there is no evidence for medium-term adverse effects of P1/P2 multi-grade classes on those second graders who are placed in these classes.

Panel B of Table 3 shows our results for first-graders' performance once they have pro-

[7]Columns (1) and (2) of Appendix Table A3 report the reduced form estimates. Not surprisingly, the ratio of our reduced-form and first stage effects is approximately equal to our 2SLS coefficients.

gressed 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. 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 2 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. 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, but Panel B of Table 3 suggests that this lack of persistence in peer effects could drive a medium-run regression to the mean.

4.2 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 set out 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 reveal 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 both statistically and economically insignificant in virtually all specifications. Similar to Leuven et al.'s (2008) study of Norwegian middle schools (but in contrast to Fredriksson et al's

(2013) findings in Sweden), the class size coefficients in Table 2 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 hardly 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.[8]

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 (De Giorgi and Pellizzari (2014) and Lavy and Sand (2019), among others), 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. Five-year old school starters will have no reference point for their experienced class structure.

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. Panel D of Table 4 breaks down our literacy assessment into its three components: reading, writing, and listening & 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

[8]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 2). 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 B elaborates on this approach.

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 et al. 2016). Multi-grade classes may exacerbate these inequalities if gains for first graders are driven by greater investment by affluent parents. While we cannot directly measure parental effort, we can explore whether there are differences in our results across socioeconomic status. Panel A of Table 4 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 those teaching multi-grade classes. First-graders in composite and multi-year classes are also taught the same curriculum. 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, Appendix Table A5 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 specialise in the instruction of multi-grade classes. But again, Panel B of Table 4 reveals little in the way of effect differences between urban and rural schools.

Finally, we stratify our sample by pupil gender. There is an extensive literature that shows gender differences (see e.g. Lavy et al. (2012b)). Panel C of Table 4 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.

5 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 local average treatment effect (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, the existing 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 implications for policymakers in the UK and beyond. 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.

References

Aizer, Anna, "Peer effects and human capital accumulation: The externalities of ADD," *National Bureau of Economic Research (NBER) Working Paper no. 14354*, 2008.

Anelli, Massimo and Giovanni Peri, "The effects of high school peers' gender on college major, college performance and income," *The Economic Journal*, 2019, *129* (618), 553–602.

Angrist, Joshua D and Kevin Lang, "Does school integration generate peer effects? Evidence from Boston's Metco Program," *American Economic Review*, 2004, *94* (5), 1613–1634.

_ **and Victor Lavy**, "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, 1999, *114* (2), 533–575.

_ , _ , **Jetson Leder-Luis, and Adi Shany**, "Maimonides' Rule Redux," *American Economic Review: Insights*, 2019, *1* (3), 309–24.

Ballatore, Rosario Maria, Marco Paccagnella, and Marco Tonello, "Bullied because younger than my mates? The effect of age rank on victimisation at school," *Labour Economics*, 2020, *62*, 101772.

_ **, Margherita Fort, and Andrea Ichino**, "Tower of Babel in the classroom: immigrants and natives in Italian schools," *Journal of Labor Economics*, 2018, *36* (4), 885–921.

Barbetta, Gian Paolo, Giuseppe Sorrenti, and Gilberto Turati, "Multigrading and child achievement," *Journal of Human Resources*, 2019, pp. 0118–9310R4.

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

Bifulco, Robert, Jason M Fletcher, and Stephen L Ross, "The effect of classmate characteristics on post-secondary outcomes: Evidence from the Add Health," *American Economic Journal: Economic Policy*, 2011, *3* (1), 25–53.

_ , _ , **Sun Jung Oh, and Stephen L Ross**, "Do high school peers have persistent effects on college attainment and other life outcomes?," *Labour Economics*, 2014, *29*, 83–90.

Black, Sandra E, Paul J Devereux, and Kjell G Salvanes, "Too young to leave the nest? The effects of school starting age," *The Review of Economics and Statistics*, 2011, *93* (2), 455–467.

_ , _ , **and** _ , "Under pressure? The effect of peers on outcomes of young adults," *Journal of Labor Economics*, 2013, *31* (1), 119–153.

Bound, John, David A Jaeger, and Regina M Baker, "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American Statistical Association*, 1995, *90* (430), 443–450.

Carrell, Scott E and Mark Hoekstra, "Family business or social problem? The cost of unreported domestic violence," *Journal of Policy Analysis and Management*, 2012, *31* (4), 861–875.

_ **and Mark L Hoekstra**, "Externalities in the classroom: How children exposed to domestic violence affect everyone's kids," *American Economic Journal: Applied Economics*, 2010, *2* (1), 211–28.

_ **, Mark Hoekstra, and Elira Kuka**, "The long-run effects of disruptive peers," *American Economic Review*, 2018, *108* (11), 3377–3415.

Cascio, Elizabeth U and Diane Whitmore Schanzenbach, "First in the class? Age and the education production function," *Education Finance and Policy*, 2016, *11* (3), 225–250.

