Class Composition and Student Achievement in Portugal

João Firmino; Luís C. Nunes; Ana B. Reis; Carmo Seabra

Nova School of Business and Economics, Universidade Nova de Lisboa, Lisbon, Portugal

joao.mgs.firmino@gmail.com; lcnunes@novasbe.pt; ana.balcao.reis@novasbe.pt; carmo.seabra@novasbe.pt

In this paper we estimate class composition effects impacting on achievement levels of Portuguese students. Endogeneity between student achievement and student non-random sorting across schools and classes may prevent the correct identification of class composition effects. Using student level cross sectional data of 6th and 9th graders (2011/12 academic year) provided by *MISI* dataset we contrast a relatively recent estimation procedure in the literature – involving a proper instrument (IV) coupled with School Fixed Effects (SFE) – with usual OLS as means to properly identify the composition effects free of endogeneity bias. Several dimensions of class composition were identified as consistently impacting national exam scores on Portuguese and Mathematics. Namely, the proportion, in a given class, of pupils: 1) under the relevant grade reference age; 2) of low income households (negative impact) and 3) with home access to internet (positive impact), to mention a few. Many of the effects are statistically significantly asymmetric (e.g. an increasing proportion of students aged at or below the relevant grade reference age in a class seems to affect positively this type of classmate while hurting those aged above it). Non-linear effects are also analysed. In turn, class size yields no significant effect on achievement, while class gender composition uniquely affects boys' achievement in Portuguese. Given that in the past recent years Portugal has been put under tight public budgetary management it is even more important to identify class compositional effects. Their identification, which this paper contributes to, can provide policy orientations capable of delivering positive increments to student achievement while, at the same time, be budget neutral. Taking the results obtained it seems that optimally allocating students across classes seems more attractive than to increase teacher spending to cut class size.

Keywords: class composition; student achievement; IV; School FE.

We thank Direção-Geral de Estatísticas da Educação e Ciência (DGEEC) for providing all relevant microdata used in this paper (*MISI* dataset).

1 Introduction

It is a demanding task to establish which determinants are relevant explaining educational achievement and how to rank them in importance. Many different inputs have been analyzed: students' characteristics and their family background, school and class characteristics, as well as teachers and type of education system, to mention a few. This work investigates the impact of different class compositions on students' achievement, among 6th and 9th grades' pupils of Portuguese public schools. The focus is on the determination of causality from class composition to student achievement. Whom each pupil shares the class with may well determine the amount of time he is able to listen to the teacher or what kind of lessons he faces or even the peers with which he will interact outside school. Being assigned to a class of brilliant peers may be the best or the worst for a given pupil. Either he may benefit from minimal class disruption or he may face a stream of non-supportable highly paced lessons whether the teacher focuses on his brilliant classmates rather than in himself. All of these considerations imply causality from certain class compositions to student performance. This work is structured as follows. A framing literature review is provided in the next section. The 3rd section depicts the appropriate dataset and its descriptive statistics. Section 4 details the OLS and IV econometric models. The 5th section presents the empirical results and section 6 concludes, pointing to potential policy implications.

2 Literature Review

Checchi (2006), chapter 4, provides a theoretical framework about class composition placing it within the supply side of the education market. After all, the type of schooling offered to students is influenced by school policies, with respect to class formation, in, at least, two ways: how many and what kind of classmates exist in each class. An illustrative model presented in his chapter, by Lazear (2001), shows that, *for a given level of students' quality* (measured as the fraction of time each pays attention to the teacher) increasing the class size would exponentially decrease class learning time – more students, more disruptions. Although it departs from the extreme assumption that students avoid synchronizing class disruption (making it as greater as possible), it is a useful concept that can be applied not only to the *amount* of students in a class but also to their *type*: *for a given class size*, increasing the proportion of disrupting-like students should also exponentially decrease overall classmates' learning.

The theoretical class size and compositional effects may be extracted from a broader input-output function mapping a wide range of educational inputs to possible educational outcomes. Those "effects" would then be the derivatives of that function with respect to the particular inputs "class composition" and "class size". This function, termed as the educational production function, is used, either as a theoretical or an econometric formulation, not only by

Checchi, but as well by many other authors, e.g. Wößmann and West (2006), Lazear (2001), Pritchett and Filmer (1999) and Hanushek (1970)[1]. And even earlier the Coleman Report (Coleman et al (1966)) already presented the idea that educational outcomes were linked to a set of inputs which included the sort of peers one finds in his school. The report pointed that although the main significant predictors of educational outcomes were family and socio-economic background also the student body composition predicted outcomes (especially those of minorities).

Empirically, when estimating class level effects, endogeneity and self-selection are issues to be tackled. These arise from possible non-random sampling of students across schools and classes which is likely to be correlated with unobserved characteristics of students or of their parents[2]. Hoxby (2000a) exploits idiosyncratic variations of gender and racial compositions in American schools, between adjacent years, due to unanticipated demographic changes, to avoid non-randomness issues. She finds, firstly, that if the class average exam score increases, unexpectedly, by 1 point, then a student from that class scores more 0.1 to 0.5 points. That is, the presence of high achievers in class increments everybody's performance. Secondly, that more female classes cause better performances in math for males. Finally, that peer effects are stronger and beneficial within racial groups. Hoxby (2000b), on the other hand, finds no significant effects from class size to pupil achievement. Hoxby uses credible random population variation as instrument for class size while also applying school fixed effects. Her rationale is based on the assumption that the residuals of a time polynomial fit on a time series of enrollment, in a given school, are unexpected population shocks not anticipated by those agents that can endogenously affect the sample selection both at school as well as at class levels (i.e. parents, teachers and principals). Hence they serve as valid instruments for class size. Although it is a rather interesting method to detect exogenous class level variables' variation it requires data that we do not have[3].

There is a stream of literature that makes use of grade-school averages as instruments for class level variables. Akerhielm (1995) premiers such procedure using average class size across a given subject within a school to instrument actual class size. She finds class size having a negative point estimate, but significant at only some subjects. Although her procedure does account for within school sorting, it did not take in account between-school sorting (no school FE). Jürges and Schneider (2004), Wößmann and West (2006) and West and Wößmann (2006), again, employ a two stage regression procedure to identify class size effects (controlling for

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

[2] Teacher sorting at both school and class levels may also be a source of endogeneity. We discuss this issue in the last section of the paper.

[3] It requires several years of enrolment data. Hoxby uses 24 years to fit a quartic time polynomial.

within school sorting) in TIMSS' database. In the first stage the exogenous variation of class size is obtained instrumenting actual class size (the endogenous variable) with the average class size of the respective grade (instead of averaging across subject). They hold the instrument as valid assuming that schools cannot respond to changes in performance of adjacent cohorts[4] by reducing or enlarging class size. On the second stage, student test score is regressed to the instrumented class size and control variables. The main difference being now they control for between-school sorting by including school FE whereas Akerhielm did not. They find that reducing class size has no meaningful effect, in particular for Portugal[5]. Beyond class size effects, i.e. with respect to class composition effects, recent literature points to gains from having homogeneous classes according to past student performance. That is concluded by Collins and Gan (2013) who constructed an index of how much students are sorted, at the class level (instrumented with the adjacent grade index) as proxies of class homogeneity in a student achievement regression. Duflo, Dupas and 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. But more interestingly they 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 positive habile-peers' effects, benefited from better tailored teaching (although the extent to which results from randomized experiments in developing countries can be linearly scaled to developed ones is disputable).

3 The Data

The *MISI* dataset is compiled by the Portuguese Ministry of Education. It encompasses all students enrolled in public schools, in continental Portugal, coursing from the 1st to the 12th grades. We use its 2011\2012 academic year cross-section. It provides information on relevant students' characteristics: their class and school membership, their scores by subject and type of examination and their academic track. National exams' scores (high-stakes scores) of Mathematics and Portuguese taken at the end of the 2011\2012 academic year provide the achievement measure, while a baseline score is collected from a low-stakes national test (with no consequences for student progression, contrary to national exams which do matter for

[4] Two consecutive grades are pooled and then regressed using a grade dummy control.

[5] This methodology will be the one we will employ in this paper, as it takes in account both between and within school sorting issues and fits well in the data we have available. The novelty we introduce are the class compositional variables. Besides being the main variables of interest in this paper they also contribute to the correct identification of class size effect due to the dual relation between quantity and type of peers, as discussed above.

student progression)[6]. Up to 2012 only students enrolled in grades 4, 6, 9 and 12 were mandated to take one of the low or high-stakes exams. Hence we discard to use the 4th grade since there is no national wide standardized baseline score. Following the Wößmann and West (2006) econometric methodology we should pool students from two consecutive grades. Since standardized exams do not happen for consecutive grades we are left with two possibilities: using the pool of 6th and 9th graders or the one with 9th and 12th graders. For the sake of concreteness we focus on the first pool. 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 out of all public schools' students, coursing the 2nd and the 3rd cycles of the Basic (where 6th and 9th grades belong to, respectively) 99% and 90% were in that track, respectively. Both (i) individual and (ii) class level variables, regarding student and classes' characteristics, are present or computable. Set (i) is composed by parents' academic background and the student's reference age[7], gender, place of birth, home access to internet and beneficiary status on both economic and academic support programs. Set (ii) contains: class size, fraction of classmates with each of the individual characteristics expressed in (i)[8] and a measure of class age dispersion. *Academic background* stands for parents' education (of the parent with the highest degree[9]): basic or no schooling, secondary and college degree. *Below Reference Age*: dummy variable distinguishing students whose age is equal or lower than their reference age. The reference is the maximum age a student is expected to have without having failed any past academic year. For 6th grade students the reference age is 12, while for 9th graders it is 15. *Place of birth*: dummy differentiating whether the student was born in Portugal or in the Community of Portuguese Language Countries (CPLP)[10]. *Beneficiary of socio-economic support* (SASE): dummy discerning whether the pupil enjoyed or not of any level of economic aid[11] under the program "schooling social action". *Beneficiary of academic*

[6] Being outcomes of national wide standardized exams\tests these avoid school and teacher specific differences embodied in internal school scores. Both scores vary from 1 to 5. Nevertheless the baseline scores were demeaned using the *grade-academic year* specific observed mean to control for time varying difficulty levels in the low stakes exams. Given that some students experience retention we observe these taking the low stakes exams in different academic years which have differing difficulties. Hence the baseline score is the position of a student relative to the average score observed in the year the student took that test, relative to the grade he belonged to.

[7] Assuming the student age as of September 15, 2011 (limiting date for the beginning of the classes).

[8] Excepting for parental background.

[9] Using this measure of parental academic background allows one to avoid missing values of a particular parent in this dimension, provided the dataset has information on the other parent. It also captures the highest academic influence to which the student is exposed at home.

