cemfi

working paper

1704

Grading on a Curve: When Having Good Peers is not Good

Caterina Calsamiglia
Annalisa Loviglio

January 2017

CENTRO DE ESTUDIOS
MONETARIOS Y FINANCIEROS
Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

CEMFI Working Paper No. 1704
January 2017

Grading on a Curve: When Having Good Peers is not Good

Abstract

Student access to education levels, tracks or majors is usually determined by their previous performance, measured either by internal exams, designed and graded by teachers in school, or external exams, designed and graded by central authorities. We say teachers grade on a curve whenever having better peers harms the evaluation obtained by a given student. We use rich administrative records from public schools in Catalonia to provide evidence that teachers indeed grade on a curve, leading to negative peer effects. We find suggestive evidence that school choice is impacted only the year when internal grades matter for future prospects.

JEL Codes: I21, I28, H75.

Keywords: Grading on a curve, negative peer effects, school choice.

Caterina Calsamiglia
CEMFI and Barcelona GSE
calsamiglia@cemfi.es

Annalisa Loviglio
UAB and Barcelona GSE
annalisa.loviglio@uab.cat

Acknowledgement

We are grateful to the staff at the Departament d'Ensenyament and IDESCAT, and in particular to Xavier Corominas and Miquel Delgado, for their help in processing the data. We are also grateful to Manuel Arellano, to Chao Fu, to participants of the "Children's health, well being, and human capital formation" workshop (Barcelona GSE Summer Forum 2016) and to participants of PhD workshops in CEMFI and UAB for their useful comments. All errors are ours. Caterina Calsamiglia acknowledges financial support by the Spanish Plan Nacional I+D+I (ECO2014-53051-P), the Generalitat de Catalunya (SGR2014-505) and the Severo Ochoa program (SEV-2015-0563) and from the ERC Starting Grant 638893. Annalisa Loviglio acknowledges financial support of La Caixa Foundation (La Caixa-Severo Ochoa International Doctoral Fellowship).

1 Introduction

Student's grades are used for two main purposes: to certify the mastery on a given subject and to compare students when selecting them into tracks, colleges or jobs. We distinguish between tests designed and graded by teachers teaching the subject in school and those tests designed and graded by external examiners, often picked by centralized authorities nationally or internationally. Tests are usually divided into different questions and each one of them is assigned a number of points. The final grade is then calculated as the percentage points earned of the total points in the exam. This is clearly the process followed when grading external evaluations, but is less clear-cut when teachers are grading.[1]

Internal evaluations capture human capital accumulation (cognitive skills), as external evaluations do, but may also capture teachers' bias. Lavy (2008) or Lavy and Sand (2015) provide empirical evidence that teachers exhibit a gender bias, often providing differential grades to females and males, and show that this bias may have long run effects. Diamond and Persson (2016) uses data from Sweden to show that teachers may inflate grades in high stakes exams for students who had a "bad test day", but do not discriminate on immigrant status or gender. They also show that teacher discretion has long term consequences for individuals in terms of level of education and earnings.

This paper provides empirical evidence of an additional source of disparity between internal and external grades and a channel through which having better peers can be harmful. In particular we show that a student in a classroom with better peers receives lower grades from the teacher than an identical student with worse peers. In principle in Catalonia – similarly to many other countries – grades in a class do not have to fit a given distribution, but shall measure absolute performance. In practice, the difficulty of lectures and exams may be at least partially adapted to the characteristics of students in the group, and teachers may be induced to grade differently depending on the quality of their students. In this paper we use a minimal definition of *grading on a curve* (GOC). We say that teachers *grade on a curve* whenever having better performing peers harms the grade provided to a given student, namely when relative performances affect the given evaluation.

Providing empirical evidence on this facts presents many challenges, both in terms of the data requirements and identification. Using a rich data set of the universe of children in primary and secondary school in public schools in Catalonia we show that grades assigned by teachers are negatively affected by average peer quality.[2] In other words, having good

[1] For instance, an article in The New York Times, "A's for good behavior" notes that teachers often reward students for their good behavior and not for their mastery in the subject. See the full article at http://www.nytimes.com/2010/11/28/weekinreview/28tyre.html

[2] Catalonia is one of the most prosperous autonomous communities in Spain with more than seven million

peers need *not be* beneficial if internal grades are important.

For identification we instrument for average peer quality through expected age at entry, and we control for school fixed effects to address selection of students into schools. Moreover we exploit the fact that students in primary school are homogeneously distributed into classrooms based on time-invariant observables. For secondary school we cannot rule out sorting into classes, but we discuss size and direction of possible biases and run a set of robustness checks that confirm the persistence of our results throughout primary and secondary school.

One of the most widely studied topics in Economics of Education is that of peer effects and how class composition may affect human capital accumulation. The literature is large and the evidence varies – see Sacerdote (2011) and Epple and Romano (2011). But in all studies, having relatively better peers is not harmful on average for human capital accumulation, and is beneficial for most individuals.[3] This paper highlights a different channel through which peer composition can affect long run educational outcomes. Although the accumulation of human capital is not harmed by the presence of better peers, the *perception* that teachers have of a student can be affected by the quality of its peers. In particular, if teachers somehow grade on a curve, then having better peers can induce teachers to give lower grades to a given student when faced with better peers.[4]

What are internal grades important for? On the one hand, attitudes of teachers towards individuals in class are said to affect students' self-image and self-confidence, and substantially influence their future educational outcomes. Such mechanisms have been widely documented in the psychology and sociology literature. Similarly Kinsler, Pavan, and DiSalvo (2014) shows that relative performance of a child in school affects parents' inference of the child's ability and parental investment in the child. Bobba and Frisancho (2014) show that students' perception of their own ability is affected by performance in exams.[5] Azmat and Iriberri (2010) and Tran and Zeckhauser (2012), on the other hand, show that students care and react to their relative position in the classroom. Hence, grades in the classroom may affect students' perceived ability, future expectations and performance.[6] But internal grades

citizens. The Catalan government has the powers to legislate in matters such as health or education, among others.

[3]Burke and Sass (2013), Carrell, Sacerdote, and West (2013), and Feld and Ulf (2016) find that a higher share of top performing peers in the group may harm performances of the low ability students. Our setting is quite different because internal evaluations of every type of students are negatively affected by the presence of better peers.

[4]Tincani (2015) explains how students preference for ranking in the classroom can explain how peers affect each other in a classroom and why that effect is heterogeneous depending on the relative position of an individual in the rank in class.

[5]Ahn, Arcidiacono, Hopson, and Thomas (2016) show that grading policies in college may affect major choice.

[6]Mayer and Jencks (1989) reviews the sociology literature and states that living in an advantageous

can also matter directly to the extent that they determine later access to school track or university. For instance in Germany or Romania, school track in secondary school depends on internal grades. Similarly, access to an excellence program for high school in Madrid, Spain, depends on the internal grades obtained in middle school. On the other hand, university admissions in Spain, Norway or Chile are determined through a centralized procedure for which a mix of internal and external grades determine priority in choosing major and university. In other countries, such as Germany, Sweden or Italy admission to some selective universities or highly demanded majors depends on a score that incorporates among other components internal grades in high school.[7] Finally applications to selective institutions in USA or Canada typically include GPA in high school. Admission committees might be able to weight this information according to the reputation of the sending institution, but most likely they cannot unravel the effect of occasional variations in peers or teachers quality.

To illustrate the implications that these differences between internals and externals may have, we simulate a selection process that selects on the basis of internal grades and compare it to one that selects on the basis of external grades at the end of primary school using our data in Catalonia. We find that the 25% top performing students are very different if selected though grades in internal or external evaluations. In particular, 30% of those selected through internal grades do not get selected through external grades, and vice-versa. Of these initial differences, about one third (10 p.p.) are explained by differences in the unexplained components of internal and external evaluations. Most of the remaining gap (from 45 to 70%) is due to grading on the curve and school grading policy. Thus differences in grading standards across schools and classes explain a large part of the differences in ranking using internal and external evaluations. Conversely teachers' biases, such as the gender bias, appear to be less relevant in this case.

In Catalonia internal grades impact academic prospects at the end of high school when applying to university, where priority in the desired major in a particular university is given as a function of a compounded grade composed 60% by average GPA (internal grades) in high school (last two years before university) and 40% by a nation wide exam.[8] Hence, students at the end of middle school, before starting high school, may be interested in moving to a school with relatively worse peers to increase internal grades towards university admissions. Changing school within the public system is difficult in Catalonia – see Calsamiglia and Güell

neighborhood may be disadvantageous, because a given student will rank worse if in an advantageous neighborhood, which may affect his or her expectations.

[7]The organization of education systems in Europe is described in https://webgate.ec.europa.eu/fpfis/mwikis/eurydice. Information about the Chilean system can be found at www.mineduc.cl.

[8]Students can undertake additional field-specific tests to improve their score. This may reduce the weight of average GPA in high school to 50%.

(2014) for a description of school choice in Catalonia. Moving is slightly more frequent among students that complete a private or semi-private middle school. Among this subsample of movers, 75% move to a school with relatively worse peers than in the previous school.

Estevan et al. (2014) analyze how the Top Ten Percent Law in Texas can generate desegregation in school because relative performance with respect to your peers is what determines access to university. Here we find that a similar effect may impact school choice at the end of middle school: better peers lead to worse internal grades, which in turn affect college admissions. This leads to some students switching schools in search of worse peers.

In the following section we present a simple model describing how external and internal grades are generated. Section 3 describes the data. Section 4 contains the empirical strategy and the results. Section 5 runs simulations on how the top selected students would change if the different sources of disparity between internal and external evaluations were controlled for. Section 6 discusses strategic change of school at the end of low secondary education. Section 7 concludes.

2 A simple model for internal and external evaluations

In this simple illustration we assume that individual human capital at a given point in time is a random variable H with expected value $E(H) = 0$. Let $\overline{H}$ be the average human capital in a class, with $E(\overline{H}) = 0$.

External evaluations measure human capital with some noise:

$$\text{ext} = \mu H + \varepsilon_E \tag{1}$$

where $cor(\varepsilon_E, H) = 0$ and the size of the coefficient μ depends on the extent to which ext captures human capital.

We assume that internal evaluations capture the same cognitive skills (as measured by the parameter μ), but they may also be affected by bias or grading standards. For simplicity we include just one bias based on gender F ($F = 1$ if student is a female, $F = 0$ if student is a male).[9] Moreover we allow teachers to consider both absolute and relative performance

[9] In the empirical analysis we test the presence of biases for several observed characteristics; however including more variables here would just complicate the exposition without providing any further insights.

when they assign evaluations.

$$\text{int} = (1-x)\mu H + x\mu(H - \overline{H}) + \delta F + \varepsilon_I \tag{2}$$

$$= \mu H - x\mu\overline{H} + \delta F + \varepsilon_I \tag{3}$$

where the error term ε_I is uncorrelated with both H and F. $x \in [0,1]$ and $1-x$ are weights given to relative and absolute performances respectively. Let's ignore for now the contribution of the bias δF; if $x = 0$, i.e. if only *absolute* performance matters, internal evaluations depend only on individual skills H, and would be completely analogous to external evaluations, except for the error component. On the other hand, if $x = 1$, the internal evaluation is based only on relative performance, as measured by the distance from the mean in the class. $x \in (0,1)$ means that both absolute and relative performance contribute to the final grade. The underlying interpretation would be that teachers adjust evaluations taking into account the average level of the class, either *ex-ante*, adapting the difficulties of lectures and tests, or *ex-post*, comparing students among them when they are assigning final grades. The magnitude of x in equation (3) tells us the relevance of grading on a curve in the school system under analysis.

Finally the parameter δ captures the additional reward (or punishment) for student gender F. It is important to stress that we do not take a stand on whether $\frac{\partial H}{\partial F} = 0$ or $\frac{\partial H}{\partial F} \gtrless 0$. If for instance $E(H|F=1) > E(H|F=0)$, this would affect in exactly the same way external and internal evaluations. δF only captures any additional difference due to gender that affects only internal evaluations. For example, if females put in more effort in school and therefore learn more contents, this would boost their human capital, increasing similarly both their internal and their external grades. However, if females, as opposed to males, are quiet in class, and teachers award some extra points for good behavior at the end of the year even if their human capital is not larger, δ would capture this. Hence, δ captures any difference between internal and external for females.