Checchi, Daniele and Maria De Paola, "The effect of multigrade classes on cognitive and non-cognitive skills. Causal evidence exploiting minimum class size rules in Italy," *Economics of Education Review*, 2018, *67*, 235–253.

Crawford, Claire, Lorraine Dearden, and Ellen Greaves, "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)*, 2014, *177* (4), 829.

Ding, Weili and Steven F Lehrer, "Do peers affect student achievement in China's secondary schools?," *The Review of Economics and Statistics*, 2007, *89* (2), 300–312.

Figlio, David N, "Boys named Sue: Disruptive children and their peers," *Education finance and policy*, 2007, *2* (4), 376–394.

Francesconi, Marco and James J Heckman, "Child development and parental investment: Introduction," *The Economic Journal*, 2016, *126* (596), F1–F27.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek, "Long-term effects of class size," *The Quarterly Journal of Economics*, 2013, *128* (1), 249–285.

_ , _ , and _ , "Parental responses to public investments in children: Evidence from a maximum class size rule," *Journal of Human Resources*, 2016, *51* (4), 832–868.

Fruehwirth, Jane Cooley, "Identifying peer achievement spillovers: Implications for desegregation and the achievement gap," *Quantitative Economics*, 2013, *4* (1), 85–124.

Giorgi, Giacomo De and Michele Pellizzari, "Understanding social interactions: Evidence from the classroom," *The Economic Journal*, 2014, *124* (579), 917–953.

Gould, Eric D, Victor Lavy, and M Daniele Paserman, "Does immigration affect the long-term educational outcomes of natives? Quasi-experimental evidence," *The Economic Journal*, 2009, *119* (540), 1243–1269.

Hanushek, Eric A and Steven G Rivkin, "Harming the best: How schools affect the black-white achievement gap," *Journal of Policy Analysis and Management*, 2009, *28* (3), 366–393.

_ , John F Kain, and Steven G Rivkin, "New evidence about Brown v. Board of Education: The complex effects of school racial composition on achievement," *Journal of Labor Economics*, 2009, *27* (3), 349–383.

_ , _ , Jacob M Markman, and Steven G Rivkin, "Does peer ability affect student achievement?," *Journal of Applied Econometrics*, 2003, *18* (5), 527–544.

Hoxby, Caroline, "Peer effects in the classroom: Learning from gender and race variation," *National Bureau of Economic Research (NBER) Working Paper no. 7867*, 2000.

Hoxby, Caroline M and Gretchen Weingarth, "Taking race out of the equation: School reassignment and the structure of peer effects," *Unpublished*, 2005.

Kleibergen, Frank and Richard Paap, "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 2006, *133* (1), 97–126.

Lavy, Victor and Analia Schlosser, "Mechanisms and impacts of gender peer effects at school," *American Economic Journal: Applied Economics*, 2011, *3* (2), 1–33.

_ and Edith Sand, "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*, 2019, *129* (617), 439–480.

_ , **M Daniele Paserman, and Analia Schlosser**, "Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom," *The Economic Journal*, 2012, *122* (559), 208–237.

_ , **Olmo Silva, and Felix Weinhardt**, "The good, the bad, and the average: Evidence on ability peer effects in schools," *Journal of Labor Economics*, 2012, *30* (2), 367–414.

Lee, David L, Justin McCrary, Marcelo J Moreira, and Jack Porter, "Valid t-ratio Inference for IV," *arXiv preprint arXiv:2010.05058*, 2020.

Lefgren, Lars, "Educational peer effects and the Chicago public schools," *Journal of Urban Economics*, 2004, *56* (2), 169–191.

Leuven, Edwin and Marte Rønning, "Classroom grade composition and pupil achievement," *The Economic Journal*, 2014, *126* (593), 1164–1192.

_ , **Hessel Oosterbeek, and Marte Rønning**, "Quasi-experimental estimates of the effect of class size on achievement in Norway," *Scandinavian Journal of Economics*, 2008, *110* (4), 663–693.

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

Neidell, Matthew and Jane Waldfogel, "Cognitive and noncognitive peer effects in early education," *The Review of Economics and Statistics*, 2010, *92* (3), 562–576.

Rossi, Gennaro, "School Performance, non-cognitive skills and house prices," *Strathclyde Discussion Papers in Economics no. 21 - 2*, 2021.

Stock, James H and Motohiro Yogo, "Testing for weak instruments in linear IV regression," *National Bureau of Economic Research (NBER) Working Paper no. 0284*, 2002.

Tables and Figures

Table 1: 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
	(0.005)	(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)	(0.005)	(0.002)	(0.002)	(0.003)
Age (in Years)	0.132***	0.135***		−0.103***	−0.105***		−0.104***
	(0.006)	(0.006)		(0.006)	(0.006)		(0.007)
1st Age Quartile			−0.013***			0.051***	
			(0.002)			(0.004)	
3rd Age Quartile			0.027***			−0.020***	
			(0.003)			(0.003)	
4th Age Quartile			0.098***			−0.028***	
			(0.005)			(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: ***/**/* indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and 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.