[10] The category of being born in Portugal includes students born in any other country than the CPLP ones. A third category differentiating those students not born neither in Portugal nor in a CPLP country would be of small size and of dubious homogeneity. Annex 1 provides a list depicting the countries that compose the CPLP category.

[11] It consisted in subsidizing, among others, the student's alimentation and cost of textbooks. Only students living in households whose earnings belonged to the two lowest categories of income could have had access to this program. This flags students living in low income families.

support: dummy distinguishing students to whom was assigned academic support (usually extra classes) given by schools[12]. The dataset presents students enrolled in classes with too reduced dimensions in relation to what was stipulated by law: minimum and maximum of 24 and 28, respectively. Nevertheless, the law also does allow exceptional lower ones (e.g. to group pupils that would overflow the limits of the remaining classes). We kept classes with, at least, 14 students[13]. *Age dispersion*: mean absolute deviation of students' age to their class average age. The final amount of appropriate observations for econometric analysis are about: 65k (6th grade) and 42k (9th grade)[14]. These numbers correspond to about 63% and 49% of the students enrolled in the regular academic track of 6th and 9th grades[15], respectively, in continental Portugal´s public schools. The appropriate descriptive statistics are presented in Table 1[16]. Regarding the descriptive statistics of class level variables note that the relevant number of observations, in this context, is the number of classes, not of students (hence the reduced number of observations for these variables). The complete distributions of the class level variables are shown in Annex 4. Although purposeful sorting of students may not be proved by simply inspecting those

Table 1. Descriptive statistics under the appropriate sample of students.

		6th Grade - Mathematics National Exams					9th Grade - Mathematics National Exams				
		N	Mean	Std.Dev.	Min	Max	N	Mean	Std.Dev.	Min	Max
Individual Level Variables	Score	65,552	2.8	1.0	1	5	42,118	2.8	1.1	1	5
	Baseline Score	65,552	0.0	0.9	-2.6	1.5	42,118	0.2	0.8	-2.2	1.9
	Tertiary (Max)	65,552	0.19	0.39	0	1	42,118	0.17	0.37	0	1
	Secondary (Max)	65,552	0.48	0.50	0	1	42,118	0.46	0.50	0	1
	Below Reference Age	65,552	0.88	0.33	0	1	42,118	0.82	0.38	0	1
	Male	65,552	0.51	0.50	0	1	42,118	0.48	0.50	0	1
	CPLP	65,552	0.02	0.13	0	1	42,118	0.02	0.13	0	1
	Internet	65,552	0.60	0.49	0	1	42,118	0.72	0.45	0	1
	SASE	65,552	0.44	0.50	0	1	42,118	0.40	0.49	0	1
	Academic Support	65,552	0.10	0.30	0	1	42,118	0.14	0.35	0	1
Class Level Variables	Class Size	4,023	23	3	14	31	2,582	22	4	14	32
	% Below Reference Age	3,998	80	14	0	100	2,411	77	14	17	100
	% Males	4,023	52	11	10	100	2,582	49	12	6	87
	% CPLP	4,020	3	6	0	55	2,582	3	6	0	57
	% Internet	4,023	55	25	0	100	2,582	68	25	0	100
	% SASE	4,023	48	19	0	100	2,582	42	19	0	100
	% Academic Support	4,023	12	15	0	94	2,582	16	21	0	96
	Age Dispersion	3,998	0.6	0.2	0.2	1.7	2,411	0.5	0.2	0.2	1.3

[12] This information flags those students struggling during the 2011\2012 academic year. Looking to the baseline score is not enough for those students that did well at the baseline exam but developed learning difficulties in the meanwhile.

[13] We also assume that students that left the class (and the school) before 1 January 2012 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.

[14] These figures refer to the students for whom there is a full set of information across all variables and that are placed in schools which have, simultaneously, at least one class of each 6th and 9th grades. This last restriction is due to the identification strategy adopted.

[15] Percentages out of the totals 104 410 and 86 416 for 6th and 9th grades, respectively, DGEEC (2013) (pages 68 and 72).

[16] Annex 2 depicts a more appealing codification of the variables. It will be used in the descriptive statistics and regression tables. These statistics refer to students with a Mathematics National exam score. Annex 3 present the same statistics for those with a Portuguese National exam score. The two populations are very similar.

distributions they interestingly point to possible sorting at the students' past achievement and economic dimensions. The former given the appearance of what seems a fat left tail, meaning that the occurrence of classes with proportionally many students above the reference age is relatively frequent. The latter given what seems an abnormal high frequency (relatively to the rest of the distribution) of classes with 0 to 5% SASE students, i.e. of low income households, which means a relatively high frequency of classes with 0 or at most 1 student with that status. Finally, to avoid being driven by outlier classes and taking in consideration the class level variables' distributions we restrict the sample to students belonging to classes satisfying the following restrictions: class size $\in$ [14, 31], % Below Reference Age $\in$ [40, 100], % Males $\in$ [10, 87], % CPLP $\in$ [0, 30], % Internet $\in$ [0, 100], % SASE $\in$ [0, 100], % Academic Support $\in$ [0, 60] and Age Dispersion $\in$ [0.17, 1.1].

4 Econometric Methodology

As mentioned in the Literature Review one needs to tackle important econometric issues to identify class composition effects on student achievement. Namely, one has to control for i) sample sorting bias, ii) all relevant explanatory variables and iii) unobserved student's characteristics. Point i) is decomposed in between and within-school non-random sampling of students. Between-school sorting arises whenever parents are stratified regionally according to professional occupation, level of educational attainment or income. One way to control for between-school sorting is to include school dummies[17]. Within-school sorting takes the form of, for example, arranging classes segregating low achievement students from the others, or segregating whether they have been retained in the past or not (as it seems to be the case in Portugal, see the previous section). Whatever the form of systematic class composition employed by school authorities, not taking it in consideration may lead us to attach a causality implication that is not true. To overcome this we first stress the inclusion, with respect to every individual student, of a baseline score and of his reference age and economic statuses, in the econometric model, as important control variables. If a student is perceived to be weak then he has a much higher probability to end up in a, what those authorities believe to be, compensatory class. The Portuguese case would be to sort them according to their past poor academic record as reflected by their lower baseline scores and retention status (above reference age status) and depressing socio-economic background in a compensatory fashion. Precisely, we observe that class size is negatively correlated with class age dispersion and fraction of students with economic aid (SASE), while positively correlated with fraction of students below or at the

[17] Which means taking out school FE.

reference age and with class average baseline score[18]. This points to the likely possibility that school authorities *believe* that less populated classes are compensatory classes[19]. Otherwise we would not see students with characteristics negatively correlated with achievement such as aged above the reference age, lower baseline score and belonging to low income households grouped in less populated classes. Lazear (2001), as pointed in the literature review, theorizes in this direction, saying that, for *fixed* student quality, decreasing class size allows a greater deal of learning time, hence greater (compensatory) achievement. From here follows the second point to identify class composition effects: the *joint* inclusion of class size and class compositional variables in the model. It is of course tautological to say that if we are interested in identifying class composition effects we must include compositional variables in the model. But it is key to accompany compositional variables with a quantity variable – class size – and, vice-versa, compositional variables if interest is on analyzing class size effects. Missing to jointly include these dual dimensions (quantity and quality) is likely to produce biased effects of each one. The *ceteris paribus* interpretation given to the effect of varying class composition (class size) may be jeopardized given that we are no longer sure that when class composition (class size) changes, class size (class composition) does not. In our case, where weak students are sorted in a compensatory fashion to less populated classes, weak students end up not only in smaller classes but also in classes with larger fractions of students *above* reference age and with SASE status, with higher age dispersion and with lower average baseline score. This is the same to say that class size variation is correlated with class composition variation. Hence both must be explicitly included in the model to avoid omitted variable bias. Other important relevant variables are parents' background and teacher quality. The first one measures the quality of academic assistance a student gets at home and the second the soundness of the teaching offered to him at school. The educational level of the parent with highest educational attainment captures the former. The latter, in turn, is already taken in account by the school dummy which provides a control for each school teacher force quality[20]. The third point one should control for is unobserved student's ability. The dummy variable *Beneficiary of academic support* should help control for it. We assume that teachers when assigning students to such program do so with an acute perception of the true (lower) ability of these pupils. Hence assignment to program is correlated with ability. We stress that the baseline score should also be seen as a control for ability and accumulated knowledge which lends more importance to its inclusion in the model.

[18] Corr(Class Size,Age Dispersion) = -0.15, Corr(Class Size,% SASE) = -0.23, Corr(Class Size,% Below Reference Age) = 0.22 and Corr(Class Size,Class Average Baseline Score) = 0.21, = 0.26, = 0.17, = 0.21 for Portuguese and Mathematics, both for 6th and 9th grades, respectively.

[19] This is the hypothesis of compensatory within-school sorting put forward by West and Wößmann (2006). They point that countries with external exams are the most prone to induce such within-school compensatory schemes. That is the case of Portugal with its (high-stakes) National Exams.

[20] As a dummy it controls for all differences between schools, including this one. Of course, a superior approach would be to include teacher FE, but currently our dataset does not allow it.

Hanushek and Rivkin (2010) provide a theoretical framework that justify the usage of a baseline score as a summary of past factors. These include intrinsic ability (ability is assumed to affect learning at every period), but also the stock of accumulated knowledge. Finally, we instrument class size with school-grade average class size as used by Jürges and Schneider (2004), Wößmann and West (2006) and West and Wößmann (2006). We employ the IV estimation as a final strategy to make sure that within-school sorting endogeneity is tackled, even considering all the controls and variables included in the model. It will be interesting to compare the results under the IV model and under the OLS one with all the above mentioned control variables. The achievement production model to be fit by 2SLS, will be:

$$Y_{icgs} = \delta . BS_{icgs} + \boldsymbol{X}'_{1,icgs} . \boldsymbol{\beta_1} + \beta_2 . \widehat{C_{ic}} + \boldsymbol{X}'_{3,ic} . \boldsymbol{\beta_3} + \gamma . G_{ig} + \boldsymbol{S}'_{is} . \boldsymbol{\alpha} + \varepsilon_{icgs} \qquad (1)$$

Where Y_{icgs} is the Mathematics or Portuguese national exam score of student "i" from class "c", grade "g" and school "s"; BS_{icgs} is his baseline score from the previous low-stakes exam took by him, of Mathematics or Portuguese, respectively; $\boldsymbol{X}'_{1,icgs}$ is a vector containing his individual characteristics[21]; $\widehat{C_{ic}}$ the fitted values (from the 1st stage) of the possibly endogenous class size variable; $\boldsymbol{X}'_{3,ic}$ is a vector containing all class compositional variables[22]; G_{ig} is a grade dummy (6th and 9th graders pooled in the regression); $\boldsymbol{S}'_{is}$ are school dummies and ε_{icgs} is that student's idiosyncratic error term. Given that schools must obey certain rules regarding class formation and face, at each academic year, at each grade, specific cohorts with a given size, then the average grade-school class size must be correlated with the actual class sizes (even though schools sort weaker students to shorter classes, 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). On the other hand, as it is put by Wößmann and 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.*" (page 700). Hence, the 1st stage is:

$$\widehat{C_{ic}} = \hat{d} . BS_{icgs} + \boldsymbol{X}'_{1,icgs} . \widehat{\boldsymbol{b_1}} + \widehat{b_2} . \overline{C_{2,igs}} + \boldsymbol{X}'_{3,ic} . \widehat{\boldsymbol{b_3}} + \hat{g} . G_{ig} + \boldsymbol{S}'_{is} . \widehat{\boldsymbol{a}} \qquad (2)$$

Where the fitted values of class size are obtained from a regression of that variable on all included (BS_{icgs}, $\boldsymbol{X}'_{1,icgs}$, $\boldsymbol{X}'_{3,ic}$, G_{ig} and $\boldsymbol{S}'_{is}$) and excluded instruments ($\overline{C_{2,igs}}$ – average class size of the respective student's grade-school). An OLS version of (1) (i.e. without instrumenting class size) will also be presented to allow the comparison between the OLS and IV approaches.

