Class Composition and Student Achievement. Evidence from Portugal.

João Firmino[a,1,*]; Luís C. Nunes[a,2]; Ana B. Reis[a,3]; Carmo Seabra[a,4]

[a]*Nova School of Business and Economics, Universidade Nova de Lisboa, Campus de Campolide, 1099-032, Lisbon, Portugal*

[1]joao.mgs.firmino@gmail.com; [2]lcnunes@novasbe.pt; [3]ana.balcao.reis@novasbe.pt; [4]carmo.seabra@novasbe.pt

[*] Corresponding author at: Nova School of Business and Economics, Universidade Nova de Lisboa, Campus de Campolide, 1099-032, Lisbon, Portugal. Tel.:+351 919010785

Abstract

Using student-level cross-sectional data of 6th and 9th graders we estimate class composition effects impacting on academic achievement. The richness of the dataset allows to tackle endogeneity stemming from between and within-school non-random sorting of students through the inclusion of many control covariates. We find that increasing the percentage of high achievers, in a 6th grade class, has a negative effect on student performance, while, in a 9th grade class, the effect is positive. Low achieving classmates leads to better performances in 6th grade classes, but to lower performances in 9th grade ones. Students with no past retentions do better with an increasing proportion of this same type of classmates. Larger shares of low-income classmates hurt performance in general. Apart from the past retention dimension where there is evidence supporting tracking students, along all other compositional dimensions it seems fairer that each class reflects the respective school-grade population heterogeneity.

Keywords: class composition; peer effects; student achievement

JEL Classification: I21

1 Introduction

All over the world students spanning elementary to upper secondary schooling are grouped in classes so that teaching can be delivered in an efficient way. It is safe to admit that, in general, the number of students at a given year-school-grade is always larger than the number of available teachers. Whereas the need for grouping students seems an innocuous aspect of any education system, *how* they are grouped is not, at all, a bland topic. With high probability, different agents within the education system will carry different beliefs, interests, and constraints with respect to what might be considered an optimal placement of students across classes. Some parents may prefer to place their children among intellectually gifted classmates while others might prefer to place them in environments where specific cultural and social characteristics are predominant. Principals may have to comply with specific legislative conditions when setting up classes or may try to mirror different personal priors with respect to efficiency and fairness considerations at the school level when forming classes.

One implicit difficulty stemming from the heterogeneity of views of what might constitute an optimal allocation of students across classes is the multitude of dimensions that the composition of a class can be looked from. Indeed, which dimension of class composition is the most important in explaining in-class behavior (and possibly the interactions between classmates outside the class) and, consequently, educational outcomes is, *a priori*, uncertain given the large number of dimensions that may be defined. Examples of such dimensions are classmates' previous attainment (Sund, 2009), gender and race (Hoxby, 2000a), or even language spoken (Yao, Ohinata & Ours, 2016). On top of this, each class compositional dimension may affect educational outcomes heterogeneously, depending on the students' individual characteristics, which adds further complexity to the topic.

In this paper, and in contrast with most of the empirical literature, we analyze, simultaneously, the effects of different class compositional dimensions on individual student cognitive achievement. We make use of a rich dataset that allows us to tackle the major endogeneity concerns when estimating class composition effects, which result from non-random allocation of students between schools and across classes within each school. The dataset allows us to control for many student characteristics, including prior achievement, to measure several class composition dimensions, and also to control for school effects. We are also able to use an instrumental variable that has been used in this literature – the school-grade average class size – for the purpose of estimating an IV model as a robustness check of the OLS specification to tackle eventual contamination bias (this will be detailed below in the methodology section). In addition, the dataset includes information on 6^{th} and 9^{th} grade students so that we can also study differences in the relevance of class composition effects across different ages.

We address the following questions: 1) which dimensions of class composition affect an individual student achievement?; 2) do these effects depend on the students' individual characteristics?; and 3) how do these results differ between 6^{th} and 9^{th} graders?

Our approach has greater value from a public policy perspective since it offers a swift comparison of the direction and magnitude of the effects stemming from several different dimensions of class composition. An evidence based clearer comparison of several different class compositional effects should help on the complex task of *how* to group students across classes.

This work is structured as follows. A literature review is provided in the next section. Section 3 details the dataset and presents some descriptive statistics. Section 4 specifies the econometric methodology. Section 5 presents the estimation results, while Section 6 discusses them. Section 7 summarizes policy implications for schools regarding class formation. Finally, Section 8 concludes.

2 Literature Review

The type of schooling offered to students is influenced by school policies, regarding class formation, in, at least, two ways: how many and what kind of classmates exist in each class. Lazear (2001), shows that, for a given level of students' *quality* (measured as the percentage of time each pays attention to the teacher) increasing class size would exponentially decrease class learning time – more students, more disruptions. But one can similarly argue that for a given class size, increasing the proportion of disruptive students should also hamper overall classmates' learning. These two similar arguments establish, then, an important link between class size and class composition to which we will come back in a moment.

Concerning, specifically, the kind of classmates found in a given class, much attention has been devoted to the case where students are tracked into homogenous classes with respect to their ability or predetermined achievement levels. Sacerdote (2011) indicates that half of the surveyed research points to positive effects from this policy. However, this tends to be less clear when one allows for heterogeneous effects with respect to own students' characteristics. For example, Burke & Sass (2013) provide evidence that different individuals (in terms of their own quality level) seem to benefit differently from being placed in classes with higher shares of top, mid, or low achievers, which adds extra complexity to whom gains more from policies related with class quality homogeneity or heterogeneity.

Our study places the class compositional effects' identification within the education production function framework following, for example, Wößmann & West (2006), Todd & Wolpin (2003), Lazear (2001), Pritchett & Filmer (1999), Hanushek (1979) and Hanushek (1970).[1] This approach may be traced back to the earlier Coleman Report (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, et al., 1966) which, already then, presented the idea that educational outcomes were linked to a set of inputs which included the sort of peers one finds in his/her school. Although it reported that the main predictors of educational outcomes were family and socio-economic background, the student body composition also helped in predicting outcomes (especially those of minorities).

[1] One of the earliest to refer to the educational production function. Hanushek is, then, a staple reference on educational production functions in both their theoretical and empirical usages. See also Hanushek (2008).

The consistent estimation of class composition effects requires endogeneity to be addressed. It arises from possible non-random sampling of students across schools and then across classes, see e.g. Wößmann & West (2006) and Bosworth (2014). Between-school sorting of students may occur if parents are stratified regionally according to a given characteristic (e.g. professional occupation, level of education or income). It causes a potential identification problem because the composition of the classes will then reflect the composition of the school which, in turn, is not independent of factors that determine, themselves, educational outcomes of the students (such as the parents' characteristics). The correlation between school and class composition is not expected to be perfect, nevertheless, since students may be, in turn, sorted across classes, within-schools, in a systematic way. Within-school sorting may take the form of segregating low from high achievers, or segregating whether they have been retained in the past or not, perhaps reflecting different priors from principals or teachers related to how a class should be formed. Consequently, students may experience class compositions that might be predicted by their own characteristics, while these, in turn, are likely to explain their educational outcomes too.

To overcome endogeneity different researchers have resorted to different approaches, conditional on the type of data at hand. Hoxby (2000a) exploits idiosyncratic variations (first differences) of gender and racial compositions in American schools, between adjacent years, due to unanticipated demographic changes, to avoid non-randomness allocation issues. She finds, firstly, that if the cohort average exam score increases, unexpectedly, by 1 point, then a student from that cohort scores more 0.1 to 0.5 points, on average. Secondly, that, in a given cohort, proportionally more females cause better performances in mathematics and reading for both males and females. Finally, that peer effects are stronger and beneficial within racial groups.[2] Hoxby (2000b), finds no significant effects from class size to pupil achievement.[3] Hoxby uses credible unexpected random population variation as instrument for class size while also applying school fixed effects.

There is a stream of literature that makes use of grade-school averages as instruments for class level variables. Akerhielm (1995) initiates such procedure using average class size across a given subject, within a school, to instrument actual class size. Although her procedure does account for within school sorting, it does not take into account between-school sorting. Jürges & Schneider (2004), Wößmann & West (2006) and West & Wößmann (2006), again, employ a two-stage regression procedure to identify class size effects (controlling for within-school sorting with school fixed effects) in the TIMSS' database. They instrument actual class size with the average class size of the respective grade.

On the nature of within-school students' allocation, West & Wößmann (2006) put forward the hypothesis of compensatory sorting. This hypothesis states that the class size reduction treatment comes hand

[2] More recently, Card & Giuliano (2016) through a regression discontinuity design find that high achievers belonging to ethnic minorities benefit from tracking to high achieving classes. Contrary to Hoxby (2000a), they hypothesize that exists negative, not positive, peer pressure that makes top achievers from minorities to underperform in a regular classroom.

[3] Although identification of class size effects is not the main goal of this paper, we review some of its literature since it may contain useful insights towards the identification of class composition effects. Moreover, class size is potentially intimately related to class composition, see Lazear (2001), West & Wößmann (2006), and Bosworth (2014).

in hand with a second treatment of sorting students with weak achievement-related inputs precisely to the classes of reduced dimensions. That is, students benefiting from less populated classes also experience, in general, a higher proportion of disadvantaged classmates (besides tending themselves to be disadvantaged students). They point that countries with external exams are prone to induce such within-school compensatory schemes.[4] The pressure to minimize the number of students with negative exam scores may dictate to a certain extent how classes are formed every year.

From the discussion so far, we acknowledge two points. First, the importance of including class size in the upcoming econometric specifications, at least as a control variable. One needs to hold constant the class size "treatment" when interpreting the class composition "treatments". Or in other words, controlling for class size is required for a *ceteris paribus* interpretation of the compositional coefficients (see Bosworth, 2014). Second, across econometric specifications that jointly include class compositional measures and class size we may regard this last one as the potential endogenous variable. This should be the case if those who are in charge, within each school, of class formation *believe* that class size reduction is a stronger compensatory policy than specific class compositions. We provide suggestive evidence in favor of the class size compensatory hypothesis at the end of Section 3.

More recent literature still points to gains from having homogeneous classes according to past student performance, see Collins & Gan (2013). Duflo, Dupas & Kremer (2011) in a randomized experiment in Kenyan schools also report positive peer effects to the achievement levels of any type of student from the presence of high achievers in class. They, interestingly, further observe that sorting students to homogenous classes with respect to their initial levels of achievement caused all types of students to perform better. They explain that low achievers, although deprived of the potential contributions of the high achieving peers, might have benefited from better tailored teaching.

Finally, some authors, e.g. Hanushek, Kain, Markman & Rivkin (2003), Sund (2009) and Burke & Sass (2013), have been able to analyze longitudinal student-level datasets with comparable students to the ones we analyze in this paper, in terms of age and grade. The panel structure of the data allows them to control for several unobserved heterogeneities by including (separately or jointly) student, teacher, and school-by-grade fixed effects or other combinations of these in their specifications. Hanushek et al. (2003) and Sund (2009) point to gains in achievement by the average student from having peers with higher levels of prior mean achievement. Burke & Sass (2013), in turn, point to gains for a given student in having better, but not too better, peers. They argue that a too large difference of prior achievement may hamper communication between students.