We can derive human capital from equation 1 ($H = \frac{1}{\mu}(\text{ext} - \varepsilon_E)$ and $\overline{H} = \frac{1}{\mu}\left(\overline{\text{ext}} - \overline{\varepsilon}_E\right)$) and replace it in (3)):

$$\text{int} = \text{ext} + \delta F - x\overline{\text{ext}} + \varepsilon_I - \varepsilon_E + \overline{\varepsilon}_E \tag{4}$$

$$= \text{ext} + \delta F - x\overline{\text{ext}} + \varepsilon \tag{5}$$

Section 4 discusses in detail how we bring an extended version of equation (5) to the data. The first obvious issue is that $cor(\text{ext}, \varepsilon) \neq 0$ at least because $cor(\text{ext}, -\varepsilon_E) < 0$. While this measurement error would downward bias the estimate, we cannot rule out any other

(positive) correlation between ext and ε, for instance if any unobserved characteristic affects in the same direction external and internal evaluations. Similar issues potentially apply to $\overline{\text{ext}}$; although errors may partially cancel out when taking average values, the typical class size is not large enough to ensure that $cor(\overline{\text{ext}}, \bar{\varepsilon})$ is 0. Thus we use an instrumental variable approach to obtain consistent estimates. In particular we use individual and average age at enrollment in primary school (A and $\overline{A}$ respectively) as instruments for external evaluations. In section 4 we extensively discuss the validity of such instruments. It is clear from equation (5) that A is a good instrument if it affects internal evaluations only through human capital but it is not a source of bias.

In principle we may rewrite (5) as

$$\text{int} - \text{ext} = \delta F - x\overline{\text{ext}} + \varepsilon \tag{6}$$

where the only endogenous variable is $\overline{\text{ext}}$. We prefer the formulation in (5) because it allows us to directly test the hypothesis that internal and external evaluations capture human capital in the same way, namely that they share the common parameter μ. If such assumption is true, the estimated coefficient of ext should be about 1; we verify this hypothesis in the data.[10]

3 Catalan school system and Data sources

Primary school (*Educació primaria*, EPRI) is the first stage of compulsory education in Catalonia; children begin primary school in September of the year in which they turn 6 years old. About 67% of students attend a public school; 30% of them attend a semi-private school, and the remaining a private school outside of the public school system.[11] Normally primary education takes 6 years, followed by 4 years of middle school (*Educació secondaria obligatòria*, ESO). After successfully completing lower secondary education, students can enroll in upper secondary education for two more years.

The core of our analysis (sections 4 and 5) focuses on students enrolled in either the last level of primary school or the last level of middle school. To be more specific, we study students enrolled in sixth grade of a public school in Catalonia from school year 2009/2010 to school year 2013/2014, and students enrolled in fourth grade of a public middle school in

[10] Moreover the model in levels can be easily modified to accommodate the case in which the coefficient for human capital in equations 1 and 3 are $\mu_{\text{ext}} \neq \mu_{\text{int}}$. In this case equation 5 would instead be $\text{int} = \frac{\mu_{\text{int}}}{\mu_{\text{ext}}}\text{ext} + \delta F - x\frac{\mu_{\text{int}}}{\mu_{\text{ext}}}\overline{\text{ext}} + \varepsilon$ and x can be backed up from the empirical estimation dividing the coefficient of $\overline{\text{ext}}$ by the coefficient of ext. For ease of exposition, we assume $\mu_{\text{ext}} = \mu_{\text{int}}$ from the beginning.

[11] Semi-private schools (*Concertadas*) are run privately and funded via both public and private sources.

Catalonia from school year 2011/2012 to school year 2013/2014.[12] In section 6 we exploit data of students enrolled in last grade of middle school and first grade of high school, in all types of schools.

We exploit data from different sources that provide us with detailed information on enrollment, school progression, academic outcomes and socio-demographic characteristics of Catalan students. The *Departament d'Ensenyament* (regional ministry of education in Catalonia) provided enrollment records for the schools in the region, from preschool to high school. The IT infrastructure that supports the automatic collection of data has been progressively introduced since the school year 2009/2010. By year 2010/2011 most of the schools have already adopted it, while we have data for about 60% of them in 2009/2010.[13]

Basic information (date of birth, school and class attended) are available for children in all types of schools, but more detailed socio-demographic characteristics (such as gender and nationality, special needs) are collected only for children in public schools. Moreover for children enrolled in public school we observe the internal evaluations that they receive at the end of the year for each subject they have undertaken. These final evaluations are assigned by teachers taking into account the progression of the child and her performance in several tests administered during the year.[14] For each class in a public middle school we also observe the identifier of teachers that taught Maths and Spanish in that class during the year; we do not have however any additional information on teacher characteristics.

The *Consell d'Avaluació de Catalunya* (public agency in charge of evaluating the educational system) provided us with the results of standardized tests taken by all the students in the region attending 6^{th} grade of primary school and 4^{th} grade of middle school.[15] Such tests are administered in the spring since 2008/2009 for primary school and since 2011/2012 for middle school. They assess basic competence in Maths, Catalan, Spanish and English and have a purely statistical purpose: they do not affect the students' final evaluations or progress to the next grades. We refer to the results in these tests, the grading of which is blind, as *external evaluations*, in contrast with the final evaluations given by teachers in the school, that we call *internal evaluations*. The four tests are administered in two consecutive days in the same premises in which students typically attend lectures. Normally every

[12] These levels correspond to ages 11-12 and 15-16 respectively.

[13] Some schools initially report data only for their lower grades, covering the entire pool of students only after two or three years. Therefore more data is available for more recent years.

[14] For primary school only evaluations at the end of second, fourth and sixth grade (i.e. at the end of "low", "medium", and "high" cycle of elementary education) are officially recorded in the centralized database and available to us. An evaluation of the child's progression is performed also at the end of first, third and fifth grade, in fact children can be retained one more year in the same level at any point of primary education.

[15] More information on these tests can be found in the following website (in Catalan): http://csda.gencat.cat/ca/arees_d_actuacio/avaluacions-consell/

student is required to take all the tests, although the school can decide to exempt students with special educational needs and children that have lived in Spain for less than two years. Moreover children that are sick one or both days and do not show up at school are not evaluated. We drop from the sample children labeled as children with special educational needs (less than 4%). We include in the analysis only classes in which results of the four tests are available for more than 80% of the children in primary school and for more than 70% of the children in middle school.[16]

Finally we collect information on the student's family background, more specifically on parental education from the Census (2002) and local register data (*Padró*).[17]

All data sources have been merged and anonymized by the Institut Català d'Estadistica (IDESCAT).

Table 1 shows some basic descriptive statistics by school year. Figure 5 plots histograms that describe the distribution of internal and external evaluations.

4 Empirical Analysis

4.1 Specification

Our analysis departs from the following empirical specification:

$$\text{int}_i = \gamma \text{ext}_i \overbrace{- x\overline{\text{ext}}_{c_i} + \sigma_{s_i}}^{\text{grading standards}} + \overbrace{\delta_F F_i + P_i \delta_P + \delta_M M_i}^{\text{bias}} \overbrace{+ \overline{X}_{c_i}\beta}^{\text{class characteristics}} + \overbrace{\tau_i}^{\text{year}} + \varepsilon_i \qquad (7)$$

where student i attends public school s_i in class c_i, and receives internal evaluations int_i and external evaluations ext_i. We study separately students in 6^{th} grade of primary school and 4^{th} grade of middle school. Both int_i and ext_i are computed as average of four subjects:

[16] We chose these two threshold in order to keep approximately 80% of the observations for both levels of school. We replicate the analysis in section 4 and 5 choosing different thresholds (in particular including all classes with more than 70% of test takers for primary school, that allow us to keep 90% of the observations) and results are basically the same.

[17] When the information can be retrieved from both sources, we impute the highest level of education, presumably the most up-to-date information. In the analysis we use dummies for "parental background" based on the average level of education of parents: "low" if both parents are early school leavers, "high" if at least one parent holds a tertiary education degree and the other parent graduated from high school, "medium" for any other case. For single-parent family we use the level of education of the single parent. We couldn't identify any of the parents for 4.5% of children in our sample; for them we use a dummy for "missing parents" in the analysis. Excluding them from the analysis does not modify the results. To compute average level of parental background in the class, we use for each student an index representing the average level of education of parents. The index takes 5 values, from 0 (both parents are early school leavers) to 4 (both parents hold a tertiary education degree).

Maths, Catalan, Spanish and English. We use z-scores for each subject and year. Using GPA rather than running separate analysis by subject is particularly convenient for two reasons. On the one hand teachers may not separately assign their evaluation, but often meet and discuss together the performance of each student. Therefore we cannot exclude that the final score in one of the subjects is somehow affected by the results in other subjects. Hence, the GPA may be the most suitable measure of skills. On the other hand the internal grade for each subject can take at most 11 different values, therefore using the GPA improves the variation of the dependent variable.[18] We also discuss results when analyses are performed by subjects. Main findings are unchanged.

According to the simple model discussed in section 2, we expect $\widehat{\gamma}$, the coefficient of individual external evaluation ext_i, to be about 1.

The coefficient of $\overline{\text{ext}}_{c_i}$, the average external evaluations in the class, will allow us to estimate the rate x of grading on the curve.[19] School fixed effects, σ_{s_i}, capture the differences in grading across schools that are constant over time and across classes. Schools may have different grading policies depending on the average pupils they face or the requirements or objectives that they may fix for the school, which may be orthogonal to pupils' observables. Unfortunately we cannot disentangle which part of the school fixed effect is determined by the quality of the students and which part depends upon other factors. Our identification exploits only within school variation: we are estimating the impact of classmates' quality on internal evaluation conditional on attending a given school. Both grading on a curve, as measured by class-level variation, and school fixed effects cause students with similar characteristics and ability to have different internal evaluations. In this paper we will refer to their joint effect as the effect of *grading standards*.

Equation (7) includes dummies for a bunch of individual characteristics: gender (F_i), foreign born status (M_i), a vector of dummies for parental education (P_i is low, medium

[18] In primary school the available evaluations are "Insufficient", "Sufficient", "Good", "Very good", "Excellent". In middle school each of these words correspond to an interval of numeric grades between 0 and 10: students receives both an integer grade from 0 to 10 and the wordy evaluation associated with it. Using the same conversion scheme, we assign to each evaluation the midpoint of its interval (and then we take z-scores); thus "Insufficient" is interpreted as 3, "Sufficient" as 5, "Good" as 6, "Very good" as 7.5 and "Excellent" as 9.5. An alternative approach for primary school would be to just use numbers from 1 to 5. If the analyses discussed in this section are replicated using this second approach results are extremely similar.

[19] Average external evaluations, as well as class characteristics, are constant in the class. In the peer effects literature the individual i is typically excluded from the computation of the average (see Sacerdote (2011)). However in this case using average at the class level is aligned with the model described in section 2, and it appears intuitively more sensible: teachers have a unique reference point (the "average performance") and compare each child with this reference point, rather than changing reference point for every student. We replicated all the analysis described in this section using average values among peers in the class, rather than at the class level, and all the results are extremely similar.

or high, or missing); their coefficients are different from zero if those characteristics directly affect internal evaluations on top of their contribution to human capital.[20] The equation controls also for their class averages (vector $\overline{X}_{c_i}$). Including regressors in $\overline{X}_{c_i}$ serves two purposes. First, controlling for any class-level bias due to class composition. For instance if on average classes with higher share of females are a more quiet, and teachers are more lenient with a class in which misbehavior is infrequent, then the coefficient of the "share of female" regressor would capture this. Second, for simplicity in section 2 we modeled relative performance as computed using the true underlying human capital. In practice if teachers' biases are not fully conscious they may interfere with their estimation of the average human capital in the class. For instance if teachers on average somehow overestimate females skills, they may set a higher reference level in classes with more females.

Equation (7) includes also year fixed effects (τ_i).

While individual characteristics are clearly exogenous regressors, given that their values are determined before the child begins compulsory education, their average in the class may be endogenous because students are not randomly matched to their peers. The same issue applies to $\overline{\text{ext}}_{c_i}$. In section 4.3 we extensively discuss possible issues related to sorting of students across classes.

As already discussed in the previous section 2, both ext_i and $\overline{\text{ext}}_{c_i}$ can be correlated with the error term ε_i. Our identification relies on the use of A, student's age at enrollment in primary school, as instrument for external evaluation. Analogously $\overline{A}$, the average age at enrollment in the class, is used as instrument for $\overline{\text{ext}}_{c_i}$. This approach is correct if A affects the human capital accumulation, but does not impact differently external and internal evaluations. In section 4.2 we provide extensive evidence in support of the validity of age at enrollment as instrument.

4.2 Age at enrollment as instrument

The fact that a unique school cut-off date determines when a child can enter school induces large heterogeneity in the age at which a child enters school and the heterogeneity of ages encountered in classrooms, with the older children being up to 20% older than their youngest peers. Older children are substantially more mature than their younger peers, which leads them to initially perform better. Work by Heckman and coauthors shows that early child development is complementary to later learning – see Cunha, Heckman, and Lochner (2006) for a review. Bedard and Dhuey (2006) use international data to show that this early relative

[20] In this paper we use "immigrant" and "foreign born" as synonyms. The dummy M_i takes value 1 if the child does not have Spanish nationality. Strictly speaking she may be born in Spain from immigrant parents.

maturity effects propagate through the human capital accumulation process and have long run effects for adults. Several papers look at the effects within a country: Fredriksson and Öckert (2014) for Sweden, Puhani and Weber (2007) for Germany, Schneeweis and Zweimüller (2014) for Austria, Black, Devereux, and Salvanes (2011) for Norway, Crawford, Dearden, and Meghir (2010) for England, McEwan and Shapiro (2008) for Chile, Ponzo and Scoppa (2014) for Italy, and Elder and Lubotsky (2009) for the US.