Table 2: Second Stage Results - First Graders (P1)

Panel A: Numeracy - Performing at Least at Level						
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	2SLS	2SLS	OLS	2SLS	2SLS
P2 Peers	0.001*	0.008**	0.011**			
	(0.000)	(0.003)	(0.005)			
Composite				−0.002	0.091**	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
No. 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		556.5	212.6		368.2	190.3
Panel B: Literacy - Performing at Least at Level						
	(7)	(8)	(9)	(10)	(11)	(12)
	OLS	2SLS	2SLS	OLS	2SLS	2SLS
P2 Peers	0.001**	0.013***	0.015**			
	(0.000)	(0.004)	(0.007)			
Composite				0.003	0.159***	0.153**
				(0.004)	(0.046)	(0.067)
Class Size	0.002***	0.001	0.004	0.002***	0.002**	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
No. 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		556.5	212.6		368.2	190.3

Notes: ***/**/* indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and 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 neighborhood in bottom 20% of deprivation (SIMD), 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 and school-year fixed effects.

The reported first-stage F-statistic is heteroscedasticity and autocorrelation consistent (HAC) and was calculated using the method developed by Kleibergen and Paap (2006).

Table 3: Second Stage Results - Performance in Fourth Grade (P4)

Panel A: Performance of Second Graders (P2) in Fourth Grade (P4)

	Numeracy				Literacy			
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS	(5) OLS	(6) 2SLS	(7) OLS	(8) 2SLS
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
No. of Schools	1449	1449	1449	1449	1449	1449	1449	1449
F-Stat		346.7		282.5		346.7		282.5

Panel B: Performance of First Graders (P1) in Fourth Grade (P4)

	Numeracy				Literacy			
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS	(5) OLS	(6) 2SLS	(7) OLS	(8) 2SLS
P2 Peers	0.004***	−0.005			0.004***	−0.000		
	(0.000)	(0.004)			(0.000)	(0.004)		
Composite			0.032***	−0.053			0.036***	−0.005
			(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	192,427	192,428	192,427	192,428	192,427	192,428	192,427
No. of Schools	1443	1443	1443	1443	1443	1443	1443	1443
F-Stat		442.7		305.5		442.7		305.5

Notes: ∗∗∗/∗∗/∗ indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and 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 neighborhood in bottom 20% of deprivation (SIMD), 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 and school-year fixed effects.

The reported first-stage F-statistic is heteroscedasticity and autocorrelation consistent (HAC) and was calculated using the method developed by Kleibergen and Paap (2006).

Table 4: 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
No. of Schools	1411	1411	1269	1269	1411	1411	1269	1269
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		376.9		338.5		376.9		338.5
Panel B: Heterogeneous Effects by School Size								
	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
No. of Schools	972	972	486	486	972	972	486	486
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		424.8		131.2		424.8		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	93,575
No. of Schools	1435	1435	1435	1435	1435	1435	1435	1435
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		479.7		489		479.7		489

Panel D: Heterogeneous Effects by Literacy Subcategory

	Literacy Subcategories					
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS	(5) OLS	(6) 2SLS
	Reading		Writing		Listening & Talking	
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)
Observations	190,704	190,704	190,704	190,704	190,704	190,704
No. of Schools	1,437	1,437	1,437	1,437	1,437	1,437
School FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		556.5		556.5		556.5

Notes: $***/**/*$ indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and year level are reported in parentheses.

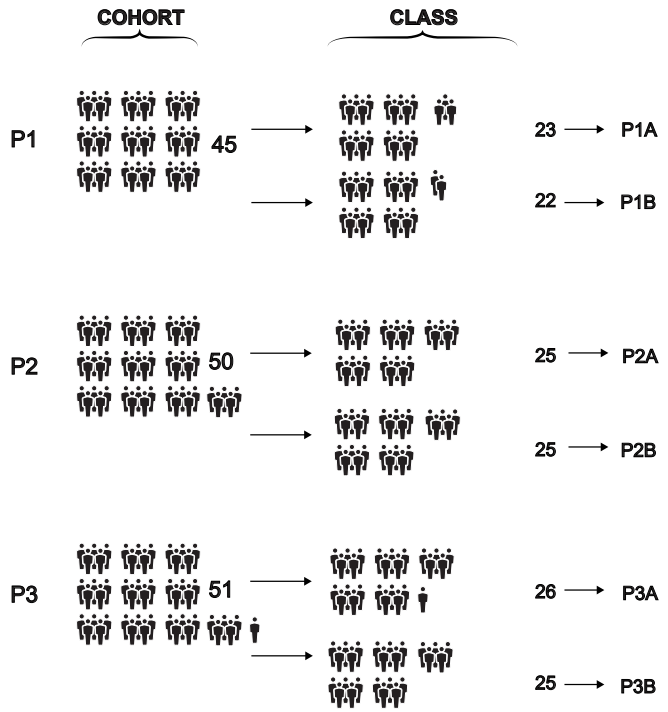
This table shows the results for our estimation of equation (1) by Ordinary Least Squares (OLS) and 2-Stage-Least-Squares (2SLS) regression. In Panels A to C, our outcomes of interest are dummy indicators for whether a pupil performs at least at the expected level in numeracy or literacy, respectively. In Panel D, it is 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).