[21] I.e. it includes: Tertiary (Max), Secondary (Max), Baseline Score, Below Reference Age, Male, CPLP, Internet, SASE and Academic Support. It also includes a generic intercept. See Annex 2 for the meaning of the naming of the variables.

[22] I.e. "% Below Reference Age", "% Males", "% CPLP", "% Internet", "% SASE", "% Academic Support" and "Age Dispersion". See Annex 2 for the meaning of the naming of the variables.

And, for the sake of the analysis, simpler models than that of (1) will also be shown to follow the evolution of the class size coefficient. This is theoretically expected to be negative. A statistically significant positive coefficient can be interpreted as a sign of endogeneity bias not duly tackled. Only under no endogeneity bias can we be confident of the class level variables' point estimates and their policy implications. Hence the first model presented will consist of OLS on equation (1) with class size as the unique class level variable and omitting the Baseline Score control variable:

$$Y_{icgs} = \boldsymbol{X}'_{1,icgs}.\boldsymbol{\beta_1} + \beta_2.C_{ic} + \gamma.G_{ig} + \boldsymbol{S}'_{is}.\boldsymbol{\alpha} + \varepsilon_{icgs} \quad (3)$$

Then we incorporate the Baseline Score to assess its impact on the class size coefficient:

$$Y_{icgs} = \delta.BS_{icgs} + \boldsymbol{X}'_{1,icgs}.\boldsymbol{\beta_1} + \beta_2.C_{ic} + \gamma.G_{ig} + \boldsymbol{S}'_{is}.\boldsymbol{\alpha} + \varepsilon_{icgs} \quad (4)$$

And after that we present the full OLS model by adding the class compositional variables onto (4), i.e. a model like (1) without instrumenting class size. Only then, the most appropriate model – the OLS or IV version of (1) – will be chosen based on a Wu-Hausman endogeneity test. We then allow for greater detail in the most appropriate model in two consecutive steps: firstly, by including interaction terms:

$$Y_{icgs} = \delta.BS_{icgs} + \boldsymbol{X}'_{1,icgs}.\boldsymbol{\beta_1} + \beta_2.C_{ic} + \beta_3.AD_{ic} + \boldsymbol{I'}_{ic}.\boldsymbol{\varphi} + \gamma.G_{ig} + \boldsymbol{S}'_{is}.\boldsymbol{\alpha} + \varepsilon_{icgs} \quad (5)$$

where $\boldsymbol{I'}_{ic}$ is the vector containing interaction terms between the class level variables and their individual dummies counterparts[23]. These interaction terms were included to better understand the possible asymmetric nature of the class composition effects, namely how a fraction of a given type of student in a class affects the student of that type and the one of the opposite type. Secondly, by adding the squares of all class level variables (interacted or not):

$$Y_{icgs} = \delta.BS_{icgs} + \boldsymbol{X}'_{1,icgs}.\boldsymbol{\beta_1} + \beta_2.C_{ic} + \beta_{2,sq}.C^2{}_{ic} + \beta_3.AD_{ic} + \beta_{3,sq}.AD^2{}_{ic} + \boldsymbol{I'}_{ic}.\boldsymbol{\varphi} + \boldsymbol{I'^2}_{ic}.\boldsymbol{\varphi}_{sq} + \gamma.G_{ig} + \boldsymbol{S}'_{is}.\boldsymbol{\alpha} + \varepsilon_{icgs} \quad (6)$$

where $\boldsymbol{I'^2}_{ic}$ is a slight abuse of notation, meaning that the interaction vector contains squared elements, e.g. (Males * "% Males"^{2}). This way we allow each group (e.g. male and female) to have its own polynomial (e.g. w.r.t. % Males and "% Males"^{2}).

Cluster robust standard errors, at the class level, are used in every model.

5 Estimation Results

Table 2 provides, for each achievement measure, the results for the OLS models of equations (3), (4) and (1) in columns (1), (2) and (3), respectively, and of the IV version of equation (1) in column (4), as described in the previous section.

[23] I.e. (Below Reference Age * % Below Reference Age), (Males * % Males), (CPLP * % CPLP), (Internet * % Internet), (SASE * % SASE), (Academic Support * % Academic Support). "Class Size" – C_{ic} and "Age Dispersion" – AD_{ic} were not interacted, hence they are not contained in the interaction term.

The simplest model in column (1) reports a suspicious positive sign on Class Size, statistically significant at the 1% level, for both achievement specifications. Controlling for between-school sorting via the school dummies, for grade specific effects (grade dummy) stemming either from differing difficulty levels of the specific grade-courses' materials or of the grade specific National Exam, for different highest parental academic influences to whom students are exposed to at home and for individual level students' characteristics, is not enough to produce the expected negative sign on Class Size. Indeed this naïve model fails to recognize the likely scenario under which schools aggregate pupils with a poor record into classes of shorter dimension, on purpose, to either allow teachers devote larger shares of their time to each pupil or to place them in a less disruptive environment. This scenario is even more plausible in countries with high-stakes national exams (which is the case of Portugal) that increase accountability pressure, forcing teachers and schools to take special care on weaker, disadvantaged students, see West and Wößmann (2006). Students assigned to larger classes are expected, *a priori*, to score more in exams than those assigned to shorter classes. We are just picking up this expected positive correlation. The following step (column (2) model) introduces Baseline Score as a further individual level control variable. As explained above it should control for within-school sorting as it is information available to school authorities when deciding the composition of classes, which is likely to be taken in account. Given that it may also be correlated with student intrinsic ability and accumulated stock of knowledge we regard it as a very important control. Looking at the class size coefficient, in fact, it is more than halved in magnitude for both achievement measures, making it closer to be negative, though it is still positive and statistically significant at the 1% level. Nevertheless, this points to lowered bias stemming from both unobserved student characteristics and within-school sorting. Next (column 3 model), we add the class compositional variables. The joint presence of class size and composition variables lends credibility to the *ceteris paribus* interpretation of any of those. And, because of this, one can be more confident that each of these coefficients are less biased with respect to the other dimension. This explains why the class size coefficient is further decreased in magnitude, to a point where it is no longer statistically different from zero. Before analysing the coefficients of all class level variables and their policy implications we look at the last model of Table 2 in column (4). It depicts the estimation results using the 2SLS estimator when instrumenting class size with grade-school average class size. Table 3 provides information about the 1st stage regression for the IV model of column (4), for each discipline specification. Basically, it shows that the instrument does predict the possibly endogenous variable as required by the rank condition for IV validity in both specifications. Their F-statistics are well above the rule of thumb of 10. Hence we are not in the presence of a weak instrument. Visual inspection of the class size coefficient across the Mathematics and Portuguese specifications indicates that

Table 2. Regression outputs w.r.t. Mathematics (Mat) and Portuguese (Pt) National Exam Score.

Explanatory Variables	Model							
	OLS						IV	
	(1)		(2)		(3)		(4)	
	Mat	Pt	Mat	Pt	Mat	Pt	Mat	Pt
Class Size	0.011***	0.008***	0.004***	0.002***	-0.001	-0.000	-0.001	-0.003
Below Reference Age	0.54***	0.34***	0.28***	0.18***	0.27***	0.17***	0.27***	0.17***
% Below Reference Age					0.0025***	0.0012***	0.0025***	0.0013***
Age Dispersion					-0.04	-0.04*	-0.04	-0.05*
Male	0.00	-0.21***	-0.07***	-0.15***	-0.07***	-0.15***	-0.07***	-0.15***
% Males					0.0001	-0.0001	0.0001	-0.0001
CPLP	-0.12***	-0.06***	-0.06***	-0.04***	-0.08***	-0.04**	-0.08***	-0.04**
% CPLP					0.0003	-0.0013*	0.0003	-0.0012*
Internet	0.14***	0.09***	0.10***	0.06***	0.09***	0.05***	0.09***	0.05***
% Internet					0.0007**	0.0006***	0.0007**	0.0006***
SASE	-0.20***	-0.11***	-0.14***	-0.07***	-0.12***	-0.06***	-0.12***	-0.06***
% SASE					-0.0024***	-0.0015***	-0.0024***	-0.0015***
Academic Support	-0.62***	-0.39***	-0.37***	-0.24***	-0.39***	-0.25***	-0.39***	-0.25***
% Academic Support					0.0015***	0.0010***	0.0015***	0.0011***
Baseline Score	--	--	✓	✓	✓	✓	✓	✓
Parent Education Dummies	✓	✓	✓	✓	✓	✓	✓	✓
Grade Dummy	✓	✓	✓	✓	✓	✓	✓	✓
School Dummies	✓	✓	✓	✓	✓	✓	✓	✓
Adjusted R2	27.7%	23.5%	46.2%	39.1%	46.2%	39.1%	46.2%	39.1%
N	107 648	106 898	107 648	106 898	100 267	99 528	100 267	99 528

Notes: 1) significance levels: * p<.10, ** p<.05, *** p<.01; 2) S.E. clustered at the class level

Table 3. First Stage information w.r.t. Mathematics (Mat) and Portuguese (Pt) models.