[4] Which is the case of Portugal, in the period studied, with its high-stakes national exams.

3 Data and Descriptive Statistics

This paper makes use of an administrative dataset maintained by the Portuguese Ministry of Education.[5] We use information on all students enrolled in public schools (only a minority is enrolled in private schools), in continental Portugal, from grades 6 and 9 in the academic year of 2011-12. The dataset provides information on students' class and school membership, their scores by subject and by type of examination, and their academic track. National exams' scores (high-stakes exams) of mathematics and reading taken at the end of the 2011-12 academic year provide the achievement measure. We also have information on previous achievement: a baseline score from a low-stakes national exam.[6]

The dataset also includes several demographic variables characterizing each individual pupil, such as gender, parents' education (we use the education level of the parent with the highest degree), and home access to internet. Using the students' birthdate, we created a dummy variable taking the value one if their age (at the beginning of the academic year – mid September) was equal or lower than the reference age for the grade the student is enrolled in (the reference being the *maximum* age a student is expected to have, at the beginning of the academic year, *without* having repeated any grade in past academic years). For 6th grade students the reference age is 12 years-old, while for 9th graders it is 15 years-old. Cultural background is proxied by a dummy variable taking the value one if the student is a foreigner, i.e. if the student was born in one of the Portuguese speaking countries excluding Portugal and zero if born in Portugal.[7] Low-income students were flagged if they received social support (equal to one if they did, zero otherwise).

Our variables of interest, related to the composition of the classes, were created in the following manner. Knowing, for each pupil, his/her class membership, we were able to compute, at the class level, the percentage of: males, pupils with home access to internet, pupils below (or at) the reference age, pupils born in a foreign country, and low-income pupils. On top of these compositional measures we further compute, for each class, the percentage of high and low achievers. We define as high achiever a student with a baseline score of 5 and as low achiever one with a baseline score of 1 or 2 (implicitly those with baseline scores of 3 or 4 compose the middle achievers). These groups, high and low achievers, vary from 3.3% to 17.2% of the respective relevant population, depending on subject and grade (see Appendix A.1). We stress that *all* class compositional measures are "leave-out-percentages" as we excluded the contribution of student *i* when

[5] We thank DGEEC for providing access to the anonymized version of the MISI administrative dataset.

[6] The scores in the high stakes national exams (the dependent variable) is recorded on a scale from 0 to 100 points. Scores from 0 to 49 correspond to a negative evaluation while scores from 50 to 100 correspond to a positive evaluation. The baseline scores are measured on a scale from 1 to 5 points, where 1 and 2 points correspond to a negative evaluation and 3, 4, and 5 points correspond to a positive evaluation. For 9th graders the baseline score refers to their 6th grade low stakes exam score, whereas for 6th graders it refers to their 4th grade low stakes exam score. For students that repeated the 4th or the 6th grade, we only use their latest score, i.e. the one immediately before they progress to the next grade.

[7] The Portuguese speaking countries (excluding Portugal) are: Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, Sao Tome and Principe, and East Timor. The baseline case includes therefore students born in Portugal or in countries not belonging to Portuguese speaking countries. A third category differentiating those students not born neither in Portugal nor in a Portuguese speaking country would be of very small size and too heterogeneous.

computing them.[8] Thus, one should interpret them as the percentage of classmates, of individual *i*, with a given characteristic, beyond individual *i* itself. Finally, we created a measure of class age dispersion[9] and counted each student's class size.

The dataset includes students enrolled in classes of extremely reduced dimensions in relation to what was stipulated by law: a minimum and a maximum of 24 and 28, respectively. Although the law allowed for exceptional cases (e.g. to group pupils that would have overflown the limits of the remaining classes), we only considered classes with at least 14 students. The number of students left out using this threshold is marginal.[10]

We note that the Portuguese educational system is divided into cycles.[11] And within each cycle the composition of a class is typically kept unchanged by schools. In particular, in the 2nd cycle, students are normally allocated to the same class in grades 5 and 6. And in the 3rd cycle, the class composition is normally kept unchanged from grades 7 to 9. A reorganization of the classes may take place when students, simultaneously, move from the 6th to the 7th grade and from a school with grades 5 to 9 to a school with grades 7 to 12.[12]

The sample is also restricted to classes of pupils enrolled under the regular academic track. This is the majority of students in continental Portugal. DGEEC (2013, page 28) reports that in 2011-12, the percentage of students in public schools enrolled in the regular academic track in the 2nd and 3rd cycles were 99% and 90%, respectively.

Finally, we only use students enrolled in schools with, at least, one class of both 6th and 9th grades (which implies schools with at least two classes). This requirement ensures, firstly, that school fixed effects can be estimated because each school will contribute with at least two classes to the estimation sample and, secondly, that the 1st stage of the instrumental variable robustness check is also estimable (more details on this on Section 4).

---

[8] Moreover, there are missing values with respect to baseline scores and place of birth for some students. We compute these leave-out percentages just using the non-missing information from a given class, but for classes with up to 3 missing values only. Students in classes with 4 or more missing values were dropped. Given a minimum class size of 14 it is unlikely that disregarding the contribution of up to 3 classmates would dramatically bias the true compositional measures for the classmates of those particular classes. Nevertheless, to control for such an effect, we considered three dummy variables which take the value of one if the pupil belongs to a class for which we bypassed missing information of 1, 2, or 3 classmates, respectively.

[9] That measure is the mean absolute deviation of the classmates' age to their class average age: $\text{Class Age Dispersion}_{ij} = \frac{\sum_{i \in j} |Age_{ij} - Class\ Average\ Age_{ij}|}{Class\ Size_{ij}}$ with *i* indexing individual students and *j* the class they belong to. When this variable takes the value, say, 0.5, then it means that, for a given class, each classmate's age is, on average and in absolute value, half year away from that class average age.

[10] We also assume that students that left the school during the first third of the academic year were not there from the beginning. Although this artificially shrinks class size it tackles the problem that stayers could only have been peer affected by leavers a small portion of the whole academic year.

[11] The 1st cycle contains grades 1 to 4; the 2nd cycle grades 5 to 6; the 3rd cycle grades 7 to 9 and the secondary grades 10 to 12.

[12] There are four main types of public schools in Portugal: elementary schools with grades 1 to 4 (1st cycle); schools with grades 5 to 9 (2nd and 3rd cycles); schools with grades 7 to 12 (3rd cycle and secondary) and schools with grades 10 to 12 (secondary).

The sample of students that could be used in our econometric analysis includes 59 thousand students in the 6th grade and 38 thousand in the 9th grade, corresponding to 56.5% and 44.0% of the students enrolled in the regular academic track, respectively, in continental Portugal´s public schools, in the academic year 2011-12.[13]

The fact that roughly half of the original population of students, of each grade, makes its way to the final estimation sample is explained by the list of restrictions enumerated in this section that we imposed on the dataset and also by missing values. The latter could introduce undesirable sample selection bias, but we argue that it should not be as large as one could, *a priori*, expect it to be. There are three main different levels at which a missing value may have been generated: at the central, school, and individual level. The latter level is mostly related with the fact that there is a non-marginal amount of missing values regarding parents' education. This may happen either because students do not know the relevant information or because parents miss to report it to the school. It is conceivable that students from more disadvantaged backgrounds may tend to be the ones that do not know such characteristics about their parents, or, more likely, that their parents tend to show up less in school and are less likely to report the required information. This could lead to an under-representation of that kind of students in our regression samples and, in the limit, to an under-representation of classes majorly composed by them. We note that the inclusion of the set of individual level regressors should control, to a large extent, for the likelihood of belonging to the regression sample. It is reasonable to suppose that higher baseline scores, not belonging to a low-income family, or even higher levels of parental education are positively correlated with the probability of disclosing all required information to the school, thus positively correlated to the likelihood of belonging to the regression sample. It is, thus, the set of coefficients associated to individual level controls the one we expect to pick the eventual sample selection bias and not the set of coefficients of the different class composition variables (the relevant ones in this paper). At the school level it may happen that the administrative services of the schools fail to export, to the central authorities, information without typos due to random daily typing incidents. Nevertheless, and irrespectively of its cause, the missing information originated at this level should be seen as school specific, thus within the school fixed effects. Lastly, we do not conceive as reasonable that losses of information that may have occurred at the central level may be systematically related to students' characteristics or, more important to us, to class composition.

Some descriptive statistics of the final dataset are presented in Table 1.[14] Regarding the statistics on class level variables note that the relevant number of observations, in this context, is the number of classes, not of students (hence the different number of observations for these variables). The distributions of the class level variables are shown in Appendix A.3.

[13] These figures refer to the number of students for whom there is a full set of information across all individual and class level variables and respect all necessary requirements enumerated in this Section. Percentages out of the totals 104 410 and 86 416 for 6th and 9th grades, respectively, DGEEC (2013, pages 68 and 72).

[14] These statistics refer to students with a mathematics national exam score. Appendix A.2 presents the same statistics for those with a reading national exam score. As expected, both populations are very similar given they differ by just a few hundreds.

Table 1. Descriptive statistics.

		6th Grade - Mathematics National Exam					9th Grade - Mathematics National Exam				
		N	Mean	Std.Dev.	Min	Max	N	Mean	Std.Dev.	Min	Max
Individual Level Variables	Score	59,359	53.2	23.0	0	100	38,046	53.0	23.5	0	100
	Baseline Score	59,359	2.8	1.0	1	5	38,046	2.9	1.1	1	5
	Reference Age	59,359	0.88	0.33	0	1	38,046	0.84	0.37	0	1
	Male	59,359	0.51	0.50	0	1	38,046	0.48	0.50	0	1
	Foreigner	59,359	0.01	0.12	0	1	38,046	0.01	0.12	0	1
	Internet	59,359	0.60	0.49	0	1	38,046	0.72	0.45	0	1
	Low-Income	59,359	0.45	0.50	0	1	38,046	0.39	0.49	0	1
	Tertiary Ed. (Parent)	59,359	0.18	0.38	0	1	38,046	0.17	0.37	0	1
	Secondary Ed. (Parent)	59,359	0.48	0.50	0	1	38,046	0.46	0.50	0	1
Class Level Variables	% High Achievers	3,552	14	12	0	81	2,334	6	7	0	45
	% Low Achievers	3,552	9	8	0	59	2,334	10	9	0	56
	% Reference Age	3,552	72	16	0	100	2,334	69	17	0	100
	% Males	3,552	43	12	0	79	2,334	40	13	0	80
	% Foreigners	3,552	2	3	0	29	2,334	2	4	0	33
	% Internet	3,552	48	23	0	100	2,334	59	24	0	100
	% Low-Income	3,552	39	17	0	95	2,334	34	17	0	95
	Age Dispersion	3,552	0.6	0.2	0.2	1.7	2,334	0.5	0.2	0.2	1.3
	Class Size	3,552	23	3	14	31	2,334	21	4	14	30

We end this section observing suggestive evidence supporting the West & Wößmann (2006) hypothesis of compensatory within-school sorting along the class size dimension (see discussion in Section 2). The statistically significant sample correlations between class size and the class percentage of: high achievers, low achievers, below reference age students, and low-income students are, respectively, 0.20, -0.14, 0.19 and -0.15. That is, we tend to find students with the, *a priori*, weakest achievement-related inputs (i.e. those in need of compensating inputs) in smaller classes. On the other hand, the statistically significant sample correlation between the class percentages of low and high achievers is -0.39. We would expect this correlation to have the opposite sign if any within-school compensatory sorting, other than through class size, say, through class "quality", was in place. In that case low achievers would be grouped in classes with high achievers at the cost of the presence of middle achievers, thus reversing the sign of the latter correlation.