The case of Catalonia deems particularly interesting as children are generally not allowed to postpone or anticipate entrance to primary school: virtually every child begins primary school in September of the year in which he or she turns 6 years old. This enrollment rule is quite sharp and exceptions are extremely rare.[21] We can verify using enrollment data for first grade of primary school that more than 99.1% of children are compliers.

Calsamiglia and Loviglio (2016), exploiting the same data sources of this paper, provide robust evidence that age at enrollment is an important determinant of educational outcomes throughout compulsory education. Figure 2 and table A-1 replicate their main results about the effect of maturity at enrollment on evaluations over time. For each school level, we regress evaluations on age at enrollment and controls, including cohort and school fixed effects.

The age effect is highly persistent over time, although decreasing in magnitude: *ceteris paribus* being born at the beginning of January rather than at the end of December increases the GPA by 0.56 standard deviations at the beginning of primary school, and by 0.32 standard deviations at the end of it. The gap is still sizeable in middle school.

The fact that the effect of maturity on school outcomes decreases over time supports the hypothesis that the difference in maturity is a strong negative shock at the beginning of formal education, that persists over time because current human capital is built on past human capital. Younger children have a learning disadvantage at the beginning of primary school: all children in a class are exposed to the same educational methods and contents, but they may have different learning capabilities due to different levels of maturity. Thus younger students create a lower stock of human capital in the earlier stage of their school career. Later on the difference in maturity is likely to fade out: a child born in January and a similar child born in December have probably the same ability to learn new contents when they are 12, therefore if they had the same level of human capital from previous period, they would be able to increase it in the same way for next period. The issue is indeed that on average they *do not* have the same level of human capital from previous period: the initial disadvantage is so large that the negative effects propagate over time and the gap is not

[21] Enrollment in primary school was regulated by Decree 94/1992, issued on April, 28 (in Diari Official de la Generalitat de Catalunya (DOCG), núm. 1593 - 13/05/1992) until school year 2008/2009 and by Decree 181/2008, issued on September, 9 (in DOGC núm. 5216 - 16/9/2008) from the following year

closed at the end of lower secondary education.[22]

Those findings are reassuring that A surely affects human capital, but it is unlikely to have any differential effect on internal and external evaluations at the end of primary school. In fact the reduced form regressions in table A-1 show that the estimated effect is identical using internal or external evaluations. As a further check, we also pool together results in external and internal evaluations, and regress them on the same covariates and their interaction with a dummy that captures when the evaluation is internal. This is the standard approach in the literature to detect biases associated with given characteristics when the evaluation is not blind (see for instance Lavy (2008)). While the interaction with other characteristics is significant and large, the coefficient for the interaction term with A is completely insignificant and extremely small in magnitude.[23]

4.3 Potential issues due to sorting of students across classes

In Catalonia, as in most countries, students are not randomly allocated to schools: school composition typically reflects neighborhood characteristics. Our analysis includes school fixed effects, so that we only exploit variation within school across classes over the time period covered in our sample. Therefore variation of regressors measured at the class level comes both from the fact that typically a given school has more than one class per year and from the fact that the school appears in the sample for more than one year. During the short period under analysis (from 2009 to 2013 for primary school, from 2011 to 2013 for middle school) there wasn't any change in enrollment rule or in the demography of the region that may suggest dramatic change in schools' composition. In fact average characteristics at the school level such as parental education or share of immigrant students are highly correlated over time. Thus time invariant fixed effects should control for sorting across schools.

While school's enrollment in Catalonia is highly regulated and based on well know priority criteria, rules on how students shall be allocated across classes in a given school are not formally defined.[24] Apparently in primary school classes are particularly designed to be homogeneous in the observables. For instance a primary school with two classes for first graders in a given year allocates female students more or less evenly in the two classes. Moreover administrators and teachers use information provided by preschool educators and parents to allocate children so that each class receive a fair number of children that showed

[22]Note that the empirical results support the hypothesis because if children continued to increase human capital at a lower rate, the estimated effect of A would be increasing rather than decreasing over time.

[23]Results are available upon request.

[24]Most primary schools in our sample have either one or two classes (about half and half), only 6% have three or more classes. Secondary schools are typically larger: almost 40% have two classes, 30% have three classes, 16% four or more, and the remaining only one.

high or low ability in the previous years. To support the anecdotal evidence, we formally test that there is no sorting in primary school, finding evidence that student's characteristics and the class the student is assigned to are statistically independent. Appendix A describes our methodology and results. Therefore although children are not assigned to classes with a random draw, their allocation is balanced and the variation in peers composition across classes can be considered *as good as random* for statistical purposes. If anything we may be concerned that the variation in class characteristics across classes of the same school in a given year is limited. Luckily we are also exploiting variation over time, and – as detailed in appendix A – a variance decomposition confirms that although some characteristics vary more between schools than within, there is reasonable variation also across classes.

Conversely a number of middle schools may sort students across classes based on their previous grades or on their intention to pursue further academic studies in the future.[25] We have no information on how teachers of a given school are assigned to classes. Thus there are at least two dimensions that may interfere with our analysis: first, allocation to classes may not be random, i.e. characteristics of students in a class are sometimes correlated; second, assignment of teachers to classes may not be random, i.e. characteristics of teachers and students in the class may be correlated.[26] While we know that students with similar ability or similar background might be more likely to be together, we have no reason to believe that the assignment of teachers to classes follows a systematic pattern, although we cannot exclude that sorting of some kind takes place. In the following paragraph we will discuss how these potential issues may affect our estimates and present the robustness checks we will perform in section 4.5. Appendix B contains a more formal illustration of the challenges to identification.

Let us first abstract from the matching of teachers to classes. A recurrent concern in the peer effects literature is that the sorting of students across classes may interfere with the identification of peer effects on the outcome of interest.[27] Sorting of students across classes is problematic if peer group composition is correlated with omitted variables that affect the dependent variable: estimated coefficients of group characteristics would spuriously

[25] We performed the same battery of tests described in appendix A using data from middle school. Although for each year a large number of schools have pretty much homogeneous classes, overall the results do not allow us to exclude sorting.

[26] This would be the case for instance if more experienced teachers are given higher performing classes, or, vice-versa, if the best teachers are assigned to group of students that lack behind. Unfortunately we have no information on the characteristics of teachers that work with students in our sample.

[27] See also Ammermueller and Pischke (2009) for a discussion of potential issues related to non-random allocation of students to classes.

capture the effect of omitted variables on the dependent variable. This is a major issue in a quite common setting in the literature: a test score is regressed on individual and peers' predetermined characteristics, to estimate the "reduced form" effect that characteristics of the group of peers have on individual outcome. Then the estimated coefficient may capture both the true effect of peers on individual performance and the fact that being with peers of given characteristics affects the probability that the individual is a high performer. In particular, if sorting is based on performance, a more able student is more likely to be enrolled in a class with high performing peers. In turn performance is typically correlated with predetermined characteristics such as parental background, thus a high performer is more likely to be in class with students with high parental background: a positive coefficient for the average parental background of the peer group may just be due to the positive correlation of this regressor with unobserved components of individual human capital.

Our setting has the advantage that we directly control for a measure of human capital (ext_i, instrumented to correct measurement error): if the model in section 4.1 is correctly specified, coefficients of other regressors only measure differences between internal and external evaluations. The fact that regressors at the class level are correlated with individual human capital would not be problematic, precisely because we control for it. An important assumption is that external and internal evaluations are meant to measure the same skills, but internal grades incorporate comparison with peers and "biases" that are orthogonal to cognitive skills. However in practice we cannot exclude that there are unobserved variables related to human capital that affect differently internal and external evaluations, and are not orthogonal to the sorting across classes. In particular teachers may observe and reward non-cognitive skills such as grit or perseverance; for instance given two children of similar cognitive ability, a teacher may decide to award a higher grade to the one that always shows interest in class and puts in more effort when doing homework.[28] The same variables might also be taken into account when students are sorted across classes, to asses whether they can benefit from a more challenging program or their willingness to attend an academic education afterwards. Thus children with high unobserved non cognitive skills would be more likely to attend a class with high performing peers (as measured by external evaluations), and they would be more likely to receive high internal evaluations. In this case the coefficient of $\overline{\text{ext}}_i$ would be upward biased. The more aligned internal and external evaluations are, the smaller the bias.

However, we can claim that our estimates provide a *lower bound* for the true relevance

[28]These variables may be correlated with the controls that we are including in the regressions, thus some of the "biases" may take care of part of their effect. However we cannot claim that the limited number of predetermined characteristics we are using fully account for non cognitive skills.

of "grading on a curve" in the system, the true effect being potentially larger than the one we find. In fact we expect the coefficient of $\overline{\text{ext}}_i$ (i.e. $-x$) to be negative, so that the sign of x, the rate of "grading on a curve", is positive. In practice $-\widehat{x}$ would also capture a spurious positive effect on internal evaluations of being with high performing peers. Thus the estimated coefficient may be smaller in magnitude than the true value.

We now discuss how non-random assignment of teachers to classes may cause further biases in our estimates. The issue here is that teachers grade their own students, and they may have different attitudes: some may be generally more lenient, other stricter, above and beyond the fact that they may compare students among them to assign grades. This is problematic for identification if the "type" of teacher is correlated with the "type" of class: in this case the estimation of $-\widehat{x}$ would be affected by any differential leniency of teachers assigned to "good" or "bad" classes. For instance, if more lenient teachers are more likely to teach in classes of high performers, then the coefficient of $\overline{\text{ext}}_i$ would be upward biased. The most problematic case for our exercise is the negative bias that would arise if strict teachers were systematically assigned to classes of high achievers. In this case students in "good" classes would be given internal grades that are low relatively to their external grades not because they are compared with their peers, but because they have a different type of teacher than students in "bad" classes. As a consequence the true "grading on a curve" would be smaller than the estimated one. Although there is no reason to believe that this very specific assignment of teachers to classes takes place, *ex ante* we cannot exclude it or any other correlation between teacher and students' characteristics.

In this paper we perform and discuss in parallel analyses for primary school and middle school. Concerns related to sorting apply only to middle school, because we have evidence that in primary school allocation to classes is "as good as random" for our purpose. The fact that results are fully aligned provides evidence that having different grading standards across classes and schools is a recurrent feature of the Catalan educational system.

Moreover we use a threefold strategy to ensure identification when working with middle school data. First, we perform the same analysis on the sample of classes whose rank correlation between internal and external grades is very high.[29] Comparison of students among them may "shift" up or down the internal evaluations depending on class composition, but does not change the relative position of students in the class. Conversely if internal evalua-

[29]For each class we compute the Spearman's rank correlation coefficient between internal and external evaluations in the class; this is equal to the Pearson correlation between the rank values of those two variables.

tions take into account (non cognitive) skills that are not measured by external evaluations, the order may change dramatically. Hence, for the subsample of classes with high rank correlation, the two evaluations are truly aligned measure of human capital.

Second, we replicate analysis for middle school using teacher fixed effects rather than school fixed effects. If results are driven by a systematic association of strict teachers and high achieving classes, then $\widehat{x}$ should be much smaller and perhaps insignificant in the regression with teacher fixed effects.

Third, we replicate the analysis using school-level rather than class-level regressors. In other words, we compute average including all the schoolmates enrolled in the same level, rather than just classmates. Average performance at the school level is obviously correlated with average performance at the class level, but is probably a less precise measure of the references group that teachers have in mind when grading children; moreover variation is limited to in school cohort variation given that we are controlling for school fixed effects. The advantage is that we can completely abstract from issues due to sorting of both students and teachers.

4.4 Results

Columns (2sls) of table 2 present the results of our estimation of equation (7) using the two stage least square approach with school fixed effects described in previous sections. For comparison columns (ols) contain coefficients of OLS estimations with school fixed effects. First stage estimation for columns (2sls) are shown in table A-2 in appendix D.

The coefficients for ext_i are close to 1, confirming that internal and external evaluations are capturing human capital in a similar way.[30] Having in mind the simple model described in section 2 we can deduce from the coefficients shown in columns (2sls) that the estimated rate of grading on a curve $\widehat{x}$ in the Catalan elementary school system is about 36%; the share raises to more than 50% in middle school.