	Adjusted R2	F statistic (robust*)		P-value
Mat	61.4%	F(1, 6030)	1815.33	0.0000
Pt	61.5%	F(1, 5988)	1778.06	0.0000

*adjusted to clustering

it is very similar in size compared to the respective ones of the full OLS specifications in column (3). We formally test the null hypothesis of exogeneity of class size by means of a Wu-Hausman endogeneity test on its coefficient. For the Mathematics specification in column (4) we obtain that the (robust) F-statistic points to a p-value of 0.9978, while for the Portuguese one to a p-value of 0.1668. Both results point to failure of rejection of the null of exogeneity at any conventional significance level. Thus we conclude that the full OLS model of column (3) is the one to keep in mind, since it delivers not only consistent estimates of the class level variables, as well as the most precise. It will be the one to analyse the estimation results[24].

Class size has no effect on achievement. At first sight this seems to contradict the theoretical result by Lazear (2001). One has to bear in mind that that result assumes students avoid synchronizing disruption in class, meaning that disruption due to class size is maximized.

[24] Furthermore, it is not the scope of this paper to disentangle the identified class compositional effects (that will be discussed below) between peer and teacher effects, in a definite and statistical way. Nevertheless we provide what we believe to be the most accurate interpretation of the compositional effects in terms of those two latent effects.

Hence it gives the worst case scenario. In turn, the best case scenario means that all students synchronize their disruption. In this case for a fraction of time $p \in (0, 1)$ of effective learning per student, the overall class time of effective learning will not be p^N but p itself (with p much larger than p^N for N large – say, larger than 14). The real effect of class size on achievement must then lie between those two scenarios. Most likely closer to p than to p^N since it is more realist to assume that classmates cease paying attention to the teacher in a partially and locally synchronized way (depending on their actual spatial distribution within the classroom). All this implies that class size effects may be actually small in magnitude, though always negative (and the estimated slope is indeed negative for both achievement specifications and both full OLS and IV models). The fact that class size coefficient is not significant may be a sign that it would be necessary to record greater class size variation in absolute terms – greater than from 14 to 31 pupils across classes – to econometrically capture a significant effect. Duflo, Dupas and Kremer (2015) find a positive significant effect from *reducing* class size in a Kenyan experiment on primary schools under some treatment conditions, but the variation they record in the experiment is quite considerable – halving class sizes of about 80 students to 40.

Students at or below their reference age seem to have an advantage in terms of achievement, compared with those above it, of 2 to 3 decimal points, depending on the specification. Additionally, the higher the fraction of students, in a class, below the reference age, the higher the achievement of classmates (irrespectively of belonging to the below or above reference age groups), in both Mathematics and Portuguese specifications. A ten percentage point (p.p.) increase in that fraction of students in a class leads to an increase of scores by about 0.01 to 0.03 decimals. This finding fits the idea that it is beneficial for a student to belong to a class majorly composed of classmates that are achievement oriented as this majority allows greater overall class time of effective learning. Note that one could argue that increasing the fraction of students below the reference age would be as decreasing the age dispersion present in a given class. After all, classes with less students below the reference age are classes with more students with at least one retention in their past, possibly more than one. I.e. classes with greater age dispersion of students. This is not a one-to-one map since it is possible to have classes with many students above the reference age that are themselves close in age. Inclusion of a measure of age dispersion clears these confounding effects. The mean absolute age deviation in a class produces no (Mathematics) or faintly (Portuguese) significant effect to achievement. Given the negative point estimates it seems that if any effect exists it is that of increasing age dispersion hurting achievement – one extra year of age dispersion in class decreases achievement by 0.4 decimals. By contrast, the effect of fraction of students below the reference age is much more precisely estimated. In turn, being male means scoring less 0.7 to 1.5 decimals in the exams of Mathematics and Portuguese, respectively. This reflects a specific

gender achievement difference given we are already taking in account parent, school and class characteristics. OECD (2015) points that boys study less and read less complex texts for own amusement (e.g. fiction) than girls at the age of 15 (which maps to our 9th graders, in general). These behavioral facts may explain why boys underperform girls in general, and in language subjects in particular. Even more interesting, in our perspective, is the rather close to zero estimated coefficient on fraction of males in class. Not statistically significant at any conventional level, and small in magnitude, that coefficient says that having proportionally more boys in class seems to not translate in losses or gains via behavioral interactions within class. Next, being born in a CPLP country translates, *per se*, to lower achievement by 0.4 to 0.8 decimals. This may be the result of either exposition to a less demanding schooling system in the origin country or to the fact that many of these students still speak a different dialect, not Portuguese, at home. Either case would be reflected in an increased difficulty to absorb the taught materials at school, reflected in lower grades in the national exams. Nevertheless the difference in achievement is not extraordinaire. The different spoken language hypothesis is reinforced with the fact that in the Portuguese specification there is mild evidence (at the 10% significance level) that a given student performance is deteriorated as the presence of CPLP born classmates increases. For an increase of 10 p.p. in the CPLP born classmates, a student scores less 0.01 decimals in the Portuguese exam. Indeed, if CPLP born students are disadvantaged in terms of Portuguese proficiency, then placing a given student in a class with many of these classmates will be synonym of placing him in a class where, most likely, the teacher will have to slowdown the class pace in order not to lose the majority of less Portuguese proficient pupils. Hence, we interpret the presence of the CPLP born student not as a source of class turbulence, but as a trigger of slowly paced lessons, which is justified by the teachers' goal to act in a compensatory fashion. Having home access to the internet means to score more 0.5 to 0.9 decimals, depending on specification. At the individual level, access to internet can be a rich source of academic content or a distraction in the form, for example, of chatting and gaming. The statistically significant estimate, at any conventional level, on the internet dummy leads us to believe that, on average, students are actually benefiting from it. At the class level, students benefit from being placed in classes where more and more classmates have home access to the internet, yet the precisely estimated effect is rather small: an increase of 10 p.p. of students with access to it leads to an increase of achievement of about 0.0065 decimals in both specifications, to any given pupil. Students flagged as of low income families (SASE students) score less 0.6 to 1.2 decimals compared to non-SASE students. It seems that income – and all the cultural and educational goods and services it can provide – plays a role for student achievement. More so with respect to Mathematics. Thus income inequality seems to be a source for achievement inequality, which in turn is likely to mean income inequality for the next generation. Sadly,

greater proportions of low income students in class further decreases achievement of any given student. A ten p.p. increase in SASE students within a given class produces lower exam scores of about 0.015 to 0.024 decimals for any given student. The presence of students with possibly lower own expected returns to education – if their expectations are explicitly or implicitly anchored on the low income level of their parents then they will expect to collect a low level of labor income during their lifetime, lowering their expected return to education – may be synonym of higher class disruption. Their focus may be biased toward present matters other than school success (which can be seen as a current sacrifice of leisure for the benefit of future career and earnings gains), relatively to non-SASE students who may perceive higher gains from academic achievement. One way to test and control for this hypothesis would be to include in the model variable(s) that would proxy school engagement and\or what each student expects to profit from thriving in school. Unfortunately we do not possess such data. Finally, if a student is flagged as having troubles in the current academic year (that is why teachers direct them to the academic support program) indeed they end up scoring much less in the exams: 2.5 to 3.9 decimals below than others without such status[25]. Paradoxically, as the fraction of these struggling students increase in a given student's class, there is an achievement gain for the latter: 0.01 to 0.015 decimals more per extra 10 p.p. of classmates with academic support. One would expect that as more classmates are flagged as having difficulties in learning this would translate to lower achievement for any student placed there, either because teachers slow down the pace of lessons or because classmates turn more disruptive. The former as a way teachers have to not "lose" their audience, the latter due to perceived lower probability to pass at the end of the academic year, by the student, making meaningless that effort. We will come back to this below.

So far the analysis constrains the effect of a given class compositional dimension to be the same irrespectively of the student status. For example, we were constraining the effect to a student's achievement of an increase in the proportion of below reference age students, in a given class, to be the same to both below and above reference age students in that class. This approach will not capture possible asymmetries that may exist across different types of classmates given a common marginal change in the composition of a class. Those asymmetries most likely reflect different peer effects or different teacher responses to different class compositions that different types of students are exposed to. Thus we present a model with interactions between the compositional and the respective individual level variables. We also expand the analysis to accommodate the presence of non-linear effects across the class level variables. The former model is that of equation (5) whereas the latter is that of equation (6).

[25] Note that neither the estimated coefficient for academic support nor for SASE programs' statuses should be interpreted as the consequence of program participation. Those statuses only help to identify where students stand in terms of household income and current academic struggle.

Both models' results are presented in Table 4, in columns (1) and (2), respectively. Meanwhile, Figure 1 plots the multiple profiles between class level variables and the predicted level of achievement, given the results of Table 4[26].

An F test on the joint hypothesis that both the first and second order terms on class size are not significant indicates that we are not able to reject this hypothesis at any conventional level (p-value of 22.0% and 28.2% for Mathematics and Portuguese, respectively). Hence the class size prediction profiles in Figure 1 depicts a *non-significant* linear relation between this variable and any of the achievement measures[27]. This result reinforces that class size is not an important achievement factor, at least for the variation recorded in our sample.

Contrasting with the results shown in Table 2, allowing differing slopes for fraction of classmates below the reference age for students below or above that age, as in column (1) of Table 4, one learns that only for the former type of student there is a gain in being placed in a class with an increasing proportion of classmates below the reference age. Across both measures of achievement in column (1), that coefficient is highly significant, but with double magnitude under Mathematics. On the other hand, the effect of increasing that same fraction of students on a pupil above the reference age is not significant for Mathematics, but significant at the 5% level for Portuguese and negative (meaning that those above the reference age do not profit, possibly even being harmed, by an increasing presence of classmates below the reference age). Moreover, there are significant non-linear effects with respect to this variable. The two F tests on the joint significance of first and second order terms of *% Below Reference Age* for both types of students and the one on the joint significant difference between the two categories' first and second order terms yield the existence of non-linear effects that are significantly different across them (at least, at the 5% significance level). The plotted prediction profiles of Figure 1 (row 2) show exactly the differences at stake: for those below the reference age it is profitable to be placed in a class with an increasing presence of classmates of his own type (though the curvatures are different across subject), while for those above the reference age it is harmful to be placed in a class with an increasing proportion of classmates of the opposite type. These findings come in line with Duflo, Dupas and Kremer (2011) who found that high (≈ below

[26] In order to avoid presenting non-significant non-linear effects we run an auxiliary regression of (6) (not shown) after performing joint significance tests, either testing, within a given compositional variable, the joint significance of the first and second order terms or the joint significance of the terms between groups. The former to see if indeed a second order specification is correct, the latter to see if indeed there are differences between groups (e.g. between males and females for a marginal change in *% Males*). These tests are always conducted on the *original* model present in column (2) of Table 4, for each measure of achievement.