4 Econometric Methodology

The benchmark model to estimate class composition effects includes the variables detailed in Section 3 and assumes that the marginal effects of the compositional variables do not depend on students' individual characteristics and are not grade specific. It is given by:

$$Y_i = \beta_0 + \boldsymbol{Comp}'_{i,(-i)}\boldsymbol{\beta_1} + \boldsymbol{C}'_i\boldsymbol{\beta_2} + \boldsymbol{X}'_i\boldsymbol{\beta_3} + \gamma G_i + \boldsymbol{S}'_i\boldsymbol{\alpha} + \varepsilon_i \qquad (1)$$

where Y_i is the standardized mathematics or reading national exam score of student *i*; $\boldsymbol{Comp}_{i,(-i)}$, $\boldsymbol{C_i}$, $\boldsymbol{X_i}$, G_i, and $\boldsymbol{S_i}$ are the explanatory variables; ε_i is the student *i* idiosyncratic error term, and β_0, $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_3}$, γ, and $\boldsymbol{\alpha}$ are parameters to be estimated. Although the explanatory variables are indexed at the student *i* level, they are calculated at specific individual, class, or school levels as explained next.

$\boldsymbol{Comp}_{i,(-i)}$ is the vector containing the percentages of classmates of student *i* with a given characteristic – high achievers, low achievers, below the reference age, males, born in a foreign country, home access to internet, and low-income. These are the class compositional variables of interest for this paper. Recall from Section 3 that these compositional measures were computed in a *leave-out* fashion and based on *predetermined* characteristics of the students. The former enables them to be interpret as peer measures, while the latter makes one avoid reflexivity bias.[15]

The $\boldsymbol{S_i}$ vector includes school dummy variables, i.e. school fixed effects. As discussed in Section 2, consistent estimation of true $\boldsymbol{\beta_1}$ requires that we credibly control for endogeneity stemming from non-random allocation of students between and within schools. One popular way in the literature to control for between-school sorting is to include school fixed effects. The idea is to control for anything that is school specific, namely the type and number of students that it attracts, but also school specific policies with respect to class formation. In turn, these and other school specificities may well impact student outcomes making their omission a source of bias. Including school fixed effects contributes to interpret the estimated $\boldsymbol{\beta_1}$ as if students had been randomly allocated across schools. Further, school fixed effects can also be seen as a step to control for teacher between-school sorting under similar arguments.

Next, $\boldsymbol{X_i}$ is the vector containing all the individual level characteristics of the students presented in Section 3[16] and we regard them as important controls with respect to within-school sorting. As discussed in the literature review – that educational systems with external exams, such as Portugal, induce compensatory policies within the schools (i.e. within-school sorting of students) – and given the suggestive evidence presented in Section 3 – that students with weaker inputs tend to be found alongside each other (in smaller classes) – we consider that, indeed, those authorities are likely to have taken into account students' characteristics during the class formation process, at least to some extent. On top of this, the information that school authorities have, at the moment of class formation, is mostly based on the dataset we actually use in this study. This means that if some sort of purposeful within-school sorting took place based on student level information, then it was a function of students' characteristics observable to us.[17] Moreover, we stress the inclusion, within this vector, of the baseline score as a control variable. Todd & Wolpin (2003) and Hanushek & Rivkin (2010) provide theoretical frameworks that justify the use of a baseline score as a summary of past factors under some technical assumptions.[18] Its inclusion, then, allows to control for possible correlations between contemporary class assignment and those past factors. Thus, conditional on the set of individual students' characteristics, one is closer to interpret the estimated $\boldsymbol{\beta_1}$ as if students had been randomly allocated across classes within schools.[19]

---

[15] Which could be the case if, for example, the percentages of high or low achievers had been defined through the outcome score and not by the predetermined baseline score as we did. In that case we would then have the outcome of student *i* influencing his/her classmates' outcome and vice-versa, hence creating reflection bias, see Hanushek et al. (2003).

[16] Dummy variables describing each student baseline score, parent education, reference age status, gender, foreign background status, internet at home status and low-income status.

[17] Students are not interviewed nor tested by the school prior to admission, hence there is no private information to schools when forming classes that could be unobservable to us.

[18] We assume these technical conditions hold, especially that (past) input coefficients are geometrically declining with the distance to the age of the measurement of the outcome.

[19] Inclusion of parent education, besides capturing the direct effect of parents' education input on students' achievement, should help to control for both between and within school sorting since it may be positively related with

The term $\boldsymbol{C_i}$ is a third vector containing other class level control variables, such as class size, age dispersion in the class, and dummies flagging classes with 1, 2, or 3 classmates whose information was not used when computing the class compositional variables of interest because of missing data.[20] As discussed in the literature review, one needs to control for class size as this is a possible confounding treatment that seems to go hand in hand with the class compositional treatments. Moreover, its inclusion contributes to a clearer *ceteris paribus* interpretation of the relevant coefficients. Inclusion of the dummies related with missing information at the class level aims at controlling for any unobserved features of such classes that might explain why that information is missing. If those unobserved features somehow relate with student outcomes then not controlling for them could bias the estimate of $\boldsymbol{\beta_1}$. And, lastly, controlling for class age dispersion means that one is closer to interpret the marginal effects of the different class compositions as if behaviors stemming from dispersion of ages at the class level were fixed. So, for example, the marginal effect of a change in the percentage of below reference age students should not capture the peer effects stemming from changing the class age structure (which is being held fixed) but peer effects stemming from a change in the structure of peers' past academic experience.[21]

Finally, G_i is a grade dummy variable (note that equation (1) is estimated using a pooled sample of students of the 6th and 9th grades) which accounts for eventual grade specific features correlated with both the students' outcomes and the composition of the classes of a given grade.

To allow the marginal effects of the class compositional variables to be heterogeneous according to students' individual characteristics we consider an alternative model allowing for interactions between each class composition variable and the corresponding individual characteristic. More precisely, we specify the model as:

$$\begin{aligned}
Y_i = \beta_0 + {} & \left[\%High\ Achievers_{i,(-i)} * \mathbb{I}_{Low\ Achiever_i=1}\right]\beta_1^{Low} + \\
& \left[\%High\ Achievers_{i,(-i)} * \mathbb{I}_{Mid\ Achiever_i=1}\right]\beta_1^{Mid} + \\
& \left[\%High\ Achievers_{i,(-i)} * \mathbb{I}_{High\ Achiever_i=1}\right]\beta_1^{High} + \\
& \left[\%Below\ Reference\ Age_{i,(-i)} * \mathbb{I}_{Below\ Reference\ Age_i=1}\right]\beta_1^{Below} + \\
& \left[\%Below\ Reference\ Age_{i,(-i)} * \mathbb{I}_{Above\ Reference\ Age_i=1}\right]\beta_1^{Above} + \\
& \left[\%Males_{i,(-i)} * \mathbb{I}_{Male_i=1}\right]\beta_1^{Male} + \\
& \left[\%Males_{i,(-i)} * \mathbb{I}_{Female_i=1}\right]\beta_1^{Female} + \\
& \left[\%Foreigners_{i,(-i)} * \mathbb{I}_{Foreigner_i=1}\right]\beta_1^{Foreigner} +
\end{aligned}$$

---

parents' ability and willingness to place their children on the "best" school and then to informally bargain within the school for the "best" class.

[20] See Section 3 for details about these control dummies.

[21] As the percentage of below reference age students increases, also, on average, the class age dispersion decreases. Nevertheless, this is not a one-to-one relationship. Some classes have a relatively high age dispersion but with few above reference age students (classes with students that entered the education system really early – 5 years old – alongside others that entered it aged almost 7 years-old and both groups never experiencing retention). And, on the other hand, classes with many students above the reference age whose ages are aligned, i.e. with a relatively small class age dispersion.

$$\left[\%Foreigners_{i,(-i)} * \mathbb{I}_{Non\ Foreigner_i=1}\right]\beta_1^{Non\ Foreigner} +$$
$$\left[\%Internet_{i,(-i)} * \mathbb{I}_{Internet_i=1}\right]\beta_1^{Internet} +$$
$$\left[\%Internet_{i,(-i)} * \mathbb{I}_{No\ Internet_i=1}\right]\beta_1^{No\ Internet} +$$
$$\left[\%Low\ Income_{i,(-i)} * \mathbb{I}_{Low\ Income_i=1}\right]\beta_1^{Low\ Income} +$$
$$\left[\%Low\ Income_{i,(-i)} * \mathbb{I}_{Non\ Low\ Income_i=1}\right]\beta_1^{Non\ Low\ Income} +$$
$$\boldsymbol{C_i'\beta_2} + \boldsymbol{X_i'\beta_3} + \gamma G_i + \boldsymbol{S_i'\alpha} + \varepsilon_i$$

or more compactly by:

$$Y_i = \beta_0 + \left[\boldsymbol{Comp}'_{\boldsymbol{i},(-\boldsymbol{i})} * \mathbb{I}_{\boldsymbol{X_i=1}}\right]\boldsymbol{\beta_1^{X_i=1}} + \left[\boldsymbol{Comp}'_{\boldsymbol{i},(-\boldsymbol{i})} * \mathbb{I}_{\boldsymbol{X_i=0}}\right]\boldsymbol{\beta_1^{X_i=0}} +$$
$$\boldsymbol{C_i'\beta_2} + \boldsymbol{X_i'\beta_3} + \gamma G_i + \boldsymbol{S_i'\alpha} + \varepsilon_i \qquad (2)$$

where $\mathbb{I}$ denotes an indicator function. That is, we assess how a (leave-out) percentage of a given type of student in a class (i.e. $\boldsymbol{Comp_{i,(-i)}}$) affects the student of that type (when $\mathbb{I}_{\boldsymbol{X_i=1}}$ takes the value one) and the student of the *other* type (when $\mathbb{I}_{\boldsymbol{X_i=0}}$ takes the value one). For example, in this model, the in-class (leave-out) percentage of males may impact differently on male ($\mathbb{I}_{Male_i=1}$ takes the value one) or female ($\mathbb{I}_{Female_i=1}$ takes the value one) students.