The coefficient for ext_i when performing OLS is downward biased and some of the estimated biases are somehow larger in magnitude; coefficients for $\overline{\text{ext}}_i$ are significant and large also for the OLS model.[31]

The specifications highlight that females are favored in internal evaluations both in primary and in middle school: being a female increases internal evaluations by 0.15 and 0.36

[30] P-values of null hypothesis that the coefficient is 1 are 0.09 and 0.21 for primary and middle school respectively.

[31] Interestingly the estimate for middle school is smaller when OLS are used: if individual human capital is not appropriately controlled for, the average performances in the class may partially capture the missing information on human capital: given that who has better peers is more likely to be a good student, coefficient is downward biased.

standard deviations respectively. Children of more educated parents receive a relatively higher score in primary school (kids of parents with University degree have internal evaluations that are on average 0.13 higher than their external), while there are no relevant differences in middle school. There is no effect associated with being foreign born in primary school, while there is a positive premium in middle school.

Class characteristics have little effect: only the coefficient of the share of female in the class is significant, but the negative effect is small in size.

To correctly interpret the results for the "biases" associated with being female, or foreign born, or parental education, it is important to recall that in this paper we call "bias" any differential effect that individual characteristics have on internal evaluations on top of their contribution to human capital. Then, for instance, the result that having highly educated parents ensures on average a positive premium on internal evaluations in primary school should not be interpreted as evidence that teachers are actively discriminating children on the basis of their parental background. What we can conclude is just that those children have their internal evaluations increased for reasons not directly related to their cognitive skills. For instance in primary school parents are actively involved in the educational process, highly educated parents might be more keen to "lobby" for their children. Moreover highly educated parents might on average emphasize more the importance of behaving in class, or make sure that children always complete their homework: the good attitude of their children may then be rewarded by teachers over and above their actual skills level.[32]

Figure 3 plots the estimated school fixed effects (y-axis) versus the average external evaluations at the school level (x-axis). The figure emphasizes that school fixed effects can be sizeable, and on average schools where external evaluations are higher appear to set stricter grading policies.

Results by subject are shown in table A-3 of appendix D.[33] Overall the qualitative results discussed in previous paragraphs are unchanged. Both for primary and middle school the estimated rate of grading on a curve is higher for Catalan and Spanish (about 45% and about 60% respectively). Results for English and Maths in middle school are close to GPA, although slightly smaller. The coefficient for Maths in primary school is still sizeable but smaller (26%), moreover it is the only coefficient that is not significant at traditional level (p-value is 0.3). These results suggest that the languages leave more room for subjective

[32]Recent immigrant may face special difficulties in adapting to the Catalan educational system, especially if they were educated abroad for a long time before. Teachers may compensate them when grading, this would explain the positive premium for being immigrant in middle school.

[33]The limitations of studying subjects separately have already been discussed in section 4.1.

evaluations of teachers, and therefore to comparison among students.

"Biases" show a similar pattern across subjects, in particular there is a positive premium associated with highly educated parents in primary school, and a positive effect of being female both in primary and middle school (particularly large for Maths). Given that the analysis by subject is extremely consistent with the main specification, we will focus on GPA from now on.

4.5 Robustness checks

Table A-4 in appendix D contains alternative specifications. In columns (1) only own external evaluation is instrumented, but not the average in the class. In columns (2) and (3) the dependent variable is the difference between internal and external evaluations; in columns (3) the average external evaluations in the class is instrumented with age, as in the baseline specification.[34] All the empirical models deliver the same qualitative message about the importance of grading on a curve: the coefficient of $\overline{\text{ext}}_i$ is significant and sizeable in all specifications, although it is slightly larger when it is not instrumented. Thus the baseline model is the model that provides the most conservative estimate.

Tables A-5 and A-6 contain results for the three robustness checks discussed at the end of section 4.3.

For the first robustness check we use only the subsample of classes for which rank correlation of internal and external evaluations is high. If the only differences between internal and external grades were grading standards, then the *rank* of students within a class would be the same using the two evaluations. In fact incorporating other students performance in the final internal evaluations does not modify the relative position in the class. In practice both biases and random errors may causes differences in the ranking within a class.

Within class rank correlation among internal and external grades is very high in primary school: the mean value is 0.82 and 75% of classes has rank correlation higher than 0.80. In middle school rank correlation is not as large as in primary school, but it is still sizable: mean value is 0.64, 75% of classes has a value higher than 0.56 and 25% are above 0.72.

Replicating our analysis on the subsample of classes with high rank correlation provides us with a further confirmation that the estimated coefficient of $\overline{\text{ext}}_i$ is not driven by spurious correlation with unobserved variables. Columns (1) of table A-5 contain results of the specification restricted to the subsample of classes whose rank correlation is larger than

[34] Specifications (2) and (3) would be equivalent to run the baseline model and (1) respectively, imposing that the coefficient of ext_i is 1.

0.75. Results are fully aligned with the baseline specification in columns (2sls) of table 2, the estimated rate of grading on a curve being even larger for primary school.[35]

The second robustness check is specific to middle school, for which we can identify Maths and Spanish teachers. This allows us to empirically address the concern discussed in section 4.3: if stricter teachers are assigned classes of high performing students, then teacher fixed effects, rather than comparison with peers, would be the reason why internal evaluations are lower than external in "good" classes and vice-versa. We can test this alternative explanation adding teacher fixed effects to our empirical specification. If the coefficient of $\overline{\text{ext}}_i$ spuriously captures the positive correlation between strict teachers and well-performing class, then controlling for teacher fixed effects should take it to zero. Obviously we can't use GPA for this analysis, therefore we work with internal and external grades in Spanish and Mathematics. Results are shown in table A-6.[36]

A limitation of our data is that we observe all the teachers that taught a given class at some point during the school year, but we cannot disentangle main teacher and substitutes. Thus some of the teachers may have spent only few days with the class, for instance while the main teacher was sick, and have no role in the evaluation of the students. In columns Spanish (1) and Maths (1) we include dummies for all the teachers in the dataset, and we allow for multiple teachers associated with students in the same class. In columns Spanish (2) and Maths (2) the sample includes only classes for which we retrieved a single Spanish or Maths teacher (about 75% of the sample used in columns (1)). Estimated coefficients are similar among them and fully aligned with the initial model. In particular the coefficient of $\overline{\text{ext}}_i$ barely changes and it is always significant at 5% level.

Finally we broaden the definition of peers, including all the schoolmates enrolled in the same level, rather than just classmates. On one hand average performances at the school level are obviously correlated with average performances at the class level, on the other hand they are probably a less precise measure of the references group that teachers may have in mind when grading children. Moreover this measure varies only over time (five years for primary school, three years for middle school). As shown in columns (2) of table A-5 results are quite close to the baseline model. Not surprisingly the estimate for the coefficient of $\overline{\text{ext}}_i$ is slightly less precise, but it is still significant at 5% for primary school and at 10% for middle school.[37]

[35] Obviously the threshold of 0.75 reduces the sample much more for middle school than for primary school. Setting a stricter threshold for primary school would deliver very similar results.

[36] Given that only a small minority of teachers change school over time, we cannot include school fixed effects in the regressions. Thus teacher dummies are capturing both the individual teacher effect and the school effect.

[37] We use only schools in which more than 80% or more than 70% (for primary or middle school respectively) of students undertook external evaluations, to make sure that the average external evaluations is a meaningful measure of students quality. Therefore sample size in first columns of table A-5 is slightly smaller than in

Although a precise estimation of heterogeneous effects of grading on the curve on different subgroups of the population is beyond the scope of this paper, in appendix C we perform some basic checks. The baseline model is estimated on specific subsamples of the population: males and females, and students with low, medium, or high predicted external evaluations (on the basis of their predetermined characteristics). Estimates are really consistent across groups, confirming that the comparison with peers affect all the students, and results are not driven by particular subsamples.

5 The impact of GOC on selection processes: a simulation

To gain a deeper understanding of the implications of the differences between internal and external evaluations we simulate a selection process that selects the top quartile of students according to either their internal or external evaluations. On one hand academic performances in primary and middle school do not directly matter for tertiary education, thus this simple exercise is just illustrative of what the impact of selecting people using either school grades or standardized tests can be. On the other hand this setting is particularly suited to study differences between internal and external evaluations because it is unlikely that parents have strategically selected school to affect internal grades of the children.

For each school year we rank students from the best to the worst according to internal and external GPA; ties are broken at random. The best 25% are "selected" while the remaining students are "excluded".[38]

In primary school 31% of students selected through internal evaluations do not get selected through external evaluations and vice-versa; this figure is almost the same (32%) for middle school. This sizeable gap suggests that the procedure used to select people can make a difference for a relevant part of the population.[39] However this finding alone does not clarify what is the main reason behind the difference in outcomes of the two procedures.

The empirical model estimated in section 4 allows us to interpret the difference between internal and external evaluations as the sum of three main components: grading standards,

table 2.

[38]For a limited number of students (less than 1% every year) the random draw can make a difference between being selected in the top quartile using internal evaluations or being excluded. Results are not sensitive to varying the share of selected students or picking thresholds that ensure that the last selected student is strictly better than the first excluded child. Thus we ignore this issue from now on.

[39]All the results we discuss in this section are weighted averages of the outcomes for each year. Yearly results are remarkably similar.

biases, and residual errors. More specifically, given the estimates reported in table 2, column (4), the individual internal evaluations can be rewritten as

$$\text{int}_i \;=\; \widehat{\text{int}}_i + \widehat{\varepsilon}_i \;=\; \widehat{\gamma}\text{ext}_i + \widehat{\delta}_F F_i + P_i\widehat{\delta}_P + \widehat{\delta}_M M_i + X_{c_i}\widehat{\beta} - \widehat{x}\overline{\text{ext}}_{c_i} + \widehat{\sigma}_{s_i} + \widehat{\tau}_i + \widehat{\varepsilon}_i \qquad (8)$$

The residual $\widehat{\varepsilon}_i$ includes all the differences between internal and external evaluations that are not taken into account in our model. In particular it contains the difference between the random component of internal and external evaluations $\varepsilon_I - \varepsilon_E$, as is clear from equation (4). Even if biases or differences due to grading standards were not relevant, ranking of students using internal and external evaluations would be different due to different random errors of the two exams. Using $\widehat{\text{int}}_i$ to rank students allows us to get rid of differences between internal and external due to that randomness.[40] The selection of the top quartile of students performed using external evaluations ext_i and predicted internal $\widehat{\text{int}}_i$ differs for 19% of the selected students in primary school and 21.5% of the selected students in middle school. Thus removing the unexplained residual closes about one third of the gap, while the remaining two thirds depend on bias and grading standards, as detailed in equation (8).

We can then "switch off" the various components of the RHS in equation (8), redo the rankings, and compare outcomes, to gain further understanding of how each dimension contributes to the difference in outcomes of the two selection processes. Table 3 summarizes the results. Interestingly the simulation delivers the same message for both primary and middle school: a large part of the difference is due to grading standards (grading on a curve and school fixed effects), while "biases" are less important.

For primary school removing the effect of grading on a curve reduces the gap by almost 5 p.p. (from 19% to 14.2%); eliminating school fixed effects reduces the gap by almost 9 p.p. (from 19% to 10%). Eliminating both simultaneously, namely removing the effect of grading standards from internal evaluations, decreases the gap by 13.2 p.p., explaining 69.5% of the differences between the selection with ext_i and the selection with $\widehat{\text{int}}_i$.

The second part of table 3 shows that for primary school getting rid of biases alone would have only minor effects on the gap in rankings. In particular removing the positive bias for female in internal evaluations reduces the gap by only 0.65 p.p., while removing the effect of parental education slightly *increases* the gap by about 0.5 p.p.[41]

[40] Being more precise, we also get rid of all the unobserved differences not captured by our model. Thus we can regard the estimated differences in ranking between ext_i and $\widehat{\text{int}}_i$ as a lower bound of the true gap if we could remove only the differences in the random errors ε_I and ε_E.

[41] This last finding results from the fact that children with higher parental background are disproportionally attending schools with high performing peers, where grading standards as captured by school fixed effects are tougher. Therefore on one hand they are favored by the internal compared with external because of having

For middle school removing the effect of grading on a curve reduces the gap by more than 7 p.p. (from 21.5% to 14.4%), while switching off school fixed effects has a slightly smaller effect than in primary school (about 4 p.p.). When both are removed the gap decreases by more than 9.5 p.p.; thus grading standards explains 44.4% of the differences between the selection with ext_i and the selection with $\widehat{\text{int}}_i$.

Biases are slightly more relevant for middle school, overall they explain 15.2% of the gap, but grading standards are by far the most important component.

Thus we can conclude that in a selection process of top students at the end of either primary or middle school most of the difference between selection using internal or external evaluations would result from grading on a curve and school grading policies.

5.1 GOC and inequality

It is important to understand how different selection systems affect minorities and children with disadvantaged background. We do not observe family income in the data, therefore we rely on parents' education and foreign-born status.