[27] The inclusion of the linear term in the auxiliary regression is justified given its presence is necessary as an important control variable, as discussed above. This auxiliary regression contains all the linear terms of every class level variable for the same reason, even if it is not individually or jointly significant within column (2) model.

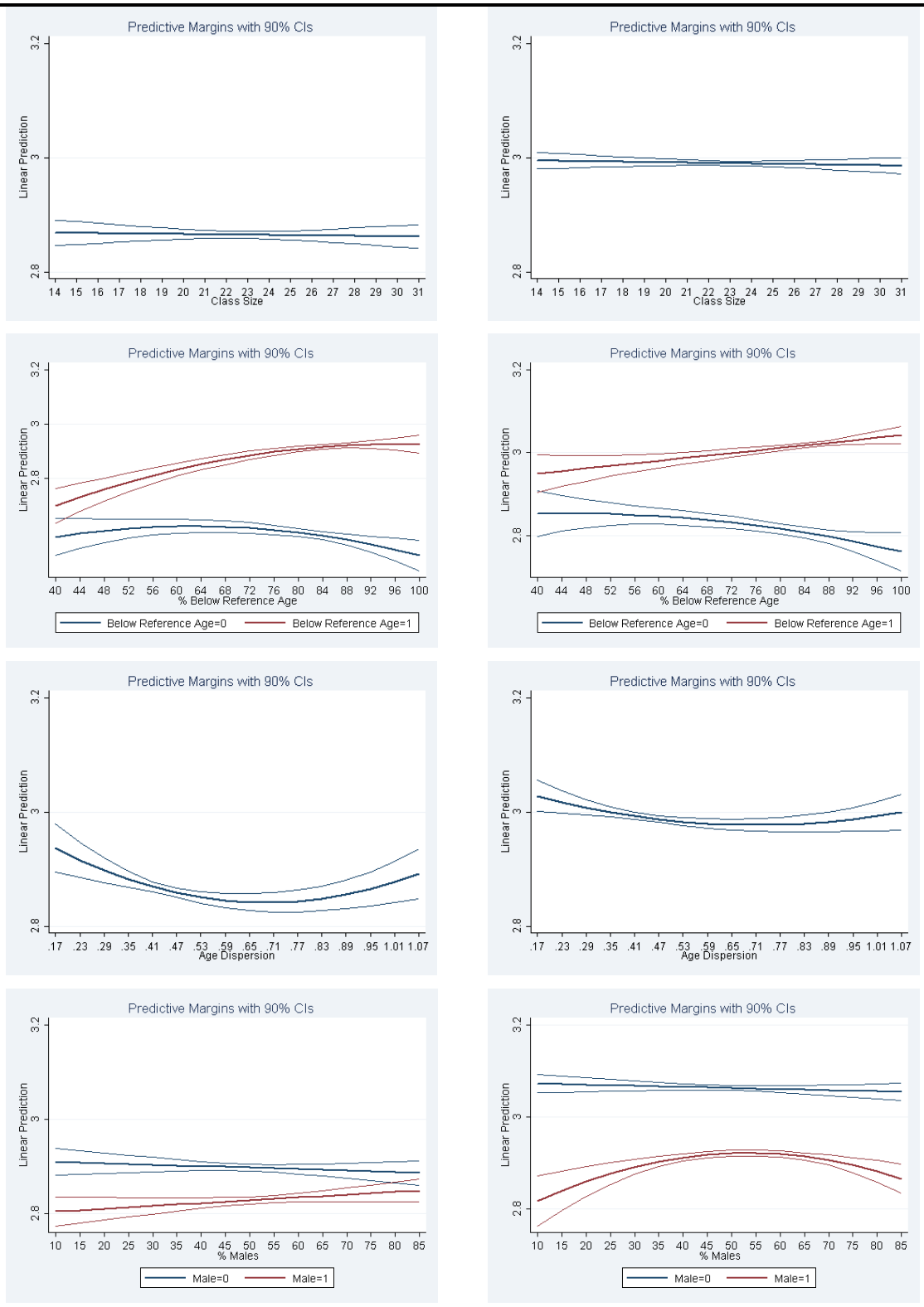
Table 4. Regression outputs w.r.t. Mathematics (Mat) and Portuguese (Pt) National Exam Score including interaction and non-linear effects.

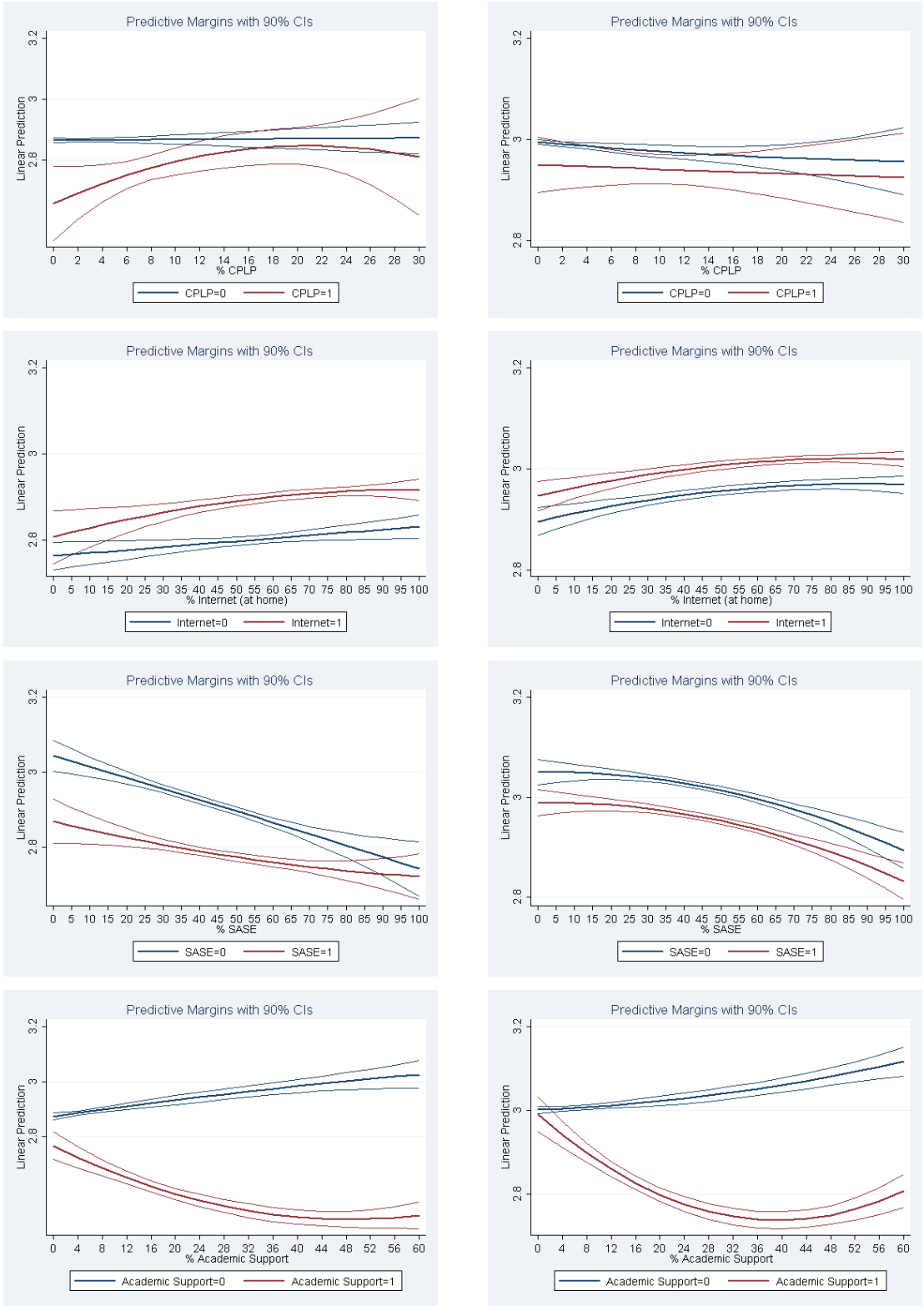
Explanatory Variables		Model			
		OLS			
		(1)		(2)	
		Mat	Pt	Mat	Pt
Class Size		-0.000	-0.000	0.024*	-0.015
Class Size Sq.		--	--	-0.0005*	0.0003
Below Reference Age		-0.05	-0.07*	-0.07	0.09
Below Reference Age = 1	% Below Reference Age	0.0035***	0.0019***	0.0136***	0.0016
	% Below Reference Age Sq.	--	--	-0.0001**	-0.0000
Below Reference Age = 0	% Below Reference Age	-0.0010	-0.0015**	0.0093	0.0027
	% Below Reference Age Sq.	--	--	-0.0001*	-0.0000
Age Dispersion		-0.03	-0.04	-0.49***	-0.24**
Age Dispersion Sq.		--	--	0.3561***	0.1702**
Male		-0.12***	-0.16***	-0.15*	-0.26***
Male = 1	% Males	0.0006	0.0000	0.0012	0.0062***
	% Males Sq.	--	--	-0.0000	-0.0001***
Male = 0	% Males	-0.0003	-0.0003	-0.0010	0.0023
	% Males Sq.	--	--	0.0000	-0.0000
CPLP		-0.14***	-0.05	-0.21***	-0.03
CPLP = 1	% CPLP	0.0058*	-0.0005	0.0185	-0.0041
	% CPLP Sq.	--	--	-0.0005	0.0001
CPLP = 0	% CPLP	0.0002	-0.0013*	0.0014	-0.0021
	% CPLP Sq.	--	--	-0.0001	0.0000
Internet		0.08***	0.07***	0.04	0.04
Internet = 1	% Internet	0.0008**	0.0005**	0.0022*	0.0019**
	% Internet Sq.	--	--	-0.0000	-0.0000*
Internet = 0	% Internet	0.0006*	0.0009***	0.0005	0.0012
	% Internet Sq.	--	--	0.0000	-0.0000
SASE		-0.20***	-0.05***	-0.17***	-0.12***
SASE = 1	% SASE	-0.0014***	-0.0015***	-0.0024*	0.0019*
	% SASE Sq.	--	--	0.0000	-0.0000***
SASE = 0	% SASE	-0.0030***	-0.0014***	-0.0029**	-0.0006
	% SASE Sq.	--	--	-0.0000	-0.0000
Academic Support		-0.20***	-0.13***	-0.11***	-0.01
Academic Support = 1	% Academic Support	-0.0037***	-0.0024***	-0.0110***	-0.0129***
	% Academic Support Sq.	--	--	0.0001***	0.0002***
Academic Support = 0	% Academic Support	0.0028***	0.0018***	0.0032***	0.0006
	% Academic Support Sq.	--	--	-0.0000	0.0000
Baseline Score		✓	✓	✓	✓
Parent Education Dummies		✓	✓	✓	✓
Grade Dummy		✓	✓	✓	✓
School Dummies		✓	✓	✓	✓
Adjusted R2		46.3%	39.1%	46.3%	39.2%
N		100 267	99 528	100 267	99 528

Notes: 1) significance levels: * p<.10, ** p<.05, *** p<.01; 2) S.E. clustered at the class level

reference age) and low (≈ above reference age) achievers profit from homogenous classes. As in their case, here both categories should benefit from better tailored teaching given their own specificities or engagement toward school achievement. Moving on to age dispersion, we see that indeed we need to introduce a second order term to capture its effects on achievement, especially for Mathematics. Whereas in Table 2 results we were unable to get strongly significant effects with just a linear effect, now, with a second order term, we get stronger joint significance of both terms (at the 5% level for Mathematics and at the 10% for Portuguese).