Unless they are the category of interest, covariates include pupil age, sex, and ethnicity, an indicator for whether pupil is from a neighborhood in bottom 20% of deprivation (SIMD), 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 and school-year fixed effects.

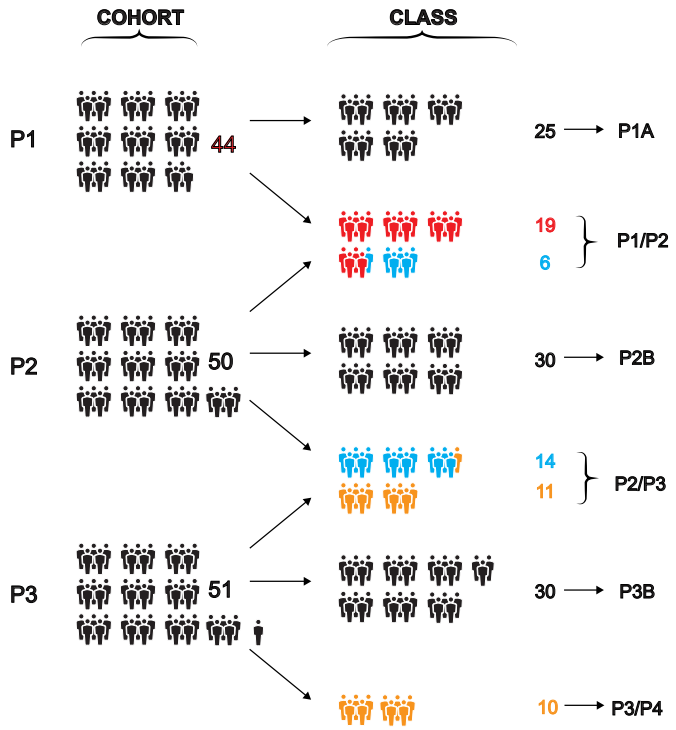
The reported first-stage F-statistic is heteroscedasticity and autocorrelation consistent (HAC) and was calculated using the method developed by Kleibergen and Paap (2006).

Figure 1: Class Planner Examples

(a) Class Planner Example - Scenario 1



(b) Class Planner Example - Scenario 2



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. $Comp_{gst}^{pred} = 0$), in scenario 2 all grades are assigned to treatment.

Appendix A

Table A1: Summary Statistics

	First-Graders (P1)		Fourth-Graders (P4)		Seventh-Graders (P7)	
	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
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 & 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
Free Meal	0.339	0.473	0.179	0.384	0.167	0.373
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
% 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
% Free School Meals	0.247	0.199	0.246	0.197	0.250	0.198
% 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
No. of Students in School	317.454	126.694	319.516	127.876	317.994	128.655
Observations	190,704		194,804		186,082	
No. of Schools	1,437		1,428		1,435	

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

Table A2: 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
$CompLow_{gst}^{pred}$	0.087***	1.038***	0.011***	−0.006	0.006	0.231***		
	(0.005)	(0.044)	(0.004)	(0.005)	(0.039)	(0.040)		
$CompUp_{gst}^{pred}$			0.005	0.021***	0.265***	−0.014	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	0.191	0.151	0.243	0.246	0.139	0.136	0.290	0.217
No. of Schools	1437	1437	1428	1428	1428	1428	1435	1435
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat	368.2	556.5	4.838	4.918	12.61	12.82	12.91	17.99

Notes: $***/**/*$ indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and 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 neighborhood in bottom 20% of deprivation (SIMD), 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 and school-year fixed effects.

The reported F-statistic is heteroscedasticity and autocorrelation consistent (HAC) and was calculated using the method developed by Kleibergen and Paap (2006).

Table A3: Reduced Form Results

	Numeracy			Literacy		
	(1) P1	(2) P4	(3) P7	(4) P1	(5) P4	(6) P7
$CompLow_{gst}^{pred}$	0.008**	0.012***		0.014***	0.008*	
	(0.003)	(0.004)		(0.004)	(0.004)	
$CompHigh_{gst}^{pred}$		−0.004	−0.008**		−0.003	−0.001
		(0.004)	(0.004)		(0.004)	(0.004)
Observations	190,704	194,804	186,082	190,704	194,804	186,082
No. of Schools	1,437	1,428	1,435	1,437	1,428	1,435
School FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: ∗∗∗/∗∗/∗ indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and year level are reported in parentheses.