Overall ranking based on internal evaluations select more children from disadvantaged background: students admitted only by the ranking based on the internal but rejected by the ranking based on the external are more likely to have less educated parents and more likely to be immigrant. However they are also attending schools in which peers have less educated parents and have low external evaluations. In other words children with less favorable parental background are more likely to attend "low performing" schools, that inflate internal evaluations more. In fact figures 4 and 5 suggest that the subgroup of children with less educated parents who attends "high performing" schools would prefer a selection based on external evaluations.

For this analysis we classified schools in 2013 in three groups ("low", "medium", and "high") on the basis of their mean performances in external evaluations the previous years. We can then compare the share of selected students among those with a given parental education and school "quality". Each graph in figures 4 and 5 focuses on a type of parental background and shows the share of students admitted for each school "quality" type if the ranking is performed using external evaluations (blue bar), and internal evaluations (red bar). Moreover the green bar displays results if internal evaluations are "corrected" removing residuals and grading standards, as explained in the previous section. The evidence is similar

highly educated parents, but on the other hand they are penalized because they are attending a school that awards less generous evaluations. Whether on average they would prefer to be ranked with internal or external evaluations depend on the relative size of the two opposite effects. However if only biases are removed, there is just the negative effect due to school fixed effects, and we find that the difference between internal and external ranking widen.

for primary and middle school. For both levels comparing the three graphs we immediately see that under any system and school type children with higher educated parents are much more likely to be selected than students with low educated parents. This is a consequence of the strong correlation between children performances and parental background in Catalonia. However the relative differences between internal and external evaluations are pretty similar across the three graphs: the share of students selected with external in 2013 is clearly increasing in school type, while it is much flatter for internal evaluations. Thus internal evaluations select relatively more children in low-type schools and relatively less children in high-type schools, while the difference is small in medium-type school.

Looking only at the aggregate statistics we may conclude that overall internal evaluations select more children with low parental background, but this evidence seems to be a consequence of the fact that those students are over represented in schools that we classified as low-type.

6 Implications for school choice

In Catalonia internal grades matter when applying to university. In fact priority in the desired major in a particular university is given as a function of a compounded grade composed by average GPA (internal grades) in high school (last two years before university) and by a nation wide exam. The weight given to internal grades is between 50% and 60%, depending on the specific tests chosen in the nationwide exam. According to our previous results, having "worse" peers may be beneficial for internal evaluations. It is then important to understand whether students and their family are aware of this fact and strategically select, when possible, worse quality peers to increase their GPA and boost their chance of admission to the preferred bachelor. This outcome may be suboptimal, especially if they select lower quality schools and end up accumulating less human capital than they would have otherwise.

In most cases the same school that provides lower secondary education also offers high school. If students want to switch to another school they have to apply to a centralized system, and the change is possible only if the chosen school has a free slot. Given that many schools are oversubscribed, especially in the public system, changing school may be difficult in practice. Thus students that we actually observe moving in our data may be a strict subset of the set of students that would be interested in changing, if they could.

For students that move we can compare results of external evaluations at the end of lower secondary education in the old school and in the new school. In particular we can compare the average results in the first and in the second school. Moreover we can assess whether

they would have improved their ranking within the school if they attended the new school the year before when the standardized test was administered.

In our data 21% of students in semi-private and private schools change school for the last two years of secondary education; about half of them chose another private or semi-private school, while the other half enroll in a public school. 75% of them move to institutions with lower average results, and 74% of them improve their position in the within-school ranking based on the external test: the average improvement in this subgroup is 18 p.p. Results are consistent when focusing only on students that move to public schools or only on students that move to private or semi-private schools. Moreover it is interesting to notice that although below average performers are more likely to move, individuals across the whole distribution are affected.[42]

Only 11% of students in public schools move to another institution for high school. 80% of them stay in the public system, while the remaining 20% select a private or semi-private school. While those in this second group on average slightly improve the quality of their new school, as measured by the average in external evaluations, those that stay in the public system on average move to schools where test scores are lower.[43]

For comparison we check what happens at the end of primary school for those children that have direct access to a middle school in the same institution. Almost all the institutions that offer both primary and lower secondary education are private or semi-private, therefore the most appropriate comparison is with the share of movers from non-public middle school. Note that in this case there is no selection in the near future based on internal grades. Movers at the end of primary school are only 6.5% and on average they chose schools with higher performing peers, contrary to what happens at the end of middle school.[44]

These results should be taken as suggestive evidence that families understand that there may be grading on a curve and reoptimize their choice in a way that does not necessarily lead to optimal human capital accumulation, but to improved bachelor assignment given that selection depends on internal grades. A rigorous analysis on this, however, would involve estimating preferences by parents under a school choice mechanism that does not provide

[42]The incentives to move depend on the relative grades obtained in one school versus the other and the required grade to access the bachelor of interest to the student. Hence, students moving may be doing so to improve their grade from 8 to 8.25 to enter Medicine or from 6.8 to 7 to enter Economics. This is why we should expect the incentives to matter throughout the ability distribution and not only at one particular threshold.

[43]However the size of the average difference in quality between old and new school is smaller than what found for students initially attending private schools.

[44]Very few students change school during primary or middle school. Their share is smaller than 4% in all levels.

incentives to tell the truth, which is beyond the scope of the present study.[45]

7 Conclusions

This paper puts forth a novel form of grading bias that suggests that having better peers may harm the grades assigned by teachers in school. In many countries grades provided by teachers in schools determine access to further studies. Hence, our findings suggest that having better peers may lead to worse educational careers and future prospects. Internal grades result from teachers following students and evaluating them in a more continuous basis, which seems to be a better evaluating procedure. But distortions due to individual or peer characteristics seem unjustified, especially when they may affect the future allocation of students into further career paths. It may be good to impose some corrective mechanism through which such systematic differences be somehow corrected for if distortions are too large. Specially if these grades have important implications for track or college assignment; otherwise school choice may be inefficiently determined.

References

Thomas Ahn, Peter Arcidiacono, Amy Hopson, and James Thomas. Equilibrium grade inflation with implications for female interest in stem majors. Mimeo, June 2016.

Andreas Ammermueller and Jörn Pischke. Peer effects in European primary schools: Evidence from the progress in International Reading Literacy Study. *Journal of Labor Economics*, 27(3): 315–348, 2009.

Ghazala Azmat and Nagore Iriberri. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435–452, August 2010.

Kelly Bedard and Elizabeth Dhuey. The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects. *The Quarterly Journal of Economics*, 121(4):1437–1472, 2006.

Sandra E. Black, Paul J. Devereux, and Kjell G. Salvanes. Too Young to Leave the Nest? The Effects of School Starting Age. *The Review of Economics and Statistics*, 93(2):455–467, May 2011.

Matteo Bobba and Veronica Frisancho. Learning about oneself: The effects of signaling academic ability on school choice. Mimeo, December 2014.

Mary A. Burke and Tim R. Sass. Classroom Peer Effects and Student Achievement. *Journal of Labor Economics*, 31(1):51 – 82, 2013.

Caterina Calsamiglia and Maia Güell. The Illusion of School Choice: Empirical Evidence from Barcelona. Working Papers 810, Barcelona Graduate School of Economics, July 2014.

[45] See Calsamiglia, Fu, and Güell (2014) for details on the mechanism used and the challenges that such estimation would entail.

Caterina Calsamiglia and Annalisa Loviglio. Maturity and school outcomes in an inflexible system: Evidence from Catalonia. Mimeo, September 2016.

Caterina Calsamiglia, Chao Fu, and Maia Güell. Structural Estimation of a Model of School Choices: the Boston Mechanism vs. Its Alternatives. Working Papers 811, Barcelona Graduate School of Economics, October 2014.

Scott E. Carrell, Bruce I. Sacerdote, and James E. West. From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. *Econometrica*, 81(3):855–882, 05 2013.

Claire Crawford, Lorraine Dearden, and Costas Meghir. When you are born matters: the impact of date of birth on educational outcomes in England. IFS Working Papers W10/06, Institute for Fiscal Studies, May 2010.

Flavio Cunha, James J. Heckman, and Lance Lochner. *Interpreting the Evidence on Life Cycle Skill Formation*, volume 1 of *Handbook of the Economics of Education*, chapter 12, pages 697–812. Elsevier, May 2006.

Rebecca Diamond and Petra Persson. The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests. NBER Working Papers 22207, National Bureau of Economic Research, Inc, April 2016.

Todd E. Elder and Darren H. Lubotsky. Kindergarten Entrance Age and Children's Achievement: Impacts of State Policies, Family Background, and Peers. *Journal of Human Resources*, 44(3), 2009.

Dennis Epple and Richard Romano. Peer effects in education: A survey of the theory and evidence. *Handbook of social economics*, 1(11):1053–1163, 2011.

Fernanda Estevan, Thomas Gall, Patrick Legros, and Andrew F. Newman. College Admission and High School Integration. Working Papers, Department of Economics 2014 - 26, University of Sao Paulo (FEA-USP), November 2014.

Jan Feld and Zölitz Ulf. Understanding peer effects - On the nature, estimation and channels of peer effects. Research Memorandum 002, Maastricht University, Graduate School of Business and Economics (GSBE), 2016.

Peter Fredriksson and Björn Öckert. Lifecycle Effects of Age at School Start. *Economic Journal*, 124(579):977–1004, 09 2014.

Claudia Goldin, Lawrence F. Katz, and Ilyana Kuziemko. The Homecoming of American College Women: The Reversal of the College Gender Gap. *Journal of Economic Perspectives*, 20(4): 133–156, Fall 2006.

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

Scott A. Imberman, Adriana D. Kugler, and Bruce I. Sacerdote. Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees. *American Economic Review*, 102(5): 2048–82, August 2012.

Josh Kinsler, Ronni Pavan, and Richard DiSalvo. Distorted beliefs and parental investment in children. Mimeo, December 2014.

Victor Lavy. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10-11):2083–2105, October 2008.

Victor Lavy and Edith Sand. On The Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers Stereotypical Biases. NBER Working Papers 20909, National Bureau of Economic Research, Inc, January 2015.

Susan E. Mayer and Christopher Jencks. Growing up in poor neighborhoods: How much does it matter? *Science*, 243(4897):1441–1445, 1989.

Patrick J. McEwan and Joseph S. Shapiro. The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates Using Exact Birth Dates. *Journal of Human Resources*, 43(1), 2008.

Michela Ponzo and Vincenzo Scoppa. The long-lasting effects of school entry age: Evidence from Italian students. *Journal of Policy Modeling*, 36(3):578–599, 2014.

Patrick Puhani and Andrea Weber. Does the early bird catch the worm? *Empirical Economics*, 32(2):359–386, May 2007.

Bruce Sacerdote. *Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?*, volume 3 of *Handbook of the Economics of Education*, chapter 4, pages 249–277. Elsevier, 2011.

Nicole Schneeweis and Martina Zweimüller. Early tracking and the misfortune of being young. *The Scandinavian Journal of Economics*, 116(2):394–428, 2014.

Michela Tincani. Heterogeneous peer effects and rank concerns: Theory and evidence. Mimeo, June 2015.

Anh Tran and Richard Zeckhauser. Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650, 2012.