The models in equations (1) and (2) rely on numerous control variables offered by the dataset to reduce possible endogeneity biases that could plague the estimates of the true class compositional effects. However, given what was argued in Section 2 and given the suggestive evidence shown in Section 3, we consider there is the chance for class size to be the within-school driver of endogeneity. In such case, an ordinary least squares (OLS) estimation of the previous models may still result in biased estimates despite the many controls used in the regressions. Hence, we also estimate the model in equation (2) using an instrumental variables (IV) approach. The class size is instrumented with school-grade average class size following Jürges & Schneider (2004), Wößmann & West (2006) and West & Wößmann (2006). We employ this IV estimation as a robustness check. If the class compositional variables correlate with the possibly endogenous class size variable (which they somewhat do) then the possible endogeneity of the latter may affect the correct estimation of the effects of the formers.

Regarding the two necessary requirements for this instrument to hold valid, we can say that, on one hand, grade-school average class size must be correlated with the actual class sizes that compose that grade. After all, even though schools may sort weaker students to shorter classes in a given grade, it must be the case that schools that have relatively more students must sort them to shorter classes that are relatively more populated than shorter classes of less populated grades-schools. Given that schools must obey certain national level rules regarding class formation and face, at each academic year, at each grade, specific cohorts with a given size, then, *conditional* on additive grade and school fixed effects (the latter capturing school specific rules or resources correlated with the number of classes opened and school specific levels of enrolment) variations on the grade-school average class size should reflect exogenous demographic variations of the respective cohort in the area of influence of the school. On the other hand, as it is put by

Wößmann & West (2006), the exclusion condition of the instrument is likely to hold too: "*There is also no reason to expect that the average class size would affect the performance of students in a specific class in any other way than through its effect on the actual size of the class of the students.*" (p. 700).

The first stage of the IV estimation consists in estimating the following equation that predicts CS_i, the class size for each individual *i*:

$$CS_i = b_0 + \left[\boldsymbol{Comp}'_{\boldsymbol{i},(-\boldsymbol{i})} * \mathbb{I}_{\boldsymbol{X_i=1}}\right]\boldsymbol{b}_{\boldsymbol{1}}^{\boldsymbol{X_i=1}} + \left[\boldsymbol{Comp}'_{\boldsymbol{i},(-\boldsymbol{i})} * \mathbb{I}_{\boldsymbol{X_i=0}}\right]\boldsymbol{b}_{\boldsymbol{1}}^{\boldsymbol{X_i=0}} + \boldsymbol{C}'_{\boldsymbol{i}}\boldsymbol{b_2} + b_{\overline{CS}}\overline{CS}_i + \boldsymbol{X}'_i\boldsymbol{b_3} + gG_i + \boldsymbol{S}'_i\boldsymbol{a} + e_i \qquad (3)$$

with $\overline{CS}_i$ representing the (excluded) instrument – student's *i* school-grade average class size. Note that to avoid perfect collinearity, between the instrument and the school fixed effects in this first stage regression, one has to pool both 6th and 9th grades' samples and make sure that each school "contributes" with, at least, two classes – one class from each of the two grades. As discussed in Section 3, our sample satisfies this requirement.

A comparison of the OLS and IV versions of model (2), detailed in the next section, will provide evidence of no significant differences of the class compositional coefficients between them. We take this as evidence that the OLS version delivers estimates of the class compositional effects free of endogeneity contamination bias stemming from the possibly endogenous class size variable. We, thus, proceed with the OLS estimation of model (2) for each sample of 6th and 9th graders, separately, with the intention of disentangling possible differences between class compositional effects at different ages.[22]

In the next section we present the estimation results of the models that pool both grades' samples and of the models that separate each sample of 6th and 9th graders.

5 Estimation Results

5.1 Results for pooled 6th and 9th grades' models

Table 2 provides, for both measures of achievement, the OLS results of model in equation (1) above. Column (1) presents the results of a simple specification of model (1), where only student and teacher between-school sorting is controlled for (via school fixed effects), and only two class compositional variables – (leave-out) percentage of high and low achievers – are included along with class size. These two compositional measures are the ones that relate more closely to what the literature defines as class student "quality". Then, column (2) adds the full set of individual student level controls which aim to account for student within-school sorting. The final column (3) adds the remaining class compositional measures which further alleviate possible confounding treatment effects, thus improving the ceteris paribus interpretation of each of the estimated compositional coefficients.

Indeed, the coefficients attached to the proportion of high and low achievers suffer significant changes (both in sign and magnitude) as each set of control and treatment variables are consecutively added. In turn, the estimated effect of class size changes from statistically significant and positive to non-significant.

[22] The grade fixed effect is taken out from these grade specific models to avoid collinearity with the intercept.

Table 2 *Estimation results using as outcomes the national exam scores in Mathematics and Reading*

Explanatory Variables	Pooled Grades					
	OLS					
	(1)		(2)		(3)	
	Mathematics	Reading	Mathematics	Reading	Mathematics	Reading
% High Achievers	0.0080***	0.0110***	-0.0026***	0.0005	-0.0045***	-0.0015***
% Low Achievers	-0.0078***	-0.0073***	-0.0005	-0.0014**	0.0013**	0.0003
% Below Reference Age	--	--	--	--	0.0027***	0.0021***
% Males	--	--	--	--	0.0001	-0.0005
% Foreigners	--	--	--	--	-0.0010	-0.0031**
% Internet	--	--	--	--	0.0009***	0.0007***
% Low-Income	--	--	--	--	-0.0033***	-0.0023***
Age Dispersion	--	--	--	--	-0.087***	-0.050*
Class Size	0.017***	0.017***	0.006***	0.004***	0.002	0.001
Below Reference Age	--	--	0.42***	0.37***	0.41***	0.37***
Male	--	--	-0.08***	-0.22***	-0.08***	-0.22***
Foreigner	--	--	-0.05**	-0.06***	-0.05***	-0.06***
Internet	--	--	0.10***	0.08***	0.10***	0.08***
Low-Income	--	--	-0.14***	-0.11***	-0.14***	-0.10***
Baseline Score Dummies	--	--	✓	✓	✓	✓
Parent Education Dummies	--	--	✓	✓	✓	✓
Grade Fixed Effects	✓	✓	✓	✓	✓	✓
School Fixed Effects	✓	✓	✓	✓	✓	✓
Adjusted R2	9.8%	7.1%	49.7%	43.9%	50.0%	44.1%
N	112,417	111,961	99,899	99,562	97,405	97,024

Notes: Significance levels: * p<.10, ** p<.05, *** p<.01. Robust standard errors clustered at the class level. Each outcome variable was standardized to have mean zero and std. dev. of one. Each model also contains dummies equal to 1 if student *i* peers' measures were computed using partial class information, i.e. if one, two or three classmates of *i* had missing information about their baseline scores or their place of birth. The class composition variables (i.e. the percentages of classmates of student *i* with a given characteristic) were computed in a leave-out fashion, i.e. excluding student *i*. Each model contains an intercept and pools students from grades 6 and 9. Only classes with 14 or more students were used and they had to belong to schools with at least one class of grade 6 and, simultaneously, another of grade 9 (i.e. each school contributed with at least two classes).

When we allow for heterogeneous class compositional marginal effects, corresponding to the model in equation (2), Table 3, column 1, the picture does not change considerably, regarding the non-significant effect of class size.[23] Overall, these changes demonstrate the importance of controlling for within-school sorting of students and for possible confounding treatment effects, which improve the causal interpretation of the class compositional effects.

In spite of the fact that we do not observe, in line with Hoxby (2000b), a statistically significant negative effect for class size even if it may be structurally small (Bosworth, 2014) or require large class size variations to be captured (Duflo, Dupas & Kremer, 2015), it may also mean that the class size coefficient may still be plagued by endogeneity bias. Recall that, as discussed in Section 4, this may cause the class compositional coefficients – the ones of interest for this paper – to also be biased, which, coupled with the suspicion that class size is the within-school driver of endogeneity, justifies our choice to be conservative. The use of the IV estimator to estimate equation (2) provides then a robustness check against this possibility. Before presenting the second stage estimation results, we discuss next two points worth of mention regarding the results of the first stage of the corresponding 2SLS procedure (see Appendix A.4).

[23] One would expect the class size effect to be detectable once sorting of students and confounding class level treatments are taken into account. Nevertheless, we do not observe it. In reality, the class size effect may not exist or if it does exist it may simply be small in magnitude and difficult to detect. For example, it may be the case that, contrary to Lazear (2001), students tend to disrupt classes with some degree of synchronization, thus diminishing its potential negative effects. Or it may be necessary to record greater class size variation in absolute terms (greater than from 14 to 31 pupils as we observe in our data) to capture a significant effect, like Duflo, Dupas & Kremer (2015) do (they observe classes halving from 80 to 40 students).

First, many of the individual and class compositional variables statistically significantly predict class size, at least at the 5% level. This reinforces the idea that, indeed, the class size each student experiences might be determined by his/her own characteristics (via purposeful within-school sorting). And it also confirms that at least some of the class compositional treatments are, in fact, related to class size, hence it is advisable to jointly include them in the educational production function.

Secondly, and more importantly, the (excluded) instrument – grade-school average class size – statistically significantly predicts class size (at the 1% level) even after "partialling-out" what the other regressors can explain about the variation of class size. Especially after what grade and school effects explain about class size, which means that the instrument should account for grade by school specificities like exogenous cohort variation of 6th and 9th graders around the influence area of the school. In turn, the usual first stage F-statistic testing the significance of the excluded instruments (which here is just one) is well above the rule of thumb of 10, hence the instrument is not weak.

The IV estimation results, which are the second stage results of the 2SLS procedure, are shown in column (2) of Table 3. The signs and magnitudes of the estimated coefficients are quite similar to those obtained by OLS in column (1). The exceptions are mild changes in the coefficients of class size and class age dispersion for the reading specification which turn negative and significant at the 5% and 10% levels, respectively. The change of the class size coefficient is expectable assuming that class size may still be plagued, to some degree, by endogeneity and the instrument is valid.[24] We conduct a Durbin-Wu-Hausman test to formally assess if, under the hypothesis that the instrument is valid, there is evidence of endogeneity, i.e. that the estimates of the 2SLS are statistically significantly different from the ones obtained by OLS. This test is presented at the bottom of Appendix A.4. For mathematics we fail to reject the null of exogeneity, whereas for reading we reject it at the 1% level of significance.[25] All in all, there is no evidence that possible endogeneity of class size, especially with respect to the reading specification, is contaminating the estimates of the class compositional variables in column (1) of Table 3 as all class compositional estimates do not change considerably when instrumenting class size or not. We then conclude that the estimation results of the model in equation (2) are reliable causal estimates of the class composition effects.