This table shows the reduced form results, i.e. the results of an Ordinary Least Squares (OLS) regression in which we regress our outcomes of interest - which are proficiency in numeracy and literacy, respectively - on our instruments - which are class planner predictions of whether a pupil's grade should contribute to a composite class.

Covariates include pupil age, sex, and ethnicity, an indicator for whether pupil is from a neighborhood in bottom 20% of deprivation (SIMD), class size, 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 and school-year fixed effects.

Table A4: Second Stage Results - Fourth (P4) and Seventh (P7) Graders

Panel A: Second Stage Results for P4

	Numeracy				Literacy			
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS	(5) OLS	(6) 2SLS	(7) OLS	(8) 2SLS
Older Peers	0.002***	0.044***			0.002***	0.024		
	(0.001)	(0.017)			(0.001)	(0.017)		
Younger Peers	−0.004***	−0.009			−0.004***	−0.008		
	(0.001)	(0.015)			(0.001)	(0.015)		
Bottom Comp.			0.033***	0.509**			0.027***	0.291
			(0.006)	(0.215)			(0.007)	(0.206)
Top Composite			−0.032***	−0.250			−0.037***	−0.180
			(0.007)	(0.257)			(0.007)	(0.247)
Class Size	0.004***	0.004	0.004***	0.001	0.004***	0.002	0.004***	0.001
	(0.001)	(0.003)	(0.001)	(0.004)	(0.001)	(0.003)	(0.001)	(0.004)
Observations	194,804	194,803	194,804	194,803	194,804	194,803	194,804	194,803
No. of Schools	1428	1,428	1,428	1,428	1,428	1,428	1,428	1,428
Class-Size Instr.	No	Yes	No	Yes	No	Yes	No	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		12.82		4.918		12.82		4.918

Panel B: Second Stage Results for P7

	Numeracy				Literacy			
	(1) OLS	(2) 2SLS	(3) OLS	(4) 2SLS	(5) OLS	(6) 2SLS	(7) OLS	(8) 2SLS
Younger Peers	−0.005***	−0.051			−0.005***	0.008		
	(0.001)	(0.035)			(0.001)	(0.031)		
Top Composite			−0.054***	−0.704			−0.051***	0.114
			(0.007)	(0.527)			(0.007)	(0.425)
Class Size	0.004***	−0.017	0.004***	−0.021	0.004***	0.005	0.004***	0.005
	(0.001)	(0.012)	(0.001)	(0.017)	(0.001)	(0.011)	(0.001)	(0.014)
Observations	186,082	186,078	186,082	186,078	186,082	186,078	186,082	186,078
No. of Schools	1,435	1,435	1,435	1,435	1,435	1,435	1,435	1,435
Class-Size Instr.	No	Yes	No	Yes	No	Yes	No	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat		17.99		12.91		17.99		12.91

Notes: ***/**/* indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and 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. Results in Panel A refer to our sample of fourth graders (P4), results in Panel B refer to seventh graders (P7).

Covariates include pupil age, sex, and ethnicity, an indicator for whether pupil is from a neighborhood in bottom 20% of deprivation (SIMD), 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 and school-year fixed effects.

The reported first-stage F-statistic is heteroscedasticity and autocorrelation consistent (HAC) and was calculated using the method developed by Kleibergen and Paap (2006).

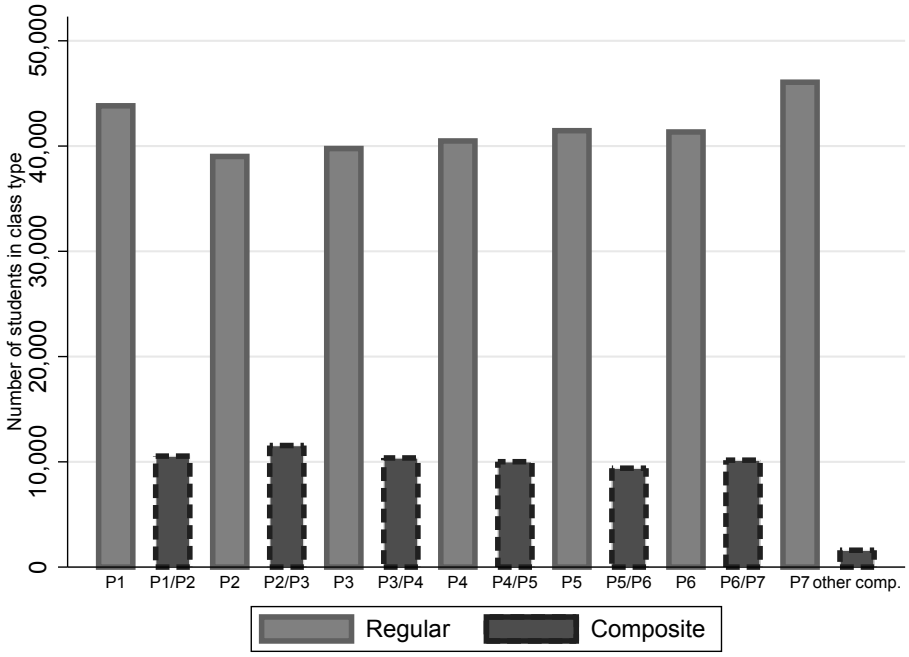
Table A5: Composites and Number of Teachers per Classroom