8 Tables

Table 1: Descriptive statistics

School year	Parents' education				Female	Immigrant	N students	N schools
	low	*middle*	*high*	*missing*				
primary school – 6^{th} grade								
2009/2010	34.1%	35.8%	25.3%	4.8%	49.5%	12.6%	10,982	372
2010/2011	32.9%	36.4%	26.5%	4.2%	49.7%	12.9%	22,273	764
2011/2012	32.2%	36.6%	27.4%	3.8%	50.0%	11.8%	28,770	927
2012/2013	32.9%	36.6%	26.6%	3.9%	49.5%	12.7%	31,975	1027
2013/2014	31.4%	36.7%	28.0%	4.0%	49.1%	11.6%	33,082	1042
middle school – 4^{th} grade								
2011/2012	35.4%	37.1%	23.6%	3.9%	50.9%	12.6%	22,533	447
2012/2013	35.3%	36.4%	24.1%	4.2%	50.8%	13.1%	25,649	470
2013/2014	35.8%	36.2%	23.9%	4.1%	50.9%	13.9%	25,717	481

Table 2: Dependent variable: internal evaluations

	primary school		middle school	
	(ols)	(2sls)	(ols)	(2sls)
external ev.	0.834	1.030	0.763	1.122
	(0.003)**	(0.018)**	(0.007)**	(0.096)**
avg external ev.	-0.445	-0.360	-0.331	-0.569
	(0.016)**	(0.117)**	(0.016)**	(0.106)**
female	0.179	0.152	0.365	0.358
	(0.003)**	(0.004)**	(0.007)**	(0.006)**
immigrant	-0.058	0.005	0.102	0.249
	(0.006)**	(0.008)	(0.011)**	(0.040)**
parents M	0.115	0.046	0.025	-0.048
	(0.004)**	(0.007)**	(0.007)**	(0.020)*
parents H	0.262	0.131	0.170	0.005
	(0.005)**	(0.013)**	(0.009)**	(0.045)
missing parents	0.091	0.047	0.044	-0.000
	(0.009)**	(0.010)**	(0.016)**	(0.019)
share female	-0.036	-0.052	-0.113	-0.148
	(0.032)	(0.027)+	(0.040)**	(0.031)**
share immigrant	-0.111	-0.063	0.005	0.117
	(0.040)**	(0.049)	(0.062)	(0.104)
avg parents edu.	0.032	0.012	0.038	-0.004
	(0.010)**	(0.023)	(0.017)*	(0.035)
Constant	-0.210	-0.108	-0.247	-0.134
	(0.025)**	(0.040)**	(0.036)**	(0.046)**
N	127,082	127,082	73,899	73,899

Note. School and year fixed effects included. Dependent variable is internal GPA at the end of primary school and middle school respectively (average of evaluations in Mathematics, Spanish, Catalan, English). Sample for primary school spans from 2009 to 2013; sample for middle school spans from 2011 to 2013. Regressors are external GPA at the end of primary school and middle school respectively ("external ev."), average external GPA in the class ("avg external ev."), individual characteristics (gender, immigrant, control for parental education – parents L, i.e. low educated, is the baseline category), average characteristics of the class (share female, share immigrant, average parental education).
In columns (2sls) own external evaluations and mean external evaluations in the class are instrumented with age at enrollment in first grade of primary school and mean age at enrollment respectively. First stage estimates are shown in table A-2.

Table 3: Selection of top quartile of students

Primary school

Selection based on:		**Diff. with external**	**Improvement**
predicted	$(\widehat{\text{int}})$	18.99%	
w/o GOC	$(\widehat{\text{int}} + \widehat{x}\overline{\text{ext}})$	14.19%	25.29%
w/o school FE	$(\widehat{\text{int}} - \widehat{\sigma})$	10.04%	47.15%
w/o school FE & GOC	$(\widehat{\text{int}} + \widehat{x}\overline{\text{ext}} - \widehat{\sigma})$	5.80%	69.49%
w/o female bias	$(\widehat{\text{int}} - \widehat{\delta}_F F)$	18.31%	3.58%
w/o parents bias	$(\widehat{\text{int}} - P\widehat{\delta}_P)$	19.49%	-2.62%
w/o all individual bias	$(\widehat{\text{int}} - \widehat{\delta}_F F - P\widehat{\delta}_P - \widehat{\delta}_M M)$	18.96%	0.18%

Middle school

Selection based on:		**Diff. with external**	**Improvement**
predicted	$(\widehat{\text{int}})$	21.55%	
w/o GOC	$(\widehat{\text{int}} + \widehat{x}\overline{\text{ext}})$	14.37%	33.35%
w/o school FE	$(\widehat{\text{int}} - \widehat{\sigma})$	17.71%	17.86%
w/o school FE & GOC	$(\widehat{\text{int}} + \widehat{x}\overline{\text{ext}} - \widehat{\sigma})$	12.00%	44.35%
w/o female bias	$(\widehat{\text{int}} - \widehat{\delta}_F F)$	18.93%	12.18%
w/o parents bias	$(\widehat{\text{int}} - P\widehat{\delta}_P)$	21.55%	0.03%
w/o all individual bias	$(\widehat{\text{int}} - \widehat{\delta}_F F - P\widehat{\delta}_P - \widehat{\delta}_M M)$	18.28%	15.19%

Note. Weighted average of results over time (school years 2009/2010 - 2013/2014 for primary school and school years 2011/2012 - 2013/2014 for middle school). Same samples of table 2.

9 Figures

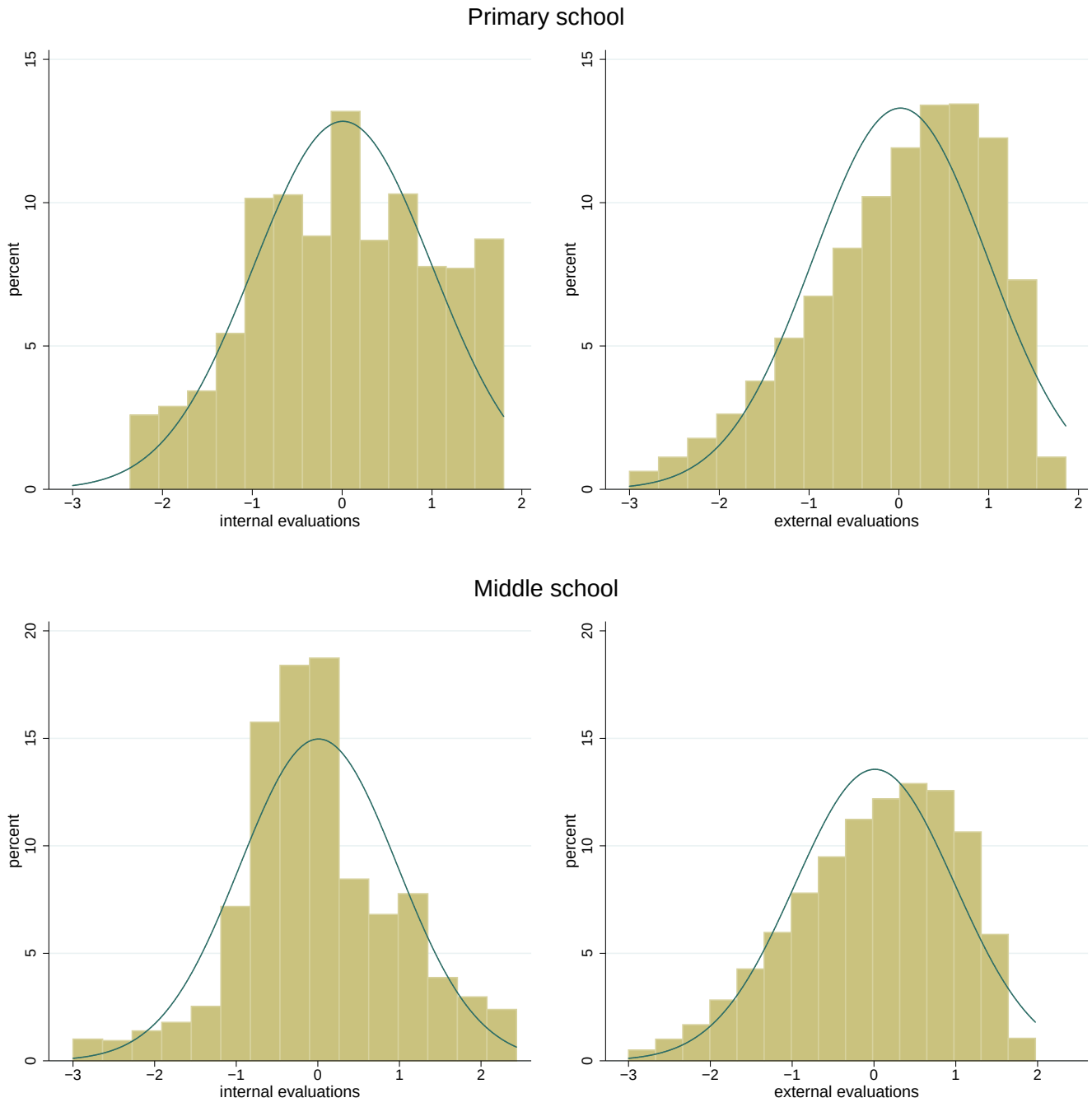

Figure 1: Empirical distribution of internal and external evaluations at the end of primary school (grade 6^{th}) and at the end of middle school (grade 4^{th}). Continuous lines are normal fits.

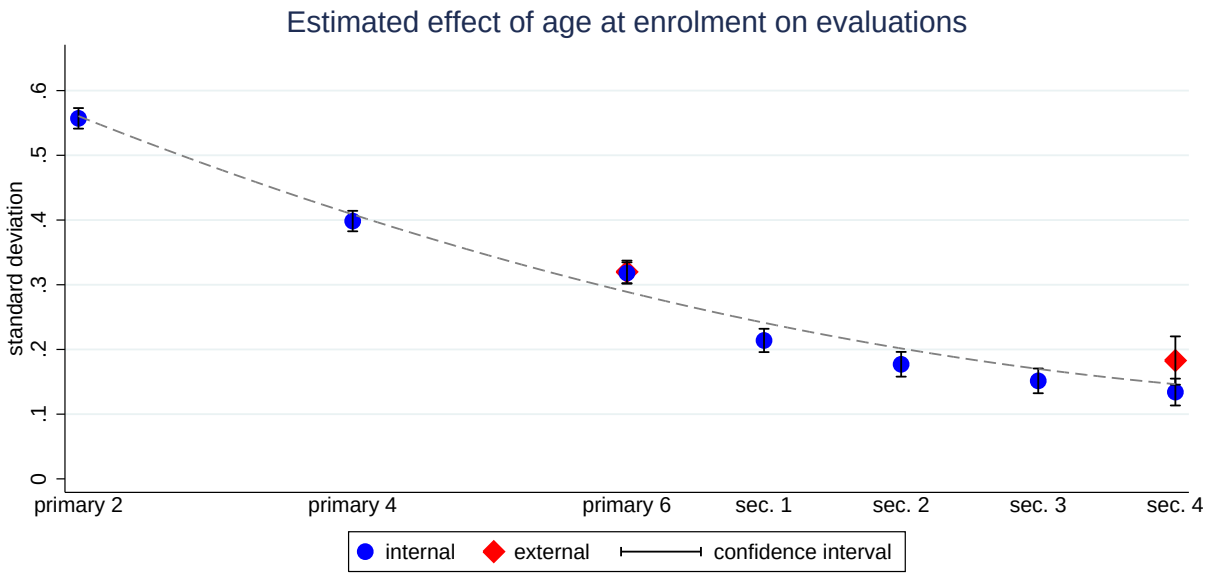

Figure 2: Each dot (diamond) is the estimated coefficient of a regression of internal (external) evaluations on age at enrollment and controls. The dotted line is a quadratic fit of all the estimates.

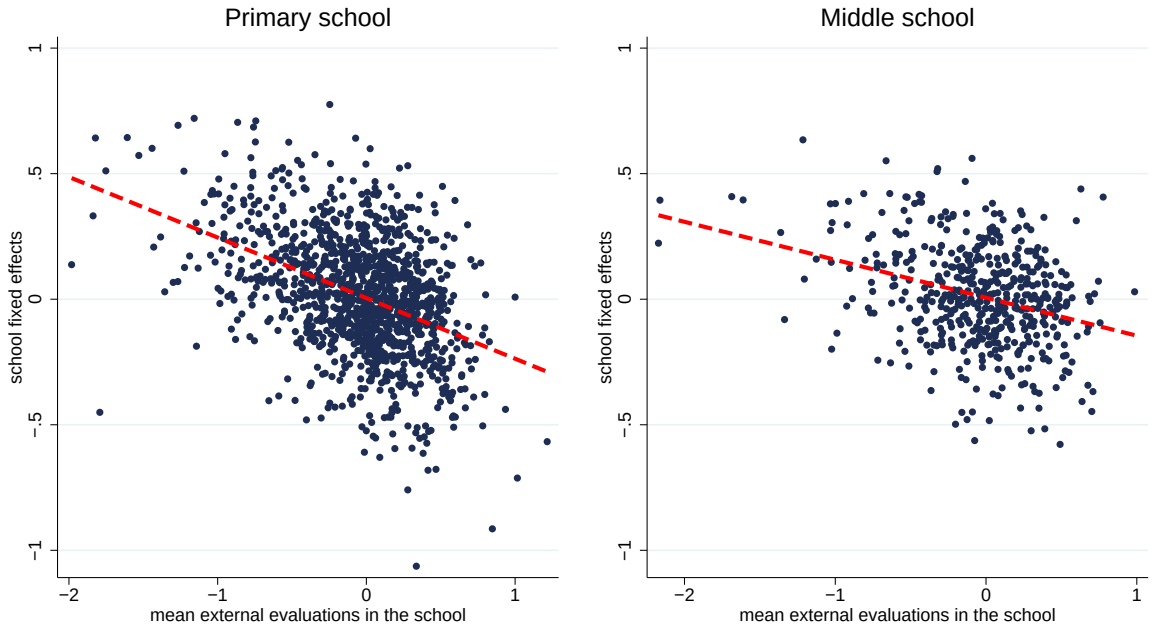

Figure 3: School fixed effects estimated by two stage least square regressions shown in table 2, columns (2sls).