[24] This result differs from Wößmann & West (2006) who find that class size, for Portuguese students, has a significant positive coefficient on mathematics using TIMSS database and the same econometric methodology. However, the fact that our estimate for class size with respect to the math specification started as significantly positive in the simplest model in column (1), Table 2, and then turned non-significant by column (2) of Table 3, comes in line with their overall results. As they include school fixed effects and instrument class size with school-grade average class size they also observe the same movement of the class size coefficient: from significant positive to non-significant and from non-significant to significant negative, in general. In turn, Akerhielm (1995) finds class size having a non-significant effect for mathematics (as we do) as well as for reading.

[25] Since reading and mathematics is taught to the same class these results suggest that there are different endogeneity biases associated to class size estimates across different subjects. That is, across subjects there may be considerable differences in the way the same input impacts on the corresponding achievement level. In this case it may be that smaller classes are relatively more incremental to reading than to mathematics. Students may be more dependent of what is discussed in-class in reading as it may be a subject of a more communicative nature and suffer more with larger and more disruptive classes.

Table 3 *Estimation results using as outcomes the national exam scores in Mathematics and Reading*

Explanatory Variables		Pooled Grades				6th Graders' Sample		9th Graders' Sample	
		OLS		IV (2nd Stage)		OLS		OLS	
		(1)		(2)		(3)		(4)	
		Mathematics	Reading	Mathematics	Reading	Mathematics	Reading	Mathematics	Reading
% High Achievers ×	Low Achiever	-0.0024***	-0.0023	-0.0024***	-0.0022	-0.0025**	-0.0018	-0.0008	-0.0014
	Middle Achiever	-0.0041***	-0.0011**	-0.0041***	-0.0010*	-0.0040***	-0.0020***	0.0020**	0.0010
	High Achiever	-0.0064***	-0.0033***	-0.0064***	-0.0033***	-0.0054***	-0.0028***	0.0003	-0.0017
% Low Achievers ×	Low Achiever	0.0016**	-0.0048***	0.0016**	-0.0051***	0.0026**	-0.0040**	-0.0009	-0.0071
	Middle Achiever	0.0012**	0.0008	0.0011**	0.0006	0.0021***	0.0014*	-0.0025***	-0.0000
	High Achiever	0.0029***	0.0003	0.0029***	-0.0001	0.0024*	0.0047***	0.0020	0.0033
% Below Reference Age ×	Below Reference Age	0.0035***	0.0029***	0.0035***	0.0031***	0.0037***	0.0026***	0.0025***	0.0038***
	Above Reference Age	-0.0006	-0.0003	-0.0005	-0.0001	-0.0011	-0.0008	-0.0008	0.0010
% Males ×	Male	0.0005	0.0002	0.0005	0.0002	0.0011**	0.0001	0.0003	0.0011*
	Female	-0.0003	-0.0011***	-0.0003	-0.0011***	0.0001	-0.0007	-0.0001	-0.0005
% Foreigners ×	Foreigner	0.0037	-0.0077**	0.0037	-0.0076**	0.0047	-0.0090**	0.0066	-0.0032
	Non-Foreigner	-0.0012	-0.0030**	-0.0012	-0.0031**	0.0004	-0.0007	-0.0007	-0.0036**
% Internet ×	Internet	0.0011***	0.0006**	0.0011***	0.0007**	0.0015***	0.0013***	0.0007	0.0004
	No Internet	0.0005	0.0008**	0.0005	0.0008**	0.0008*	0.0012***	-0.0010*	-0.0004
% Low-Income ×	Low-Income	-0.0025***	-0.0026***	-0.0025***	-0.0027***	-0.0026***	-0.0027***	-0.0017***	-0.0007
	Non-Low-Income	-0.0039***	-0.0020***	-0.0039***	-0.0021***	-0.0046***	-0.0025***	-0.0022***	-0.0005
Age Dispersion		-0.070**	-0.036	-0.071**	-0.044*	-0.032	-0.056*	-0.082*	-0.100**
Class Size		0.002	0.000	0.001	-0.007**	0.002	0.003	-0.001	0.003
Below Reference Age		✓	✓	✓	✓	✓	✓	✓	✓
Male		✓	✓	✓	✓	✓	✓	✓	✓
Foreigner		✓	✓	✓	✓	✓	✓	✓	✓
Internet		✓	✓	✓	✓	✓	✓	✓	✓
Low-Income		✓	✓	✓	✓	✓	✓	✓	✓
Baseline Score		✓	✓	✓	✓	✓	✓	✓	✓
Parent Education Dummies		✓	✓	✓	✓	✓	✓	✓	✓
Grade Fixed Effects		✓	✓	✓	✓	--	--	--	--
School Fixed Effects		✓	✓	✓	✓	✓	✓	✓	✓
Adjusted R2		50.1%	44.2%	50.1%	44.2%	53.2%	46.7%	49.8%	43.8%
N		97,405	97,024	97,405	97,024	59,345	59,098	38,023	37,888

Notes: Significance levels: * p<.10, ** p<.05, *** p<.01. Robust standard errors clustered at the class level. Each outcome variable was standardized to have mean zero and std. dev. of one. Each model also contains dummies equal to 1 if student *i* peers' measures were computed using partial class information, i.e. if one, two or three classmates of *i* had missing information about their baseline scores or their place of birth. The class composition variables (i.e. the percentages of classmates of student *i* with a given characteristic) were computed in a leave-out fashion, i.e. excluding student *i*. Each model contains an intercept. Models of columns (1) and (2) pool students from grades 6 and 9, while those of columns (3) and (4) separate them. Only classes with 14 or more students were used and they had to belong to schools with at least one class of grade 6 and, simultaneously, another of grade 9 (i.e. each school contributed with at least two classes). The model of column (2) reports the second stage IV estimates of equation (2), see Section 4 in the text; the corresponding first stage output can be found in Appendix A.4.

5.2 *Results for separate 6th and 9th grades' models*

The estimation results of the model in equation (2) applied separately to the 6th and 9th grades samples are presented in Table 3, columns (3) and (4), respectively. We assume that the evidence in favor of the no endogeneity bias contamination provided by the comparison of the pooled sample estimates of columns (1) and (2) of Table 3 extends to the 6th and 9th grades samples' estimates of columns (3) and (4) since, as explained above, the IV procedure cannot be implemented when estimating separate models for each of the grades. However, since actual classes belong uniquely to a 6th or a 9th grade it is more valuable from a policy perspective to analyze in detail grade specific class compositional effects rather than composite ones that mix the specific effects of both grades.

The results suggest that, in mathematics, a given 6th grader is harmed when facing a higher percentage of high achievers in his/her class. This negative effect increases in magnitude as that given 6th grade student is defined as a better performer (i.e. as his/her baseline score is higher), see column (3) of Table 3. An increase of 20 percentage points (p.p.) of high achievers in his/her class (about 4 more high achievers and 4 less middle achievers in an average sized class of 20) leads to a *loss* in performance of 5%[26], 8% and 10.8% of a SD, depending if he/she is defined as a low, middle or high performer, respectively. In turn, for reading the respective values are 0% (not significant), 4%, and 5.6% of a SD, hence the same pattern is observed.

In turn, the percentage of low achievers deliver opposite results, in general. An increase of 20p.p. of low achievers leads to a statistically significant *gain* in mathematics performance around 4% to 5% of a SD, whether he/she is defined as a low, middle or high performer in the 6th grade. On the other hand, looking at the reading specification, positive variations in the proportion of low achievers *negatively* affect those 6th graders that are themselves low achievers: 20p.p. more low achievers translates to a *loss* of about 8% of a SD. The same increment of low achievers produces a *gain* to middle and high achievers of about 2.8% and 9.4% of a SD, respectively.

Next, the impact of increasing by 20p.p. the percentage of below reference age classmates (which is the reciprocal of a 20p.p. *decrease* in the percentage of students *above* the reference age) is to increase the achievement of 6th graders that are below the reference age ranging from 5.2% to 7.4% of a SD, depending on the subject we look at. Students above the reference age, i.e. with at least one retention in their past schooling trajectory, have no statistical evidence of being affected in any particular direction by the proportion of students below or above the reference age.

Sixth grade males seem to partially benefit from being exposed to a higher concentration of males, whereas for females there is no statistical evidence supporting any effect. In fact, for males, the effect of sharing the class with an increasing proportion of other males, with respect to mathematics achievement, is statistically significant (at 5% level) but small: their performance level rises by 2.2% of a SD given an increase of 20p.p. in the percentage of male classmates. Nevertheless, this effect vanishes when looking at the 6th grade reading specification.

---

[26] These values result from multiplying by 20 p.p. the respective estimate of the marginal effect of the class compositional variables (which are measured in percentage points). For example, $0.05 = 0.0025 * 20$, i.e. 0.05 standard deviations (SD) or, equivalently, 5% of a SD.

As in the previous case we only observe one particular significant effect regarding the impact of the in-class percentage of foreigners on 6th grade achievement. It is found on the reading specification: 6th graders who are themselves foreigners are harmed by larger shares of foreign classmates. This effect is around 4.5% of a SD given an increase of 5p.p.[27] in the percentage of students born abroad. Regarding mathematics, no significant effects are found relative to this compositional dimension for this grade.

One can notice, as well, that varying the proportion of students with home access to the internet, in a given class, seems to positively (but weakly) affect 6th grade students. The gain in having more 20p.p. in the percentage of classmates with internet varies from 1.6% to 3% of a SD, depending on subject and whether the student has internet at home or not.

Finally, sixth graders seem to be hurt with larger shares of low-income classmates. A 20p.p. increase in the proportion of low-income students, in a given 6th grade class, hampers the performance of a given low-income student from that class by about 5.3% of a SD, in both subjects; and that of a given non-low-income student from that class by 5% to 9.2% of a SD, depending on subject.

The estimation results concerning 9th grade students (Table 3, column 4) interestingly reveal that only approximately half of the considered class compositions have a statistical significant effect compared to those that have a significant effect on 6th grade students. In other words, it seems that class compositional effects seem to be less relevant explaining 9th grade achievement variation.[28] Nevertheless, there still remain a few important class compositional dimensions that deliver significant effects.

Both specifications of mathematics and reading for 9th graders present a similar pattern to those for 6th graders with respect to the marginal effects of the percentage of below reference age classmates. The impact of increasing by 20p.p. the percentage of below reference age classmates is to improve the achievement of 9th grade pupils that are themselves below the reference age by the same order of magnitude as for 6th graders. Ninth grade above the reference age students again have no statistical evidence of being affected in any particular direction by the proportion of students below or above the reference age.

Also, as in the case of 6th graders, the proportion of low-income classmates hampers the performance of both low-income and non-low-income 9th graders. Nevertheless, these effects are only significant in the mathematics specification and with a lower magnitude as compared with those that were estimated for 6th graders. We also note that higher concentrations of low-income students seem to do more harm to mathematics than to reading achievement (within each grade) and to impact more heavily 6th graders rather than 9th graders (within each subject).

The significant effects stemming from the 9th grade gender and foreign background class compositions are quite particular to the outcome variable used and the type of student, as they were with the 6th grade models.