	First Graders (P1)		Fourth Graders (P4)		Seventh Graders (P7)	
	(1)	(2)	(3)	(4)	(5)	(6)
	>1 Teacher Present		>1 Teacher Present		>1 Teacher Present	
Composite Class (Binary)	−0.008**	−0.008**	0.004**	0.004**	0.008***	0.008***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)
Female		0.000		−0.001**		0.000
		(0.001)		(0.000)		(0.000)
White		0.001		0.001		0.000
		(0.001)		(0.001)		(0.001)
Native English Speaker		0.000		−0.001		−0.001
		(0.001)		(0.001)		(0.001)
Bottom 20% SIMD		−0.001		0.000		−0.000
		(0.001)		(0.000)		(0.000)
Age (in Years)		0.000		0.000		0.001
		(0.002)		(0.001)		(0.000)
% Female in School		0.096		−0.005		0.043
		(0.109)		(0.047)		(0.039)
% White British		0.163		−0.074*		−0.080
		(0.115)		(0.044)		(0.054)
% Native English Speakers		−0.073		0.098**		−0.026
		(0.138)		(0.047)		(0.057)
% in Bottom 20% SIMD		−0.146		−0.081**		0.073
		(0.108)		(0.040)		(0.056)
Number of Students in School		0.000*		0.000***		0.000
		(0.000)		(0.000)		(0.000)
Observations	190,040	190,040	193,627	193,627	185,416	185,416
R-squared	0.424	0.424	0.274	0.276	0.228	0.229
School FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: ∗∗∗/ ∗∗/∗ indicate significance at the 1%/5%/10%-level. Heteroscedasticity-robust standard errors adjusted for clustering at the school and year level are reported in parentheses.

This table shows the results of an OLS regression where the dependent variable a dummy indicator for whether there is more than 1 teacher present in the classroom.

Figure A1: 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

Appendix B: Estimating the Effect of Class Size

Throughout this paper, we control for the effect of class size. By virtue of a lower cap, multi-grade classes tend to be smaller than single-year classes (see Section 2 for maximum class size rules). As a result, they may affect achievement not just through peer effects but also due to a lower pupil to teacher ratio. In order to disentangle these two competing mechanisms, we have included a control variable for class size in all specifications.

While not the primary focus of this paper, the effect of class size is also interesting in itself. However, class size may well be endogenously determined even after controlling for school fixed effects. We therefore also construct an instrument for class size based on enrolment counts in the mould of Angrist and Lavy's (1999) seminal study. Our approach exploits that at an enrolment level just above a maximum class size cut-off, a new class needs to be created.

The key condition in such a regression discontinuity design is that enrolment counts are as good as randomly determined, for instance, by natural fluctuations in birth rates within catchment areas. Bunching in enrolment count on the other hand indicates a violation of the main identifying assumption. Figure B1a shows the enrolment counts for P1 for the school years 2011/12 to 2018/19 when the maximum class size for this grade was 25. We can see obvious sorting at multiples of 25. For instance, there are almost twice as many schools with enrolment counts of exactly 50 than with 51. We can also see bunching in P4 (Figure B1c) and P7 (Figure B1e) at multiples of 33.

Angrist et al. (2019) find similar patterns for their data from Israeli schools. In their case, financial incentives lead to enrolment count manipulation. Israeli schools receive further funding for every additional class that needs to be created. School head teachers selectively use deferment and retention or class skipping, to create enrolment counts that are just large enough to trigger additional classes. In Scotland, the incentives line up exactly in reverse. Scottish head teachers have virtually no discretion over their enrolment counts. Grade retention is also almost unheard of. The sorting that is apparent in Figures B1a, B1c, and

B1e is instead driven by strategic acceptance of placing requests by Councils. As mentioned in the main text (see Section 2), parents can request their children go to schools outside their catchment area, but councils will only grant such placing requests if the requested school has space available. In practice, councils will often accept placing requests for oversubscribed schools up to the point at which the enrolment count is equal to a multiple of the class size limit. Because funding to schools is on a per-class basis (rather than a per-pupil basis), this reshuffling and "filling-up" approach helps councils to cut costs.