Figure 1. Profiles of predicted achievement given class composition (left column – Mathematics' score prediction; right column – Portuguese's score prediction).





Across both achievement measures the minimizer level of age dispersion is about 0.7 years of mean absolute age deviation in a given class. It seems that as age dispersion decreases below that threshold we are back to the case of a homogenous class (this time with respect to the age composition of classmates) which is again associated with an increase in achievement.

Unexpectedly, also for increasing levels of age dispersion, for a given class, beyond that threshold we obtain a positive effect on achievement. We regard the former finding as a result of positive peer and teacher effects due to class homogeneity. While the second (less precise given the enlarged confidence interval in that section of the profiles) as a result of diminished (negative) peer effects. After all, classes with large age dispersion are classes with very young and very old students relatively to the class mean age (e.g. for age dispersion = 1 we have that, on average, each student is 1 year old younger or older than the class mean age, i.e. a class where half of the students are 1 year younger and the other half is 1 year older than that class mean age would produce that level of dispersion). It may be the case that, for large enough age differences between those two groups within the class, they start acting as two distinct subclasses, where disruptive behaviours are contained within each one. If this is the case then overall effective time of learning should be greater than in a class where age dispersion is about 0.7 (which translates a class where few pupils are expressively younger or older than the majority which may create an environment propitious to overall disruption). In turn, the fraction of male students continues to yield no significant impact on achievement, now for both male and female classmates, as column (1) specifications indicate. However, with non-linear effects present in column (2), the joint significance of the first and second order terms on *% Males* for male student can be established at the 1% level, but only for the Portuguese achievement measure[28]. This is the unique identified effect relatively to class gender composition (the individual level gender dummy still points to significant lower achievement due to the student being male). Additionally, the plotted Portuguese predicted score profile for varying fractions of males in class shows that the implied optimizer is around 53% males in class. This is to say that it would be optimal, at least for Portuguese achievement, to evenly split the classes in terms of gender. The mechanisms through which male students perform poorly in the presence of too few or too many other male classmates are unclear from the results presented in this paper. In any case, we consider that behavioural and possibly sociological reasons may be driving this result. Recall that the individual level gender dummy points to lower achievement for males, especially in the Portuguese achievement measure. Assuming that, in general, male students care less about the language subject relatively to girls – which is in line with what was said earlier about the fact that boys read less complex texts as girls do, for ages similar to those of the sample used in this paper – then *too few* males in a class for a given male pupil may mean Portuguese lessons not suited toward what could still interest him in that subject thus lowering his effort level in that subject, whereas *too many* male classmates may mean higher disruption among males (after all, Portuguese lessons are not their highest interest by assumption) that dominates the benefit of higher level of effort due to hypothetically better suited lessons. With

[28] For female student those two terms are not jointly significant, at any conventional significance level.

respect to the proportion of CPLP born students, column (1) estimates indicate that the previously reported (Table 2) negative, mildly significant, effect to Portuguese achievement is driven by the non-CPLP born students who are the ones actually affected by that fraction of students. Regarding Mathematics we now see also a mildly, but positive, significant effect from higher fractions of CPLP born students to students of that type. The former finding fits the earlier reasoning that as the fraction of CPLP born students increase in a class, the more the Portuguese lessons will be focused to the needs of this audience, at the cost of those born in Portugal. The latter comes in line with Hoxby (2000a) who finds that peer effects are stronger and beneficial within cultural groups. Furthermore, allowing for non-linear effects in this compositional dimension we arrive at the same, mildly significant, results. For Mathematics achievement, the first and second order terms are jointly significant (just at the 10% level) only for the CPLP born student. For Portuguese, those two terms are jointly significant (at the 10% level) only for the non-CPLP born student, implying, for him, an achievement loss for increasing shares of CPLP students. The plotted achievement profiles show these mildly significant relations. In particular, CPLP born students profit, but at a decreasing rate, with a higher share of the same type of classmate, at least up until a share of about 20%, in Mathematics. Whether it is a maximizer is difficult to establish given the mild significance of the parameters and the enlarged confidence interval from then on. Concerning the proportion of students with Internet at home, column (1) shows that higher fractions of such students in class produce better achievement results for each type of student, i.e. for the one that has such good and for the one that does not. This mimics the finding from Table 2. There may be going on not only potential benefits from increased access to contents in the web for the individual student (positive and significant individual level dummy for Internet) as well as benefits from enhanced communication between classmates (positive and significant *% Internet* class level variable). Even students without home access to internet may be benefiting from it given the possibility that colleagues possessing it may share its access with them for group assignments or group study activities. F tests for the joint significance of the first and second order terms across both categories of students reveal that those terms are jointly significant for the two categories under the Portuguese specification and for students with internet under the Mathematics specification. On top of that, under the Portuguese specification, we are not able to reject the null (at any conventional level) that the first and second order parameters are jointly the same across the two categories of students. Hence the plotted achievement profile for this specification depicts the same curves for both categories of students, shifted by the individual effect of having or not Internet at home (in favour for those who have it). The curvatures for both categories of students under the Portuguese plot and for those with internet under the Mathematics one depict the positive, but marginally decreasing, impact of increasing the proportion of students with web

access. It is possible that as the fraction of classmates with internet increases, the likelihood to realize achievement gains from internet sharing decreases. The "first" ones sharing that good may make the greatest impact. Exposure to classmates belonging to low income households, as discussed using Table 2 results, was predicted to harm a given generic student. Indeed that result carries on to Table 4 results, but now we are able to see that larger fractions of SASE students impacts negatively on both categories of students, i.e. on SASE and non-SASE students. In fact, from column (1) we further see that the negative marginal impact is double for non-SASE students than that for SASE students with respect to Mathematics achievement. For Portuguese achievement the negative marginal effects are similar between the two categories of students and close to the one for SASE students in the Mathematics specification. Higher fractions of students that are likely to value less class learning time are portions of students likely to be more disruptive, at the cost of either type of student. And that is what is being mirrored by these results. The introduction of non-linear effects done in column (2) does not alter the picture. For both categories of students F tests for the joint significance of the first and second order terms deliver p-values below 1% for each measure of achievement. Given that non-linear effects are significant we then test if the two terms are significantly the same between the two categories of students. And in fact they are for Portuguese (null not rejected even at the 10% level), but not for Mathematics (null rejected at any conventional level). Hence, the respective achievement profiles shown in Figure 2 depict the same curve for both categories of students under the Portuguese achievement measure, but two distinct curves (concave for non-SASE students, convex for SASE students) under the Mathematics one. The general case seems to be that it is *increasingly* harmful to have higher shares of low income classmates, in a given class. The exception is the SASE student which is harmed, at a *decreasing* rate, with increasing shares of SASE students, in Mathematics. Finally, allowing for asymmetric and non-linear effects does not change the puzzling fact that students *not* signalled as troubled during the academic year benefit from an increasing proportion of struggling students. Looking at column (1) coefficients, the student flagged as having troubles to succeed during the academic year is, as what could be expected, harmed by larger fractions of struggling classmates. This expectation can be based on the assumption that due to having troubles to succeed (or due to the latent reasons making them prone to such troubles) those students are more disruptive or on teachers slowing down the pace of lessons. But these hypothesis should produce a negative effect, from higher levels of *% Academic Support*, on those that are not flagged as struggling. Those should also be affected by class disruption or less paced lessons. That is not supported by the empirical results. The introduction of non-linear effects contributes with yet another puzzling result. The usual F tests indicate very significant non-linear effects across these last two categories of students and very significant different first and second order parameters on *% Academic*

Support between them. But looking at the Portuguese profile, there is a significant inversion in the tendency of lower achievement given increases in the fraction of struggling students. For a high enough fraction of struggling students (more than 40%) this type of student seems to be able to profit out of it. Again, this does not fit our hypothesis that a class containing a large proportion of struggling students impacts negatively in both types of students' achievement levels.

6 Conclusions and Policy Implications

This work estimates class composition effects within the Portuguese public schools using the *MISI* student level dataset. Contrasting with non-significant class size effects, which is in line with Hoxby (2000b), Jürges and Schneider (2004), Wößmann and West (2006) and West and Wößmann (2006), we find that some class compositional dimensions cause different levels of pupil achievement. These effects are, in turn, in many cases, significantly asymmetric (and in fewer cases also non-linear) between the relevant categories of students.

Irrespectively of the measure of achievement we look at, the significant parameters estimated with respect to the fraction of students below the reference age point to potential achievement gains under class homogeneity: classes increasingly composed by pupils above the reference age (i.e. classes decreasingly composed by below reference age students) help students with this status to perform superiorly, whereas classes majorly composed by pupils below the reference age contribute to increase the performance of below reference age students. Shifting a pupil aged above his grade reference age from a class where merely 10% of the classmates share that characteristic (thus 90% are below the reference age) to a class where there is a significant presence of classmates sharing that characteristic, say 40% of them (thus with 60% below the reference age), should make him achieve a higher Mathematics' or Portuguese score by about 0.7 or 0.5 decimals[29], respectively. Under monotonic decreasing scores w.r.t. *% Below Reference Age* (supported by the Portuguese achievement profile, and not entirely possible to reject for the Mathematics' one) those gains should maintain even for fractions smaller than 60%. In other words, students above the reference age can still profit if allocated to classes uniquely composed by such category of classmates. Complementarily, shifting a below reference age student from a class with 60% to another with 90% below reference age classmates will allow him to score more 1 and 0.5 decimals in Mathematics and Portuguese, respectively. Better achievement results under class homogeneity can be explained by better tailored teaching as Duflo, Dupas and Kremer (2011) or Collins and Gan (2013) noted.

[29] Using the parameters from the auxiliary regression. These and following stated achievement gains can be visually inspected in Figure 1 relevant profiles.