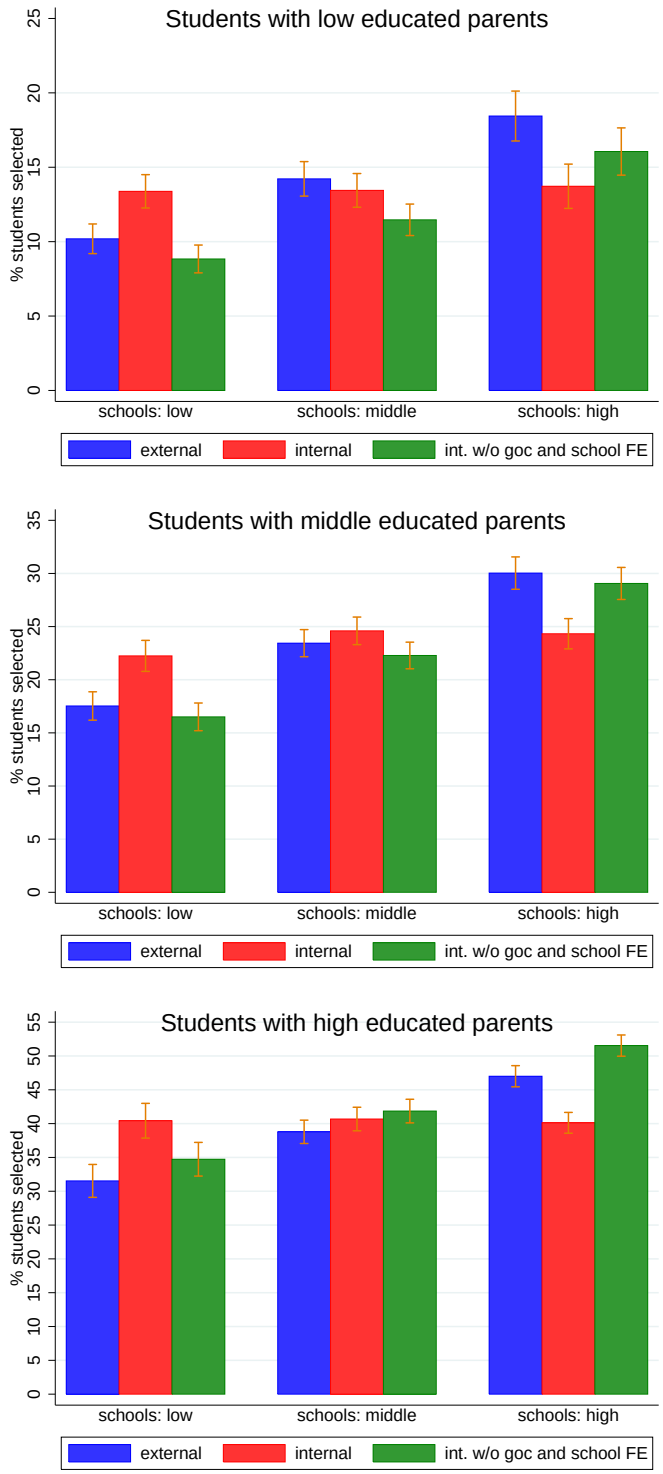

Figure 4: **Simulated selection of top quartile of students at the end of primary school.** The graphs plot share of admitted students under different selection process in 2013. School quality is defined using school average outcomes in external evaluations from 2009 to 2012: low quality schools are in bottom 33% for at least half of the years, and never above 66 percentile; high quality schools belong to the best one third for at least half of the years, and they never belong to the bottom 33%. The top graph concerns students with low educated parents (both have at most middle school diploma), the bottom graph focus on students with high educated parents (at least one with tertiary education and the other with high school diploma).

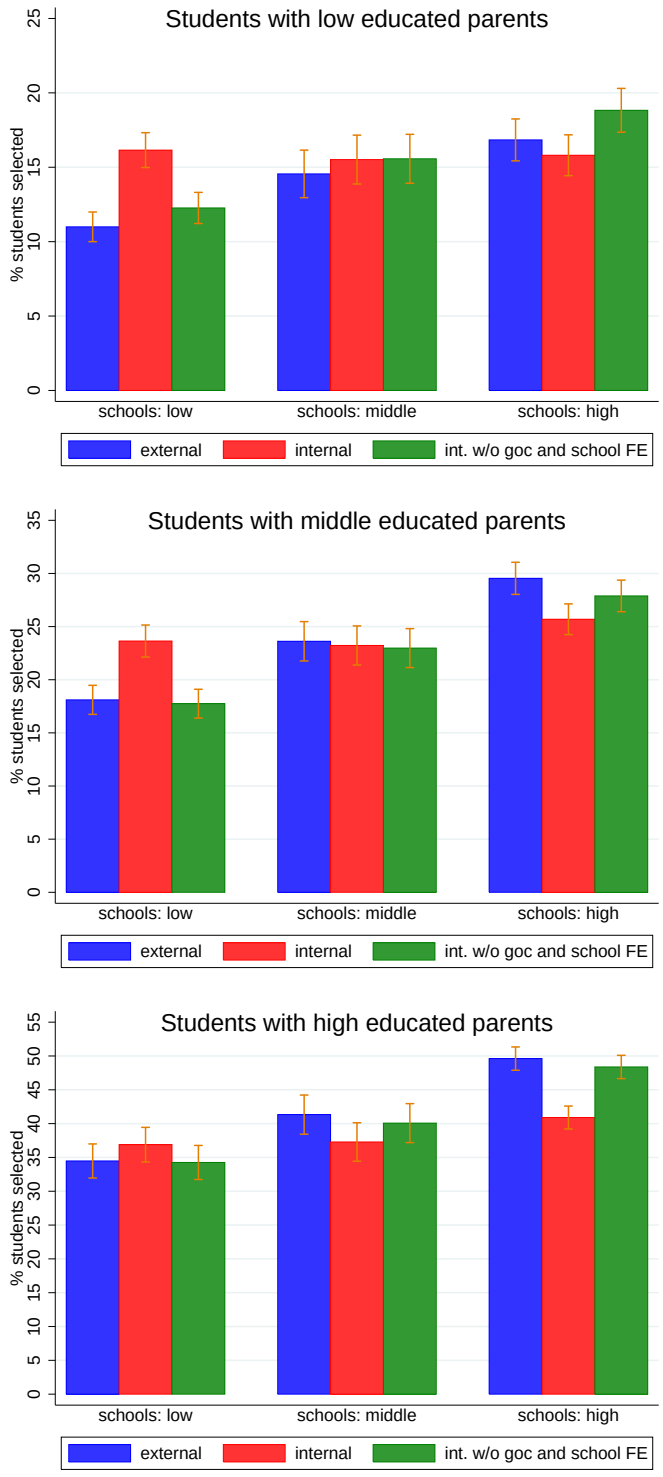

Figure 5: **Simulated selection of top quartile of students at the end of middle school.** The graphs plot share of admitted students under different selection process in 2013. School quality is defined using school average outcomes in external evaluations from 2011 to 2012: low quality schools are in bottom 33% for at least one year, and never above 66 percentile; high quality schools belong to the better one third for at least one year, and they never belong to the bottom 33%. The top graph concerns students with low educated parents (both have at most middle school diploma), the bottom graph focus on students with high educated parents (at least one with tertiary education and the other with high school diploma).

All appendices are for online publication

A Class formation in primary school

Although there is no specific regulation on how children should be allocated in primary school, anecdotal evidence suggests that classes are particularly designed to be homogeneous in the observables. For instance a primary school with two classes for first graders in a given year allocates female students more or less evenly in the two classes. Moreover administrators and teachers use information provided by preschool educators and parents to allocate children so that each class receive a fair number of children that showed high or low ability in the previous years. Therefore although children are not assigned to classes with a random draw, their allocation is balanced and the variation in peers composition across classes is *as good as random* for our purposes.

For each primary school in a given year, we can formally verify that there is no sorting, testing whether students characteristics and the class the student is assigned to are statistically independent. Following the procedure described in Ammermueller and Pischke (2009), we perform Pearson χ^2 test for discrete characteristics such as female, immigrant, parental education.[46] Moreover we implement a Kruskall-Wallis test for age at enrollment, which is a continuous variable.

We replicate the same battery of tests for both first and last grades of primary school, to make sure that not only there is no sorting at the beginning of primary school, but that classes are still balanced in sixth grade.[47].

For each characteristic, we reject at 5% level the null hypothesis of "random" allocation less than 4% of times, both in first and in sixth grade. This percentage drops to 0.5% when gender is the characteristic under analysis. We interpret these results as strong evidence that sorting is not in place in primary school; if anything there are interventions to smooth out differences, designing classes to be homogeneous among them.

A natural question that may arise is then whether there is enough variation across classes (and over time) to properly identify the effect of grading on a curve. The variance decomposition in table A-9 shows that although some characteristics vary more between schools than within, there is a reasonable amount of variation also across classes. The decomposition is computed following Ammermueller and Pischke (2009): first we compute the class averages

[46] Given that sample size for each school is relatively small, we also performed Fisher's exact tests, which do not rely on any asymptotic assumption on the distribution of the variables. We find extremely similar results.

[47] In fact some schools shuffle classes either at the beginning of third or of fifth grade. In our sample less than 20% of primary schools do so.

of each variable, and then we decompose the total variance in these class averages into within school and between school variances.[48]

B Sorting in middle school

This appendix discusses biases that may affect our estimation when students are sorted across classes. To simplify the notation let us rewrite the model in equation (7) as follows

$$\text{int} = \gamma \text{ext} - x\overline{\text{ext}} + X\delta + \overline{X}\beta + \varepsilon \tag{A-1}$$

where we omit the indexes (i and c_i), we ignore school and year fixed effects, and we use the vector X for the individual predetermined characteristics and $\overline{X}$ for their average at the class level. Our goal is to understand how sorting can bias the estimated coefficients of the class-level regressors, particularly of $\overline{\text{ext}}$. As extensively discussed in sections 4.1 and 4.2, to solve measurement error issues we can instrument ext and $\overline{\text{ext}}$ with age at enrollment A and average age at enrollment $\overline{A}$. Obviously the model is well specified if the error term is uncorrelated with the instruments and the other regressors. Therefore sorting affects the coefficient of $\overline{\text{ext}}$ if the other regressors or $\overline{A}$ are correlated with the error term. After instrumenting for ext and $\overline{\text{ext}}$ sorting affects the coefficient in $-\widehat{x}$ in the same way it affects any other coefficient in the vector β. Thus to simplify the exposition from now on we will treat ext as if it is a predetermined regressors, without need to repeat every time that it has been instrumented for.

As explained in sections 2 and 4, the model in (A-1) relies on the assumptions that internal and external evaluations measure the same cognitive skills, but internal evaluations are modified by comparison with peers and biases that are orthogonal to cognitive skills. However there might be unobserved variables related to human capital (say non-cognitive skills) that may affect differently internal and external evaluations. Moreover we do not explicitly model the fact that some teachers may be generally more lenient or stricter than other, above and beyond the "grading on the curve". The following model incorporates these

[48] The formula we use is

$$\frac{1}{N_C}\sum_{s=1}^{S}\sum_{c=1}^{C_s}(x_{cs}-\overline{x})^2 = \frac{1}{N_C}\sum_{s=1}^{S}\sum_{c=1}^{C_s}(x_{cs}-\overline{x}_s)^2 + \frac{1}{N_C}\sum_{s=1}^{S}(\overline{x}_s-\overline{x})^2$$

where x is the variable under analysis, $s = 1, ..., S$ is the school indicator, $c_s = 1, ..., C_s$ is the class indicator (there are C_s classes for school s in our sample), and N_C is the total number of classes in the sample. The first part of the RHS gives the variance within school, the second part the variance between schools. We pool together classes of a given school over time. For instance if school s appears in the sample from 2011 to 2013, with two classes each year, then $C_s = 6$.

two aspects in a simple way:

$$\text{int} = \gamma \text{ext} - x\overline{\text{ext}} + X\delta + \overline{X}\beta + \psi N + T_k + \eta \tag{A-2}$$

where N is a measure of non-cognitive skills and T_k is teacher fixed effects, namely a shift up or down of the evaluation depending on the leniency of teacher k. Thus in equation (A-1) $\varepsilon = \psi N + T_k + \eta$. If students are randomly allocated to classes regressor are uncorrelated with ε, and equation (A-1) can be consistently estimated. Conversely if students are sorted across classes, the correlation need not to be 0. Suppose that there are "more difficult" and "easier" classes, and students are allocated based on their cognitive and non cognitive skills. Then a student in a class with high average external evaluations probably belong to a "more difficult" class, thus she probably has high cognitive or non cognitive skills (or both). In particular $E(N|\overline{\text{ext}} = y_1) > E(N|\overline{\text{ext}} = y_2)$ if $y_1 > y_2$ and therefore $cor(\overline{\text{ext}}, \eta) \neq 0$. Abstracting from teacher effects, the correlation is surely positive, and would upward bias the coefficient of $\overline{\text{ext}}$; thus $\widehat{x}$ would be downward biased, namely it would underestimate the true effect of GOC.