Angrist et al. (2019) remedy the bunching issue by calculating an imputed enrolment count that assigns each pupil to the grade in which they should be in, had birthday cut-offs been strictly adhered to. We follow this approach in spirit and assign each pupil to the school they should attend based on the catchment area they reside in. In order to do so, we exploit information on each pupil's postcode area.In other words, we calculate each school's (imputed) enrolment count as if placing requests were not an option. This creates two issues. First, a small set of pupils who are in a school might be there by virtue of a placing request, but we cannot identify them as such because only part of their postcode area overlaps with the catchment area. Second, if we identify a pupil who is attending a school by virtue of a placing request but whose postcode area stretches over multiple catchment areas, it is not obvious against which school's imputed enrolment count such a student should count.

We address both issues by calculating postcode area frequency distributions for all schools in all years, as well as school frequency distributions for each postcode area in all years. If a pupil's postcode area makes up less than 5 percent of her school's pupil population, we re-assign the pupil to her catchment area school[9]. In other words, we assume that students from infrequent postcode areas are in a school due to placing requests. If that same pupil's postcode feeds into two schools, we assign her to the first school with a probability equal to the percentage of pupils from the same postcode area that attend this first school; and to the second school with a probability equal to the percentage of pupils from the same postcode

[9]We also experimented with slightly lower and higher thresholds, the results are qualitatively identical.

area that attend this second school.[10]

Figures B1b, B1d, and B1f show our imputed enrolment counts for P1, P4, and P7 respectively. We can see that our imputed enrolment counts no longer suffer from bunching. All three distributions are smooth and the heaping at multiples of maximum class sizes has disappeared. Appendix Figures B2a to B2f show the corresponding density plots that accompany McCrary's (2008) formal test for sorting. We firmly reject the null hypothesis of no discontinuity for the original enrolment counts but fail to do so for imputed enrolment counts.

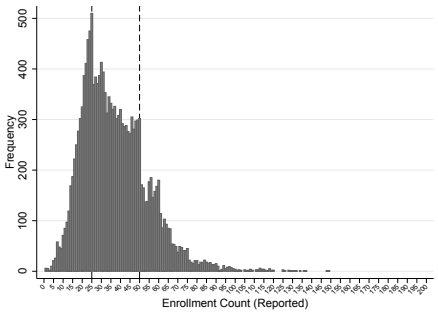
We therefore use the imputed grade enrolment counts rather than the actual enrolment counts to predict the class sizes for class c in grade g in school s and year t as:

$$f_{cgst} = \frac{r_{gst}^{imp}}{int(\frac{r_{gst}^{imp}-1}{cutoff_{gt}}) + 1}$$

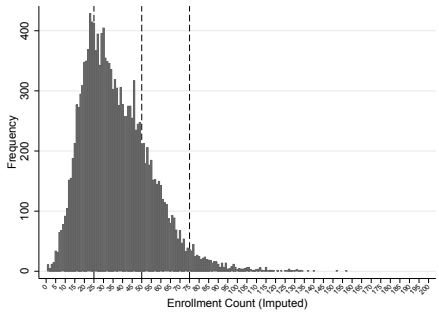
Where r_{gst}^{imp} is the imputed enrolment in school s' g^{th} grade as of September in year t. In both first-stage and second-stage regressions we also flexibly control for r_{gst}^{imp}. Lastly, cutoff$_{gt}$ represents the class size limit, which varies by grade g and school-year t Ultimately, we use the predicted class size f_{cgst} as an instrument for actual class size, $\widehat{CS_{cgst}}$.

[10]Pupils on placing requests are excluded from these probability calculations such that the probabilities add up to 1.

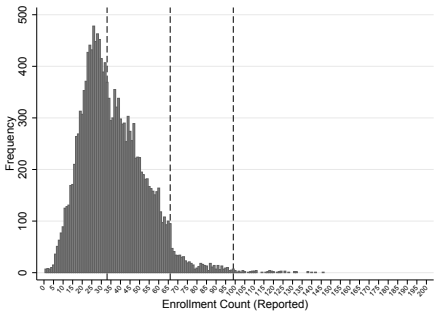
Figure B1: Enrollment Distributions 2007-2018