Moreover, targeting homogenous classes with respect to reference age status compounds to also targeting classes with lower levels of age dispersion. After all, pupils below the reference age are pupils of the same cohort, generally sharing the same year of birth. Those above the reference age would be grouped with classmates born one or more years before the reference age year of birth, making that class to have, in principle, one less possible year of birth represented in there – the reference one – lowering age dispersion. Assuming the non-linear parametrization of this variable as the most accurate, lower levels of age dispersion cause modest achievement gains (see Figure 1 for the non-linear case) in both measures of achievement (especially in Mathematics). These gains will compound with those from homogenizing classes according to reference age status.

Irrespectively of his own income status a student is harmed if placed in a class with an increasing proportion of low income students. Displacing a low income student from a class with 20% low income classmates to one with 80% makes him to score less 0.9 and 1.1 decimals in Mathematics and Portuguese, respectively. The same class displacement exercise with respect to a non-low income student yields achievement losses, for him, of 1.8 and 1.1 decimals in Mathematics and Portuguese, respectively. Peer effects can be an explanation: poorer pupils may fail to recognize the importance of school success, hence disrupting the class relatively more often, in disfavor of everyone present there. Two policy implications are in place given these facts. On one hand school authorities should spread the most low income students throughout the relevant grade classes. This would ensure the inexistence of classes populated by too many students likely to not grasp the complete actual benefit of exceling at school, hence likely to be relatively more disruptive. Taking as reference for a typical school population the 40% and 44% figures of SASE students as depicted in Table 1 for the whole sample of 6^{th} and 9^{th} graders, then this typical school should produce classes each containing those same proportions of low income students. Looking at the distributions of the share of classmates with SASE status shown in Annex 4, one realizes that, although a decent number of classes presented shares close to those figures, roughly almost half of them recorded shares higher than 50%. Students placed in these classes may well have been harmed just by the placement policy of the school. Curiously, and as noted in the Data section, the distributions show an abnormal frequency of classes reporting zero to one SASE classmates. This suspicious sorting of some non-SASE students to classes completely void of SASE students may reflect strategic behavior from the parents of the former who, by pressing school authorities to make arrangements to allow their children to seat in a class free of low income classmates, anticipate the negative achievement consequences from the presence of the latter type of classmate. Even if such kind of parent behavior can be seen as rational schools should not allow such free SASE student classes to exist. It imposes the existence of other classes with an inflated proportion of such

students, at the cost of either type of student present there. On the other hand, if indeed it is true that low income students fail to recognize their full future potential benefits of exceling at school[30] (propelling, as discussed above, a higher propensity to disrupt the class as a byproduct), then there is room for policies helping low income students (and their parents) to realize them. Information disclosure about actual labor market benefits related to increased levels of education can raise awareness among that subpopulation of students and parents concerning their actual cost of opportunity from not exceling at school. And, as a byproduct, induce lower levels of class disruption caused by them.

Gender class composition results point to no effect from varying the fraction of males present in class to either a girl or a boy pupil. The only exception comes from the non-linear effects model using the Portuguese achievement measure, which yields an optimal proportion of males in class of roughly 50% for the male student achievement. Boys were the ones identified as being benefiting from more female classes in math by Hoxby (2000a), but in this paper we find that it is true not for math but for Portuguese and only up to the point where females are about half of the class.

Regarding class composition by place of birth it is the case that pupils born in Portugal seem to be somewhat hurt by placement in classes with increasing proportion of CPLP born students, at least in Portuguese achievement. In turn, more CPLP born students in class seems to benefit the CPLP born student himself. Nevertheless these effects should not be seen as definitive since they are only mildly significant. Taking the estimated effects at face value schools should group students according to their place of birth status as this should improve the CPLP born student in Mathematics and the non-CPLP born student in Portuguese.

Finally, the proportion of classmates with home access to the internet contributes to higher scores in both Mathematics and Portuguese for both the student with it and the one without it. The implication is then to produce 6th and 9th grade classes containing a proportion of classmates with home Internet around the one existing in the overall population. This way, this type of student is spread across classes equally, making everyone benefiting from similar proportions of this "good".

All in all the results obtained in this paper point to the conclusion that schools should purposefully sort students across classes in *different ways depending on their characteristics*. Whereas Collins and Gan (2013) concluded in favor of segregating low and high achievers, across classes, for the benefit of both types, we, in turn, similarly conclude that also students registering at least one retention (thus falling to the above reference age status) and those never registering a retention should be grouped *homogeneously* in different classes, for the sake of the

[30] See Portugal (2004) for a reference on the estimated actual premiums for higher levels of education completed, in Portugal.

achievement levels of both. Nevertheless our results also point to the need of sorting students *heterogeneously* along the household income and home access to the internet dimensions, and, to a lesser extent, along the gender dimension too. Sorting students *cleverly* across classes should improve their achievement levels, more so than reducing class size for which we did not find a significant effect. And, on top of being superior, with respect to student achievement, to class size reduction, wise student sorting is also cheaper. It is just a matter of wisely sorting students across an existing number of classes, for a given level of teacher payroll expenditure.

There is scope for future research. Methodologically, it would be an improvement to formally address both student and teacher unobserved heterogeneity. A panel data framework would allow to include student and teacher fixed effects yielding a more consistent estimation of the class compositional effects. Nonetheless the inclusion of a baseline score should control for student unobserved ability a great deal insofar they are strongly correlated. Additionally, controlling for teacher unobserved quality would allow to interpret the class compositional effects as if teachers had been randomly assigned to classes. If teacher assignment to a given class is somehow related with his unobserved characteristics, then class composition and teacher characteristics may correlate. If this is true then the class compositional effects may be biased. Newer versions of *MISI* will hopefully have that structure in the near future.

References

Akerhielm, K. 1995. "Does class size matter?". Economics of Education Review, 14(3), 229–241.

Checchi, Daniele. 2006. "The supply of education". In *The Economics of Education: Human Capital, Family Background and Inequality*, 84-105. New York: Cambridge University Press.

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, Courtney and Gan, Li. 2013. "Does Sorting Students Improve Scores? An Analysis of Class Composition". *National Bureau of Economic Research*, Working Paper No. 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, Esther, Pascaline Dupas and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya". *American Economic Review*, 101: 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, Eric. 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, Eric. 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 Series, (Working Paper 15816).

Hoxby, Caroline. 2000a. “Peer Effects in the Classroom: Learning from Gender and Race Variation”. *NBER Working Paper*, 7867.

Hoxby, C. M. 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. 2001. “Educational production”. *Quarterly Journal of Economics*, 116 (3): 777–803.

OECD. 2015. “The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence, PISA”, OECD Publishing.

Portugal, P. (2004). “Myths and facts regarding the portuguese labour market - the tragic fate of college graduates. *Economic Bulletin, Bank of Portugal*, (March), 69–76.

Pritchett, Lant and Filmer, Deon. 1999. “What educational production functions really show: a positive theory of educational spending”. *Economics of Education Review*, 18 (2): 223–39.

West, M. R., & Wößmann, L. 2006. “Which School Systems Sort Weaker Students into Smaller Classes? International Evidence”. European Journal of Political Economy, 22(4)

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

Appendix

Annex 1. Countries under CPLP category.

Angola
Brazil
Cape Verde
Guinea-Bissau
Mozambique
Sao Tome and Principe
East Timor

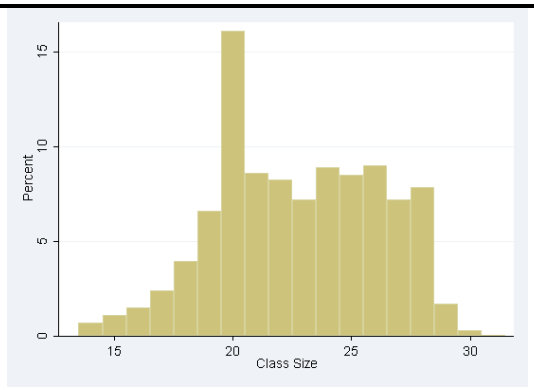
Annex 2. Codification.

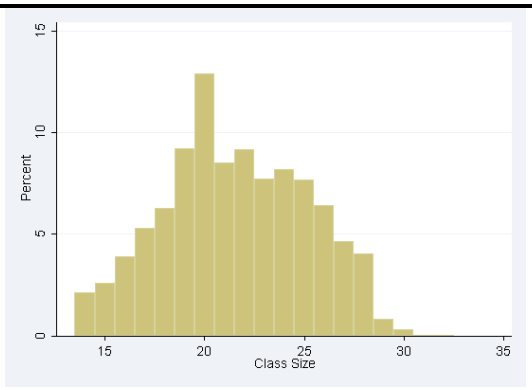
Variable Name	Category	Code
National Exam Score	High Stakes	Score
	Low Stakes	Baseline Score
Parent with the Highest Academic Background	Secondary Education	Secondary (Max)
	Tertiary Education	Tertiary (Max)
Age Dispersion		Age Dispersion
Below Reference Age	< or =	Below Reference Age
Gender	Male	Male
Place of Birth	CPLP country	CPLP
Home access to the Internet	if yes	Internet
Beneficiary of Socio-Economic Support	if yes	SASE
Beneficiary of Academic Support	if yes	Academic Support
Class Size		Class Size
Fraction of Students Under or At the Reference Age		% Below Reference Age
Fraction of Male Students		% Males
Fraction of CPLP born Students		% CPLP
Fraction of Students with Internet		% Internet
Fraction of SASE Students		% SASE
Fraction of Students with Academic Support		% Academic Support

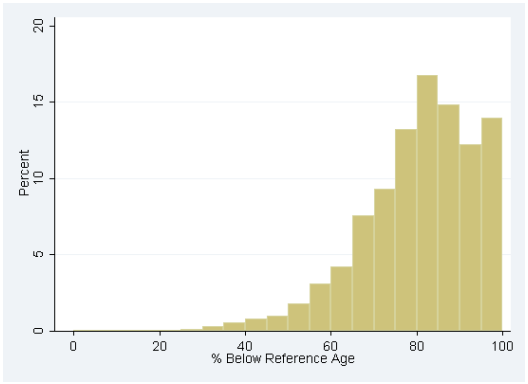
Annex 3. Descriptive statistics under the appropriate sample of students.