[27] Although a 20p.p. variation in the proportion of foreign born students, in a given class, is within its sample variation (on both grades) it is still quite large compared to its standard deviation. Increasing the percentage of foreign born students, in a given class, by 5p.p. is closer to its standard deviation, see Table 1.

[28] We note that it is likely that as students get older their true network of peers may tend to substantially differ from the peers they have as classmates as a consequence of a more salient self-selection onto specific groups of peers. That is, the group of classmates, at more advanced grades, may be a noisy measure of the actual peer group each student interacts with. Consequently, it may not surprise that older students seem more insensitive to the composition of the class.

However, a common pattern seems to arise: the proportion of males seems to be beneficial for 9th grade males (as it was in the 6th grade), and the proportion of foreigners seems to be harmful with respect to reading achievement (as it was in the 6th grade).

The effect of the percentage of 9th grade classmates with internet at home differs from the one for 6th graders: the formers seem insensitive to this compositional dimension and in one case it is even negative, but faintly significant. Ninth graders with no internet at home are estimated to be slightly harmed as the proportion of classmates with internet at home increases.

Perhaps, the results related with the impact of high and low achievers on 9th grade classmates are the ones most markedly different from the results using the 6th grade sample. Ninth grade middle achievers are estimated to be significantly leveraged (around 4% of a SD) in mathematics given an increase of 20p.p. of the percentage of high achieving classmates. On the other hand, those same 9th grade middle achievers are harmed (around 5% of a SD) given the same increase of the proportion of low achieving classmates. The 9th grade reading specification yields no significant results which also contrasts with the respective 6th grade specification.

Lastly, the effect of the class age dispersion is more salient for 9th graders. Across both outcomes' specifications the effect is more significant and larger in magnitude than for 6th graders. Increasing the age difference (in absolute value) between the classmates of a 9th grade class and the corresponding class' average age by 0.2 years (its sample standard deviation) it is estimated that achievement falls by 1.6% to 2% of a SD, depending on subject.

6 Discussion

Looking back at what has been reported by the literature (e.g. Sacerdote, 2011), the empirical findings concerning the impact of high and low achievers can be seen as odd, in particular those related with 6th graders. Duflo, Dupas & Kremer (2011) indicate that tracking students according to their levels of achievement delivered the best outcomes; nevertheless, we observe the contrary. High achieving 6th graders break the tracking hypothesis as they have their outcomes *negatively* associated with an increasing proportion of other high achievers (in both subjects, but especially in mathematics where it reaches a sizeable effect of 10.8% of a SD). Low achievers of that same grade only report to profit out of an increasing proportion of other low achievers (hence supporting tracking) for mathematics. The tracking hypothesis, again, fails in reading for 6th grade low achievers – in face of more low achieving classmates they record a sizeable and precise negative effect of 8% of a SD. Hoxby (2000a), Hanushek et al. (2003) and Sund (2009) point to gains in achievement by the average student from having better peers (as measured by higher current or prior peer achievement levels) or, conversely, to losses from having weaker peers. Our results show that this is only fully observed for 9th grade middle achievers in mathematics. Although those authors can control for teacher heterogeneity which we can only partially here, they include less class compositional dimensions than we do. Still, if our results indeed incorporate possible bias associated with non-random teacher sorting within-schools then this is a rather strong confounding effect, at least for 6th graders. However, our results regarding 6th graders align somewhat with recent research – Burke & Sass (2013) – which documents that increasing too much class peer quality leads to a decrease in performance by the average student.

The fact that 9th graders do not seem to follow this pattern – they seem to fit better within the view that proportionally more high achievers (or less low achievers) improves educational achievement, at least in mathematics – may point to important differences in the way classmates interact with each other (and even with the teacher as a group) as they progress in age.

Looking at class gender composition its most precise estimate reveals that 6th grade males achieve slightly more in mathematics' national exams when placed in classes populated by relatively more males. This deviates from Hoxby (2000a) who estimates that relatively more females in the respective cohort help both males and females in mathematics. She also finds that larger shares of females help in reading achievement which contrasts with our faintly significant positive effect for 9th grade males of having more male classmates. In turn, she documents that peer effects are stronger and beneficial *within* cultural groups and our results align with her partially. Indeed, looking at the 6th grade sample, the effect of increasing the proportion of students with a foreign country cultural background is strong for those within that cultural group. But it is negative, not positive. Moreover, looking at the 9th grade sample, we observe that the unique significant effect is *across* cultural groups and negative.

Contrary to our results in Table 3 regarding the proportion of low-income classmates, Hanushek et al. (2003) reported evidence not supporting the view that lower income peers harm achievement. They argued that eligibility to a reduced-price lunch is a noisy measure of actual income differences which could in part explain their findings. Even if our binary variable flagging low-income students incorporates some measurement error we see our results as evidence that indeed low-income classmates have a detrimental effect on achievement, especially in mathematics (both grades) and on 6th graders (both subjects). We argue that the statistically significant results from Table 3 are then, in the worst-case scenario, lower bounds of the absolute value of the true effect of low-income classmates due to attenuation bias.

7 Policy Implications

So far, we detailed and discussed the estimated effects of several class compositional dimensions. We now provide policy implications based on those that seem to have the largest impact at improving overall education achievement.

It is important to recognize first, however, that distributing a potentially positive educational input (e.g. classmates with characteristics that facilitate classroom learning) across classes seems fairer from the perspective of equality of opportunity, that is, of not letting particular groups of pupils be denied exposure to it. In turn, distributing potentially negative inputs (such as classmates who tend to disrupt the classroom) can be defended by the symmetric argument of not letting particular groups of pupils be over exposed to it. One such case is the distribution of high and low achievers across 6th grade classes. That is, taking our results at face value, it should be fairer to target heterogenous 6th grade classes, contrary to what Collins & Gan (2013) point, since those types of classmates seem to act, respectively, as negative and positive inputs, at least to the majority of students. For 9th graders, class heterogeneity seems also the fairest since high and low achievers still affect achievement in contrary directions: the proportions of high and low achievers act as positive and negative inputs, respectively. In other words, both 6th and 9th grade classes of a given school should not deviate, in terms of the proportions of high and low achievers, from the respective proportions observed in that given school-grade population. Otherwise would signal the existence of groups of pupils placed in classes under or over exposed to these particular positive or negative inputs. The same reasoning must be applied to the class level proportion of low-income pupils. This input has been estimated as being harmful to both low and non-low-income students of both grades hence it should be socially fairer to evenly spread low-income students across all classes of a given school-grade. Again, each class should reflect the heterogeneity of low and non-low-income students found at the respective school-grade population.

Finally, the proportion of students, in a given class, below or above the reference age is the class compositional dimension that calls for homogeneity. According to the estimates, tracking and grouping students below the reference age in the same class should increment achievement for those that are themselves below the reference age. At the same time, tracking students above the reference age to other distinct classes should not bring any positive nor negative effect upon them. Tracking students whether their current age is below or above the grade reference age should also capitalize on the positive effect of a smaller class age dispersion (especially for 9th grade classes).

In a nutshell, we report here evidence that seems to support the tracking of students merely along the past retention dimension as an educational efficiency improving policy, whereas along all the other dimensions it seems socially fairer to reflect school-grade population heterogeneity. As a comparison, reducing class size by 4 (comparable to the hypothetical 20p.p. variations of each class compositional variable with respect to an average sized class of 20 as performed in Section 5) entails a performance gain in reading[29] of 2.8% of a SD which is lower than many of the effects of class compositional variations discussed earlier. That is, class compositional rearrangement is estimated to provide, in some dimensions, a larger performance improvement than comparable variations of class size. On top of this it is necessary to add that class compositional rearrangement should be cost-benefit superior to cutting class size as rearranging the composition of the classes brings no financial costs to schools, whereas class size reduction implies, in principle, hiring extra teachers with its associated costs.

8 Conclusion

Exploiting a relatively recent Portuguese educational micro dataset, in particular its 6th and 9th grades' samples of students enrolled in public schools in the 2011-12 academic year, this paper documents evidence supporting: 1) the existence of relevant class compositional effects, 2) the heterogeneity of these effects with respect to students' individual characteristics and 3) grade and subject specific differences of the class compositional effects. We list the class compositional measures related to the proportion of high and low achievers, the proportion of students below or at the reference age, and the proportion of students belonging to low-income households as the ones more relevant for policy implications.

[29] Looking at the best causal, statistically significant, estimate of the class size effect present in the 2SLS model of column (2) in Table (3).

We find that low, mid and high 6th grade achievers perform *worse*, in mathematics and reading, given an *increase* of high achieving students. These same types of 6th graders perform, in general, *better*, in both subjects, given an *increase* of low achieving students. In turn, we find that, contrasting with the 6th grade, the 9th grade middle achiever performs *better* in mathematics, given an *increase* of high achieving students. This type of 9th grader performs *worse* in mathematics, given an *increase* of low achieving students. A larger share of low-income classmates deteriorates performance levels in general. Finally, the performance of students with no prior retentions increments with a higher proportion of classmates who have also not been retained in the past.

All in all, the results obtained in this paper point to the conclusion that schools have room to improve the composition of the classes. In general, class composition heterogeneity seems fairer from a social point of view, especially along specific dimensions such as the proportions of high and low achievers and of low-income students. That is, the composition of each class, along these dimensions, should reflect the respective school-grade population heterogeneity. Nevertheless, we also find evidence that tracking students according to whether they have recorded a retention in the past or not, that is, targeting homogeneous classes in this specific dimension, is incremental to performance. These are worthwhile policy implications as they entail sizeable positive increments to student achievement at virtually no financial cost. Contrasting to class composition rearrangement, class size reduction was estimated to deliver a weaker increment to performance. Given that class size reduction is likely to require a non-marginal financial effort from the part of the educational system, class composition rearrangement seems superior as an alternative policy.

Acknowledgments

We acknowledge financial support from National Funds through FCT – Fundação para a Ciência e a Tecnologia – under project PTDC/IIM-ECO/6813/2014 and by POR Lisboa under the project LISBOA-01-0145-FEDER-007722. João Firmino gratefully acknowledges the support of the grant SFRH/BD/122573/2016 from FCT. All funding sources had no involvement in any phase of this study. We thank DGEEC-Ministry of Education of Portugal for providing the data and Diogo Pereira for excellent assistance with the dataset. We also thank participants at the XXV Meeting of the Economics of Education Association (AEDE), the Fourth Lisbon Research Workshop on Economics, Statistics and Econometrics of Education (ISEG) and the Fórum Estatístico (21.ª Sessão, DGEEC) for helpful comments. The views expressed here are those of the authors and do not necessarily reflect the views of FCT or DGEEC-Ministry of Education of Portugal. Any errors are our own responsibility.

References

Akerhielm, K. (1995). Does class size matter?. *Economics of education Review*, *14*(3), 229-241.

Bosworth, R. (2014). Class size, class composition, and the distribution of student achievement. *Education Economics*, *22*(2), 141-165.