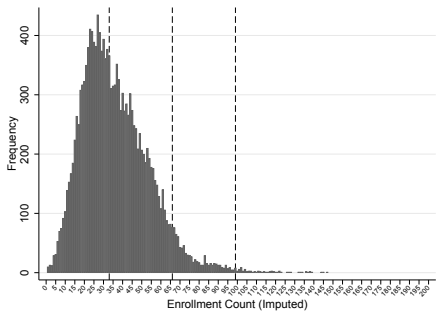
(a) P1 - Reported



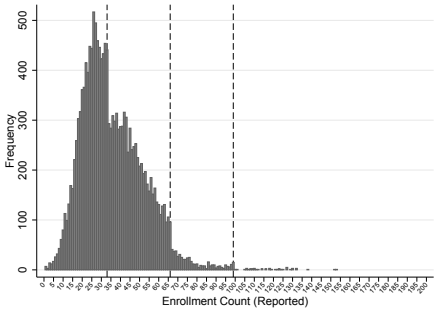
(b) P1 - Imputed



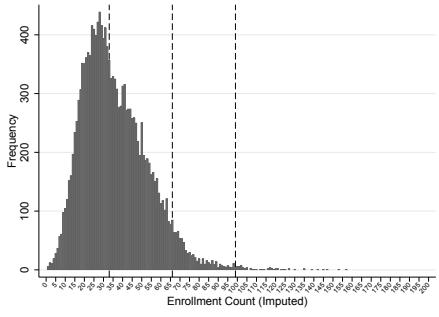
(c) P4 - Reported



(d) P4 - Imputed



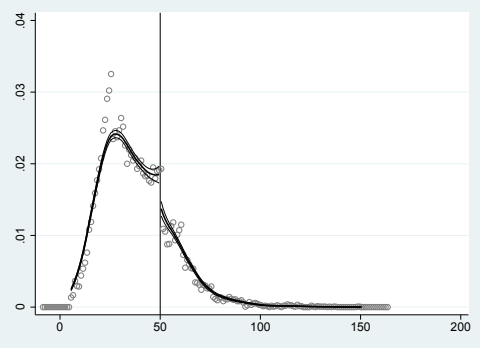
(e) P7 - Reported



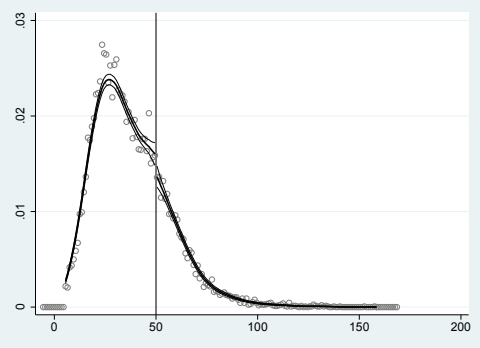
(f) P7 - Imputed

Notes: These figures show the enrolment counts for all schools in our data from 2007/08 to 2018/19, separately for first grade (P1), fourth grade (P4), and seventh grade (P7). On the left are the original enrolment counts which show bunching at multiples of the corresponding maximum class size. On the right are the corresponding imputed enrolment counts in which pupils who we believe are in a school due to placing requests were re-allocated to their catchment area school.

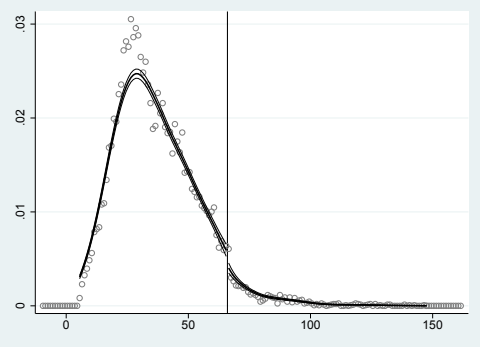
Figure B2: Density Tests - Illustrations



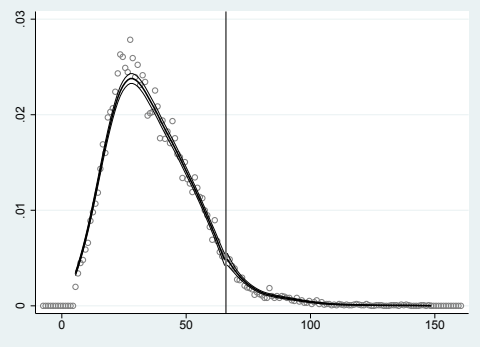
(a) P1 - Reported



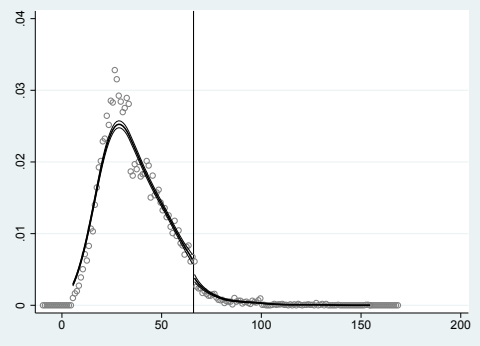
(b) P1 - Imputed



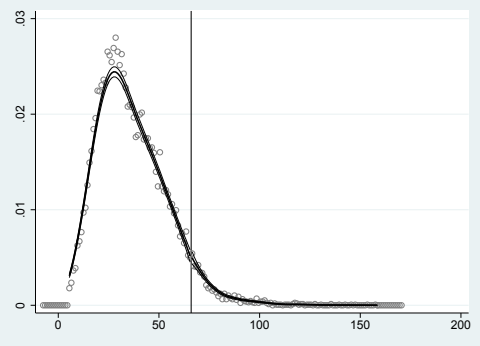
(c) P4 - Reported



(d) P4 - Imputed



(e) P7 - Reported



(f) P7 - Imputed

Notes: These figures show the density in cohort size distributions. Vertical lines are placed at the second multiple of the respective maximum class size thresholds.