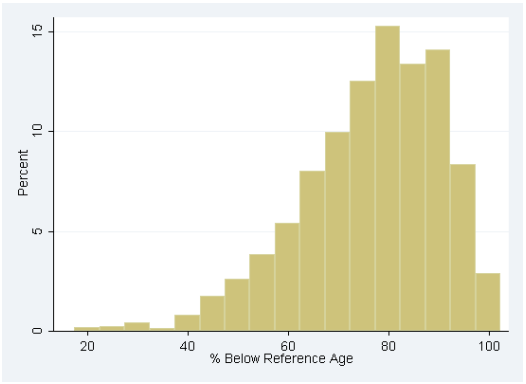
		6th Grade - Portuguese National Exams					9th Grade - Portuguese National Exams				
		N	Mean	Std.Dev.	Min	Max	N	Mean	Std.Dev.	Min	Max
Individual Level Variables	Score	65,054	3.1	0.8	1	5	41,866	2.8	0.7	1	5
	Baseline Score	65,054	0.0	0.8	-2.5	1.6	41,866	0.2	0.7	-2.4	1.7
	Tertiary (Max)	65,054	0.18	0.39	0	1	41,866	0.17	0.37	0	1
	Secondary (Max)	65,054	0.48	0.50	0	1	41,866	0.46	0.50	0	1
	Reference Age	65,054	0.88	0.33	0	1	41,866	0.82	0.38	0	1
	Male	65,054	0.51	0.50	0	1	41,866	0.48	0.50	0	1
	CPLP	65,054	0.02	0.13	0	1	41,866	0.02	0.13	0	1
	Internet	65,054	0.60	0.49	0	1	41,866	0.72	0.45	0	1
	SASE	65,054	0.44	0.50	0	1	41,866	0.39	0.49	0	1
	Academic Support	65,054	0.10	0.30	0	1	41,866	0.15	0.35	0	1
Class Level Variables	Class Size	3,989	23	3	14	31	2,575	22	4	14	32
	% Reference Age	3,964	80	14	0	100	2,403	77	14	17	100
	% Males	3,989	52	11	10	100	2,575	49	12	6	87
	% CPLP	3,986	3	6	0	55	2,575	3	6	0	57
	% Internet	3,989	55	25	0	100	2,575	68	25	0	100
	% SASE	3,989	48	19	0	100	2,575	42	19	0	100
	% Academic Support	3,989	12	15	0	94	2,575	16	21	0	96
	Age Dispersion	3,964	0.6	0.2	0.2	1.7	2,403	0.5	0.2	0.2	1.3

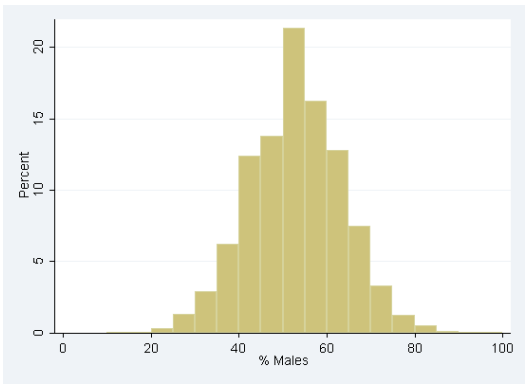
Annex 4. Distributions of class level variables - 6th (left column) and 9th (right column)) grades' classes.

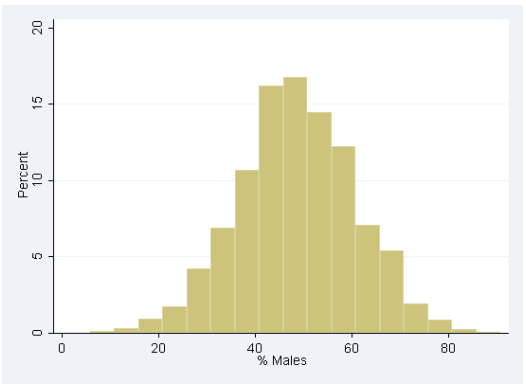


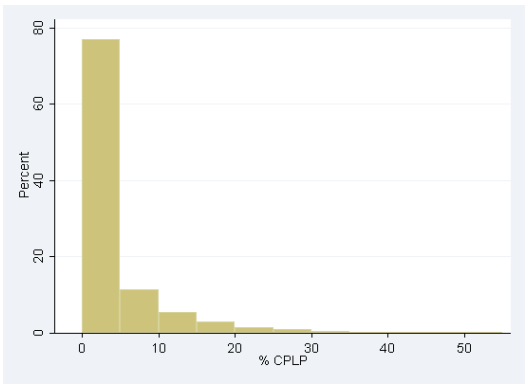


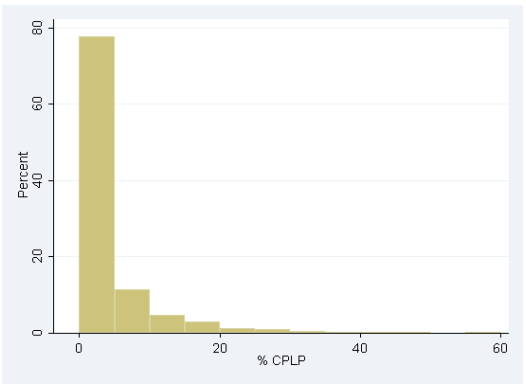


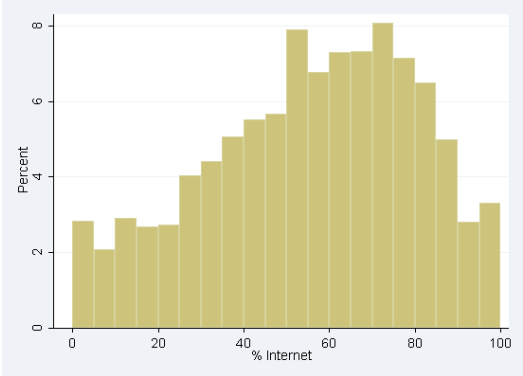


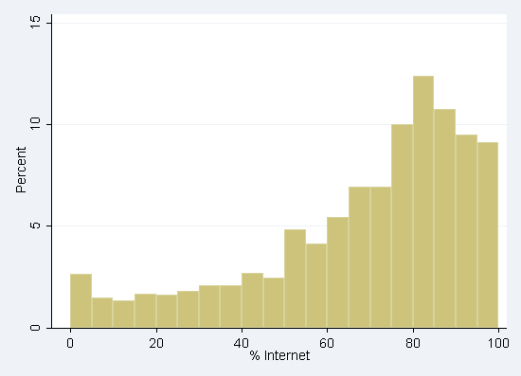


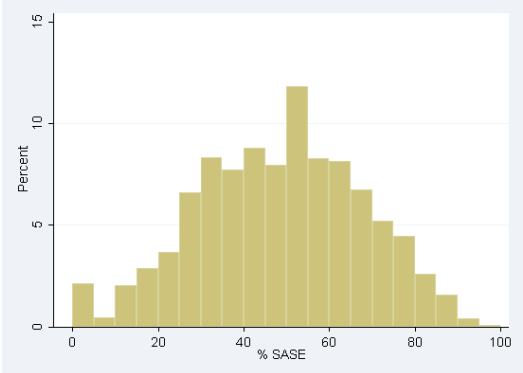


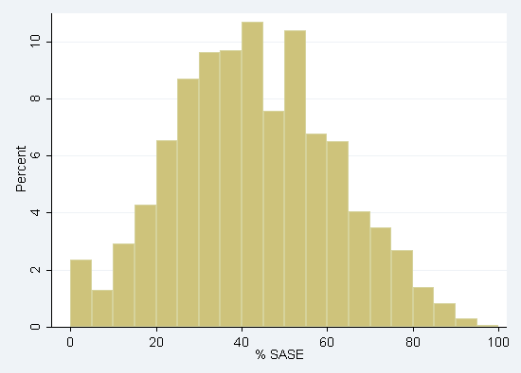


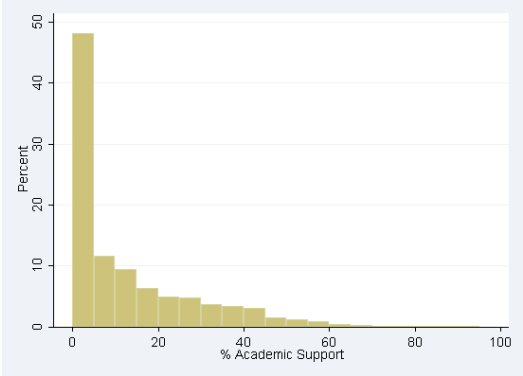


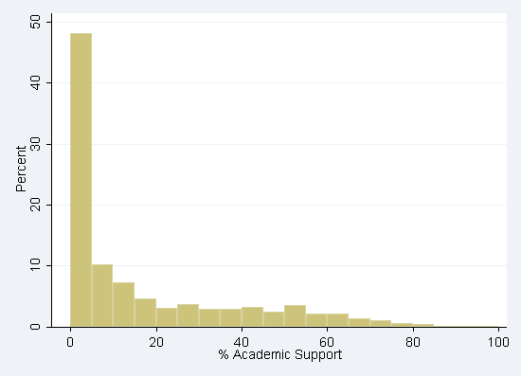

Percent
0
2
4
6
8
0
20
40
60
80
100
% Internet

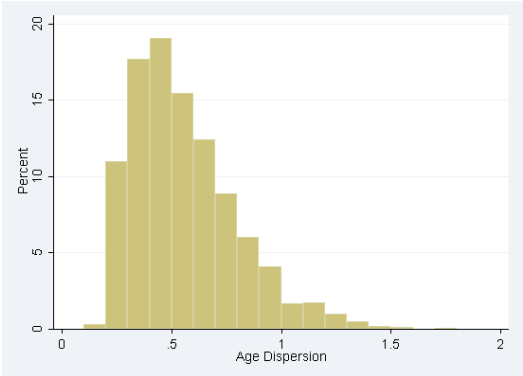

Percent
0
5
10
15
0
20
40
60
80
100
% Internet

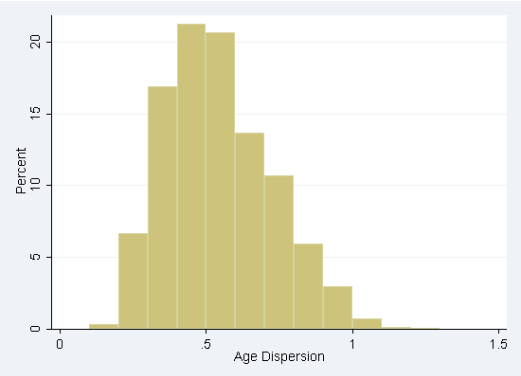

Percent
0
5
10
15
0
20
40
60
80
100
% SASE


Percent
0
2
4
6
8
10
0
20
40
60
80
100
% SASE


Percent
0
10
20
30
40
50
0
20
40
60
80
100
% Academic Support


Percent
0
10
20
30
40
50
0
20
40
60
80
100
% Academic Support


Percent
0
5
10
15
20
0
.5
1
1.5
2
Age Dispersion


Percent
0
5
10
15
20
0
.5
1
1.5
Age Dispersion