Burke, M. A., & Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics*, *31*(1), 51-82.

Card, D., & Giuliano, L. (2016). Can Tracking Raise the Test Scores of High-Ability Minority Students?. *The American Economic Review*, 106(10), 2783-2816.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. (1966). Equality of educational opportunity (Coleman Report), *U.S. Department of Health, Education, and Welfare*.

Collins, C. A., & Gan, L. (2013). *Does sorting students improve scores? An analysis of class composition*. NBER. Working Paper 18848.

Direção-Geral de Estatísticas da Educação e Ciência (DGEEC) (2013). *Estatísticas da Educação 2011/2012 – Jovens*. Lisboa: Direção-Geral de Estatísticas da Educação e Ciência.

Duflo, E., Dupas, P, & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5): 1739–1774.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, 123, 92-110.

Hanushek, E. (1970). The Production of Education, Teacher Quality, and Efficiency. In *U.S. Office of Education, Do Teachers Make a Difference?*, 79-99. Washington, D.C.: Government Printing Office.

Hanushek, E. A. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *Journal of Human Resources*, 14(3), 351-388.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement?. *Journal of applied econometrics*, 18(5), 527-544.

Hanushek, E. (2008). Education Production Functions. In *The New Palgrave Dictionary of Economics Online*, ed. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan.

Hanushek, E. A., & Rivkin, S. G. (2010). *Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools?*. NBER. Working Paper 15816.

Hoxby, C. (2000a). *Peer Effects in the Classroom: Learning from Gender and Race Variation*. NBER. Working Paper 7867.

Hoxby, C. (2000b). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, 115(4), 1239-1285.

Jürges, H., & Schneider, K. (2004). International differences in student achievement: An economic perspective. *German Economic Review*, *5*(3), 357-380.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, 116(3), 777-803.

Pritchett, L., & Filmer, D. (1999). What education production functions really show: a positive theory of education expenditures. *Economics of Education review*, 18(2), 223-239.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far. *Handbook of the Economics of Education*, 3(3), 249-277.

Sund, K. (2009). Estimating peer effects in Swedish high school using school, teacher, and student fixed effects. *Economics of Education Review*, 28(3), 329-336.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485).

West, M. R., & Wößmann, L. (2006). Which school systems sort weaker students into smaller classes? International evidence. *European Journal of Political Economy*, 22(4), 944-968.

Wößmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3), 695-736.

Yao, Y., Ohinata, A., & van Ours, J. C. (2016). The educational consequences of language proficiency for young children. *Economics of Education Review*, 54, 1-15.

Appendix A

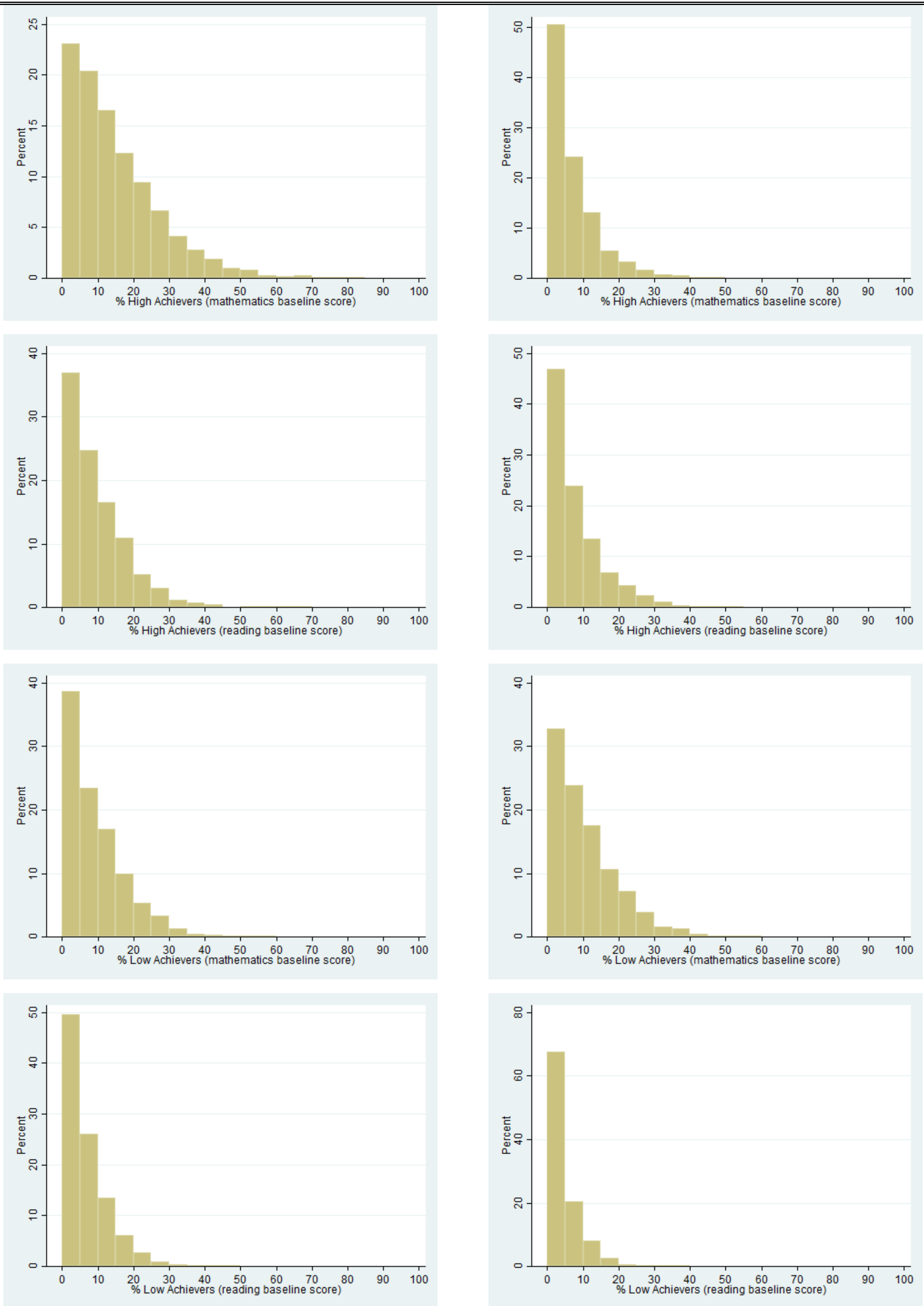
Appendix A.1. Decomposition of 6th and 9th graders according to baseline score per subject.

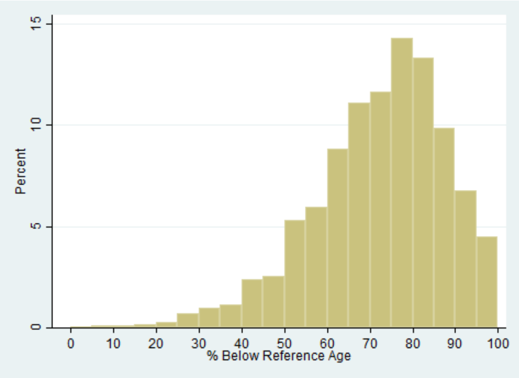
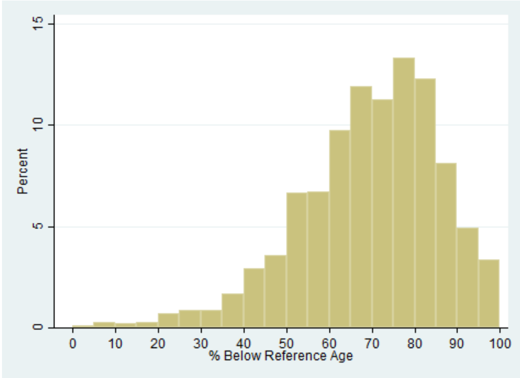
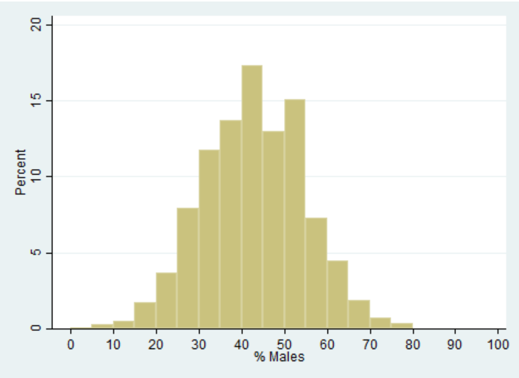
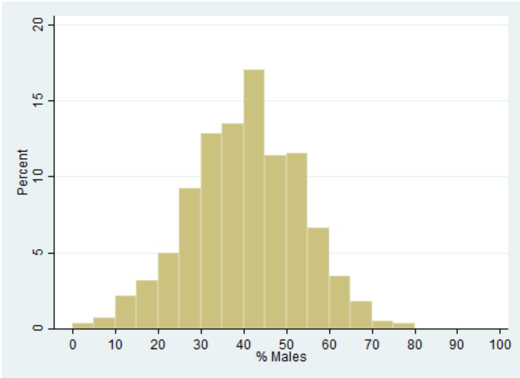
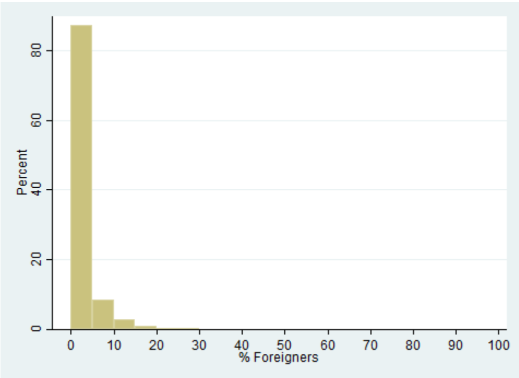
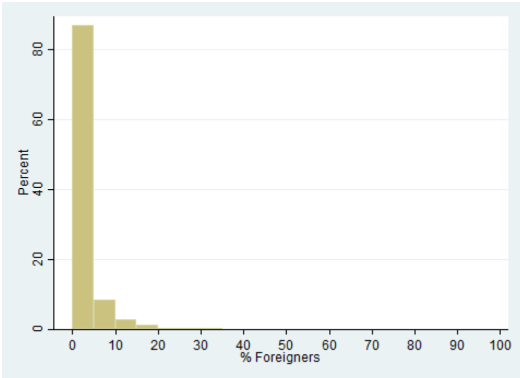
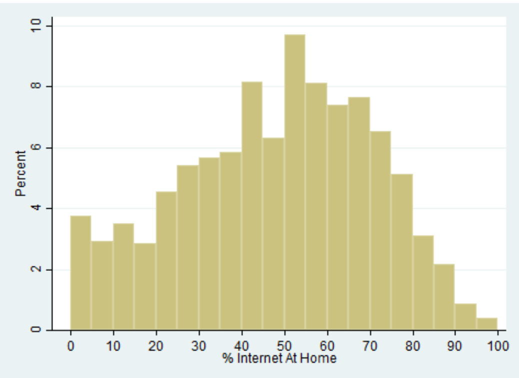
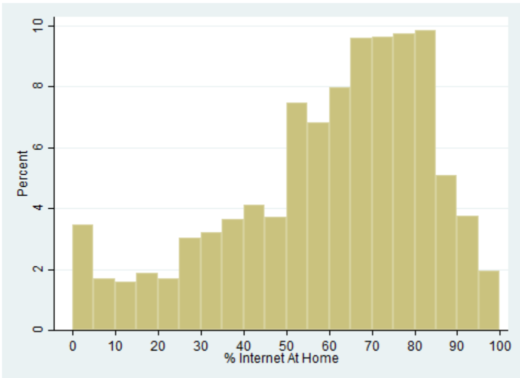
6th Mathematics	N	Percent	9th Mathematics	N	Percent
Low Achievers (Baseline Score = 1 or 2)	5,416	9.12	Low Achievers (Baseline Score = 1 or 2)	3,956	10.4
Middle Achievers (Baseline Score = 3 or 4)	43,733	73.68	Middle Achievers (Baseline Score = 3 or 4)	31,139	81.85
High Achievers (Baseline Score = 5)	10,210	17.2	High Achievers (Baseline Score = 5)	2,951	7.76
Total	59,359	100	Total	38,046	100
6th Reading	**N**	**Percent**	**9th Reading**	**N**	**Percent**
Low Achievers (Baseline Score = 1 or 2)	3,850	6.51	Low Achievers (Baseline Score = 1 or 2)	1,236	3.26
Middle Achievers (Baseline Score = 3 or 4)	48,902	82.73	Middle Achievers (Baseline Score = 3 or 4)	33,330	87.91
High Achievers (Baseline Score = 5)	6,360	10.76	High Achievers (Baseline Score = 5)	3,346	8.83
Total	59,112	100	Total	37,912	100

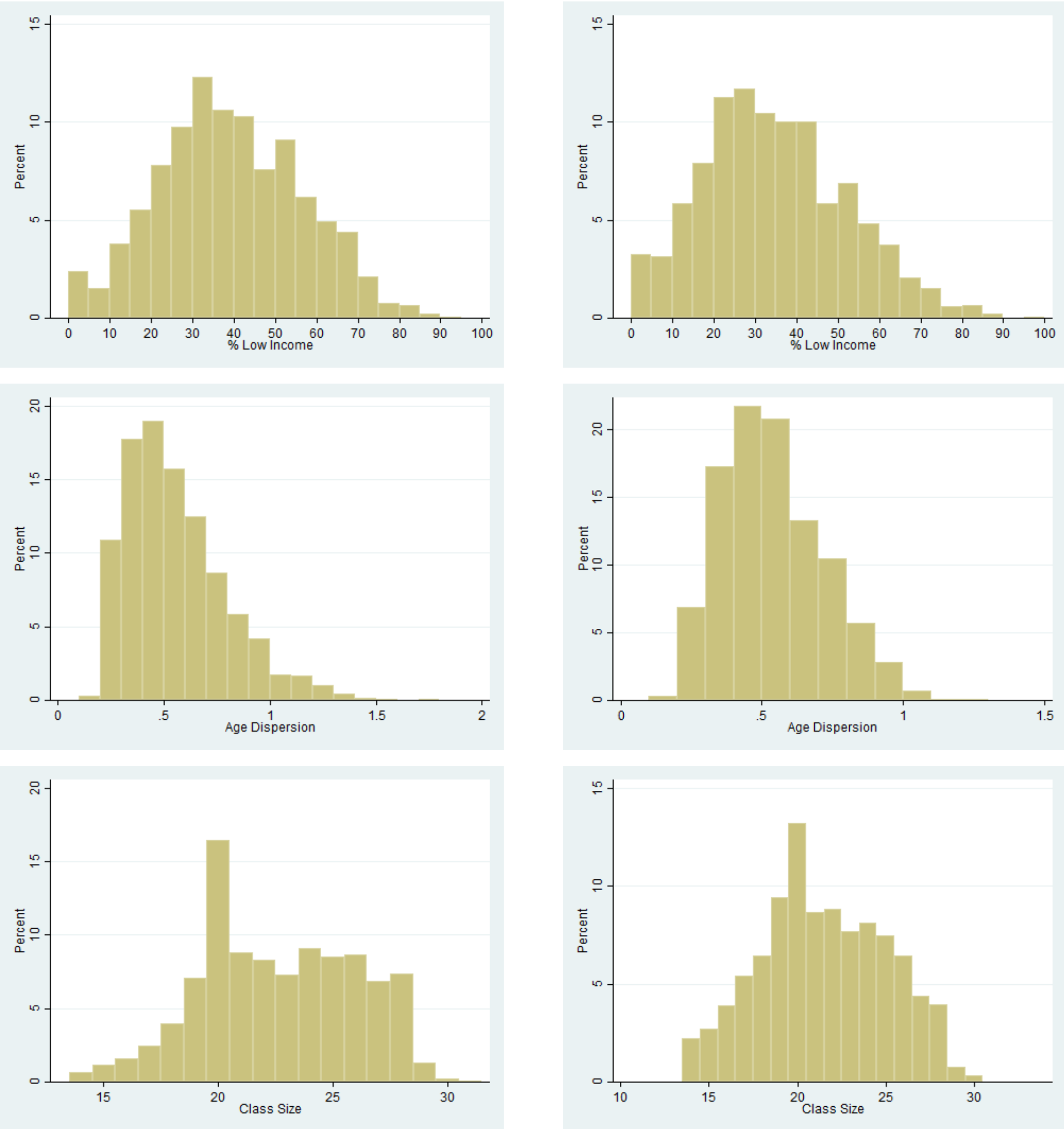
Appendix A.2. Descriptive statistics.

		6th Grade - Reading National Exam					9th Grade - Reading National Exam				
		N	Mean	Std.Dev.	Min	Max	N	Mean	Std.Dev.	Min	Max
Individual Level Variables	Score	59,112	59.1	15.9	1	100	37,912	52.9	14.8	0	100
	Baseline Score	59,112	3.1	0.8	1	5	37,912	2.8	0.7	1	5
	Reference Age	59,112	0.88	0.33	0	1	37,912	0.84	0.37	0	1
	Male	59,112	0.51	0.50	0	1	37,912	0.47	0.50	0	1
	Foreigner	59,112	0.01	0.12	0	1	37,912	0.01	0.12	0	1
	Internet	59,112	0.60	0.49	0	1	37,912	0.73	0.45	0	1
	Low-Income	59,112	0.45	0.50	0	1	37,912	0.39	0.49	0	1
	Tertiary Ed. (Parent)	59,112	0.18	0.38	0	1	37,912	0.17	0.37	0	1
	Secondary Ed. (Parent)	59,112	0.48	0.50	0	1	37,912	0.46	0.50	0	1
Class Level Variables	% High Achievers	3,552	14	12	0	81	2,336	7	8	0	53
	% Low Achievers	3,552	9	8	0	59	2,336	4	5	0	36
	% Reference Age	3,552	72	16	0	100	2,336	69	17	0	100
	% Males	3,552	43	12	0	79	2,336	40	13	0	80
	% Foreigners	3,552	2	3	0	29	2,336	1	3	0	33
	% Internet	3,552	48	23	0	100	2,336	59	24	0	100
	% Low-Income	3,552	39	17	0	95	2,336	34	18	0	90
	Age Dispersion	3,552	0.6	0.2	0.2	1.7	2,336	0.5	0.2	0.2	1.3
	Class Size	3,552	23	3	14	31	2,336	22	4	14	30

Appendix A.3. Distributions of class level variables - 6th grade classes on left column and 9th grade classes on right column.




Percent
% Below Reference Age

Percent
% Below Reference Age

Percent
% Males

Percent
% Males

Percent
% Foreigners

Percent
% Foreigners

Percent
% Internet At Home

Percent
% Internet At Home


Percent
% Low Income
Percent
% Low Income
Percent
Age Dispersion
Percent
Age Dispersion
Percent
Class Size
Percent
Class Size

Appendix A.4. First stage regression outputs using the Mathematics and Reading samples.

Explanatory Variables			IV (1st Stage) Pooled Grades (1) Mathematics	(2) Reading
Average Class Size (excluded instrument)			0.9455714***	0.9464064***
% High Achievers	×	Low Achiever	0.0108246**	0.0134733*
		Middle Achiever	0.0137038***	0.0157996***
		High Achiever	0.004842	0.0030827
% Low Achievers	×	Low Achiever	-0.0201416***	-0.0223342**
		Middle Achiever	-0.0197364***	-0.0301222***
		High Achiever	-0.0270461***	-0.0407922***
% Below Reference Age	×	Above Reference Age	0.0306045***	0.0307392***
		Below Reference Age	0.032289***	0.0309265***
% Males	×	Female	-0.0015243	0.000782
		Male	-0.0013183	0.0012115
% Foreigners	×	Non-Foreigner	-0.0056713	-0.0058593
		Foreigner	0.0212817	0.0148748
% Internet	×	No Internet	0.0016562	0.0014092
		Internet	0.0023914	0.0022271
% Low-Income	×	Non-Low-Income	-0.0019125	-0.0025982
		Low-Income	-0.0045517	-0.0047458*
Age Dispersion			-1.087461***	-1.141959***
Below Reference Age			-0.0486207	0.043228
Male			-0.0359362	-0.0306013
Foreigner			-0.2478417***	-0.2012908**
Internet			-0.0531369	-0.0607908
Low-Income			0.0875046	0.0734998
Baseline Score		if 2	0.092764	0.8299373***
		if 3	0.1596036	0.9895933***
		if 4	0.1515098	1.039625***
		if 5	0.4007196**	1.270225***
Parent Education		if Tertiary	0.0991892***	0.0944203***
		if Secondary	0.0490241***	0.0490555***
Grade Fixed Effects			✓	✓
School Fixed Effects			✓	✓
Adjusted R2			61.7%	61.7%
F statistic (overall significance)			452.10	1063.84
(P-Value)			0.000	0.000
F statistic (excluded instrument significance)			1822.67	1862.4
(P-Value)			0.000	0.000
F statistic (test of endogeneity)			0.0552	7.8316
(P-Value)			0.814	0.005
N			97,405	97,024

Notes: Significance levels: * p<.10, ** p<.05, *** p<.01. Robust standard errors clustered at the class level. Each model also contains dummies equal to 1 if student *i* peers' measures were computed using partial class information, i.e. if one, two or three classmates of *i* had missing information about their baseline scores or their place of birth. The class composition variables (i.e. the percentages of classmates of student *i* with a given characteristic) were computed in a leave-out fashion, i.e. excluding student *i*. Each model contains an intercept and pools students from grades 6 and 9. Only classes with 14 or more students were used and they had to belong to schools with at least one class of grade 6 and, simultaneously, another of grade 9 (i.e. each school contributed with at least two classes).