If students are grouped according to some characteristics, we cannot exclude that the assignment of teachers as well is non-random. This would be a problem if $\overline{\text{ext}}$ (and other average characteristics in the class) are correlated with T_k. In particular if $cor(\overline{\text{ext}}, T_k) < 0$, for instance because stricter teachers are assigned to high performing classes, then $cor(\overline{\text{ext}}, \eta)$ may be negative. Consequently the coefficient of $\overline{\text{ext}}$ would be downward biased, and $\widehat{x}$ would overestimate the true effect of GOC.

C Heterogeneous effects

As often found in the literature, female students in our sample tend to perform better in the languages when undertaking external evaluations. Conversely boys over perform girls in mathematics – see, for example, Goldin, Katz, and Kuziemko (2006). As discussed in previous section, internal evaluations grant girls a positive bias over and above their average better performances in external evaluations.

A possible concern is that internal or external evaluations capture human capital in a somehow different way for boys and girls. We test whether there are heterogeneous effect by gender in table A-7 in appendix D. We simply estimate the baseline model (columns (2sls) of table A-7) on the two subsamples of female and male students; results in table A-7 show that the estimated rate of grading on the curve is similar across genders for boys and girls. The estimated coefficient for external evaluations is close to 1, although it is somehow larger

and significantly different from 1 for girls in middle school.[49]

Hoxby and Weingarth (2005) argue that the linear-in-means model is not necessarily the right model of peer effects. More recent findings such as Burke and Sass (2013), Imberman, Kugler, and Sacerdote (2012) or Carrell, Sacerdote, and West (2013), among others, find heterogeneous effects of peers. Our source of peer effect is quite different in nature, but it is still worth studying whether the effect is heterogeneous for students of different ability levels. Hence, in this section we complete the analysis by differentiating three groups of students who have different expected human capital, given their observable characteristics. In particular we predict GPA in external evaluations using only individual exogenous regressors (age at enrollment, parental education, gender, foreign born, and school year). Then for each school year we classify students in three quantiles: low, middle and high predicted GPA.[50] The classification is quite rough, given that we rely only on time invariant characteristics and we cannot control for past measure of ability. In fact the R^2 of the regression is only 0.14 ; nonetheless actual and predicted GPA have a positive and significant correlation (0.37).

We perform the usual specification on the three subsamples. Table A-8 display the estimated coefficients for own evaluation and average in the class. Results are qualitatively similar across groups. All the coefficients of $\overline{\text{ext}}_i$ are statistically significant at least at 10% level, besides one whose p-value is 0.101. In primary school we cannot reject the null hypothesis that they are all 0.40, in middle we cannot reject the null hypothesis that they are all 0.5.[51]

D Additional tables

[49] In middle school the coefficient of $\overline{\text{ext}}_i$ is slightly larger for girls than for boys, but the ratio with the coefficient of ext_i is about 0.5 for both.

[50] Grouping students on the basis of their predicted outcome is similar to what Carrell, Sacerdote, and West (2013) do.

[51] For primary school the coefficient for own evaluation is significantly smaller than 1 for group "low" and significantly larger for group "high" but the differences are very small in magnitude. In middle the coefficient for the group "high" is significantly larger than 1. As shown in figure 5, the distribution of internal and external is more similar in the middle, while externals have a larger negative skewness; this may explain this small difference across subsamples. Getting rid off 486 observations whose external are lower than -3 (about 1% of the sample) the coefficient for group "low" in primary school is not significantly different from 1.

Table A-1: Effect of entry age over time.

	prim. 2	prim. 4	prim. 6 (int.)	prim. 6 (ext.)	sec. 1	sec. 2	sec. 3	sec. 4 (int.)	sec. 4 (ext.)
age at enrolment	0.557	0.398	0.318	0.320	0.214	0.177	0.151	0.134	0.183
	(0.008)**	(0.008)**	(0.008)**	(0.009)**	(0.009)**	(0.010)**	(0.010)**	(0.011)**	(0.019)**
female	0.166	0.215	0.290	0.126	0.331	0.328	0.331	0.317	0.050
	(0.005)**	(0.005)**	(0.005)**	(0.005)**	(0.007)**	(0.007)**	(0.007)**	(0.007)**	(0.012)**
parents M	0.398	0.427	0.412	0.348	0.368	0.347	0.294	0.221	0.321
	(0.006)**	(0.007)**	(0.007)**	(0.007)**	(0.008)**	(0.009)**	(0.009)**	(0.008)**	(0.014)**
parents H	0.744	0.808	0.821	0.655	0.779	0.760	0.677	0.572	0.661
	(0.008)**	(0.008)**	(0.009)**	(0.008)**	(0.013)**	(0.012)**	(0.012)**	(0.012)**	(0.018)**
missing parents	0.175	0.176	0.189	0.182	0.153	0.137	0.095	0.067	0.180
	(0.015)**	(0.015)**	(0.015)**	(0.015)**	(0.018)**	(0.021)**	(0.020)**	(0.019)**	(0.035)**
immigrant	-0.336	-0.342	-0.344	-0.304	-0.314	-0.292	-0.326	-0.396	-0.619
	(0.010)**	(0.010)**	(0.010)**	(0.010)**	(0.012)**	(0.012)**	(0.012)**	(0.012)**	(0.022)**
R^2	0.21	0.21	0.22	0.25	0.20	0.18	0.17	0.16	0.30
N	163,982	151,582	139,486	120,018	110,500	105,897	98,448	87,472	24,867

$+\ p < 0.1$; $*\ p < 0.05$; $**\ p < 0.01$

Note. This table replicates finding of Calsamiglia and Loviglio (2016). In each column the dependent variable is internal or external GPA (average of evaluations in Mathematics, Spanish, Catalan, English) at the end of the school year. School and cohort fixed effects are included.

Table A-2: First stage estimates

	primary school		middle school	
	external ev.	avg ext. ev.	external ev.	avg ext. ev.
entry age	0.327	0.000	0.104	0.000
	(0.009)**	(0.003)	(0.011)**	(0.005)
avg entry age	-0.081	0.239	0.912	1.017
	(0.044)	(0.012)**	(0.052)**	(0.022)**
female	0.138	0.001	0.021	0.001
	(0.005)**	(0.001)	(0.006)**	(0.003)
immigrant	-0.320	-0.007	-0.415	-0.007
	(0.008)**	(0.002)**	(0.010)**	(0.004)
parents M	0.350	0.007	0.219	0.022
	(0.006)**	(0.002)**	(0.007)**	(0.003)**
parents H	0.666	0.004	0.468	0.007
	(0.007)**	(0.002)*	(0.009)**	(0.004)
missing parents	0.226	-0.004	0.119	-0.008
	(0.014)**	(0.004)	(0.016)**	(0.007)
share female	0.020	0.158	0.304	0.320
	(0.031)	(0.009)**	(0.029)**	(0.012)**
share immigrant	-0.087	-0.364	-1.581	-1.937
	(0.040)*	(0.011)**	(0.040)**	(0.017)**
avg parents edu.	0.022	0.186	0.565	0.682
	(0.010)*	(0.003)**	(0.010)**	(0.004)**
Constant	-0.502	-0.440	-1.383	-1.361
	(0.032)**	(0.009)**	(0.034)**	(0.014)**
N	127,082	127,082	73,899	73,899

Note. First stages of 2sls regressions shown in table 2, columns (2sls). "entry age" is the student's (expected) age at enrollment in first grade of primary school. This variable has been scaled in the interval $[0, 1]$ (it is 1 for a child born on January, 1; it is 0 for a child born on December, 31). "avg entry age" is the mean value at the class level. Other regressors, school and year fixed effects are as described in table 2.

Table A-3: Estimates by subject

	primary school				middle school			
	Maths	Spanish	Catalan	English	Maths	Spanish	Catalan	English
external ev.	1.213	0.984	1.051	1.137	1.059	1.105	1.304	1.090
	(0.034)**	(0.024)**	(0.026)**	(0.030)**	(0.155)**	(0.112)**	(0.144)**	(0.126)**
avg external ev.	-0.258	-0.449	-0.449	-0.344	-0.486	-0.600	-0.621	-0.517
	(0.270)	(0.157)**	(0.186)*	(0.167)*	(0.168)**	(0.133)**	(0.165)**	(0.136)**
female	0.258	0.164	0.137	0.043	0.507	0.227	0.205	0.178
	(0.008)**	(0.007)**	(0.007)**	(0.009)**	(0.047)**	(0.020)**	(0.026)**	(0.019)**
immigrant	0.047	0.025	0.038	-0.082	0.182	0.296	0.485	0.047
	(0.012)**	(0.012)*	(0.012)**	(0.008)**	(0.049)**	(0.053)**	(0.082)**	(0.023)*
parents M	0.031	0.055	0.034	0.044	-0.016	-0.029	-0.067	-0.028
	(0.011)**	(0.009)**	(0.010)**	(0.010)**	(0.025)	(0.019)	(0.025)**	(0.025)
parents H	0.096	0.161	0.116	0.093	0.050	0.075	-0.020	0.018
	(0.020)**	(0.014)**	(0.017)**	(0.018)**	(0.061)	(0.036)*	(0.052)	(0.058)
missing parents	0.035	0.048	0.045	0.036	-0.003	0.060	0.023	-0.011
	(0.015)*	(0.013)**	(0.013)**	(0.013)**	(0.026)	(0.020)**	(0.022)	(0.025)
share female	-0.052	-0.018	-0.044	-0.050	-0.206	-0.036	-0.132	-0.080
	(0.053)	(0.036)	(0.043)	(0.061)	(0.057)**	(0.051)	(0.059)*	(0.041)+
share immigrant	-0.026	-0.156	-0.094	0.019	0.208	-0.072	0.116	0.147
	(0.086)	(0.069)*	(0.080)	(0.047)	(0.116)	(0.147)	(0.175)	(0.089)+
avg parents edu.	-0.014	0.037	0.028	-0.008	-0.051	0.071	-0.014	-0.020
	(0.039)	(0.025)	(0.032)	(0.030)	(0.043)	(0.039)	(0.045)	(0.037)
Constant	-0.116	-0.171	-0.123	-0.012	-0.141	-0.232	-0.068	-0.043
	(0.035)**	(0.045)**	(0.059)*	(0.072)	(0.045)**	(0.056)**	(0.065)	(0.053)
N	127,082	127,082	127,082	127,082	73,899	73,899	73,899	73,899

Note. Dependent variable is the internal evaluation in the subject reported above the column; similarly for individual and average external evaluations. Other regressors, school and year fixed effects are as in table 2.

Table A-4: Alternative specifications

	primary school			middle school		
	(1)	(2)	(3)	(1)	(2)	(3)
external ev.	1.038			1.175		
	(0.018)**			(0.100)**		
avg external ev.	-0.648	-0.610	-0.330	-0.740	-0.566	-0.448
	(0.019)**	(0.016)**	(0.114)**	(0.099)**	(0.014)**	(0.044)**
female	0.151	0.156	0.156	0.357	0.360	0.360
	(0.004)**	(0.004)**	(0.003)**	(0.006)**	(0.007)**	(0.006)**
immigrant	0.006	-0.006	-0.004	0.270	0.199	0.199
	(0.008)	(0.007)	(0.005)	(0.042)**	(0.013)**	(0.009)**
parents M	0.045	0.058	0.056	-0.056	-0.022	-0.024
	(0.007)**	(0.005)**	(0.004)**	(0.021)**	(0.007)**	(0.007)**
parents H	0.127	0.152	0.151	-0.019	0.062	0.061
	(0.013)**	(0.005)**	(0.005)**	(0.047)	(0.010)**	(0.008)**
missing parents	0.044	0.052	0.054	-0.008	0.014	0.015
	(0.010)**	(0.009)**	(0.009)**	(0.020)	(0.016)	(0.015)
share female	-0.008	-0.013	-0.056	-0.107	-0.109	-0.150
	(0.020)	(0.032)	(0.026)*	(0.028)**	(0.040)**	(0.030)**
share immigrant	-0.166	-0.156	-0.055	-0.135	-0.076	0.160
	(0.026)**	(0.041)**	(0.048)	(0.053)*	(0.063)	(0.095)+
avg parents edu.	0.065	0.059	0.007	0.084	0.064	-0.018
	(0.007)**	(0.010)**	(0.022)	(0.016)**	(0.017)**	(0.032)
Constant	-0.197	-0.199	-0.110	-0.235	-0.240	-0.137
	(0.015)**	(0.025)**	(0.040)**	(0.024)**	(0.036)**	(0.044)**
N	127,082	127,082	127,082	73,899	73,899	73,899

Note. Dependent variable in regression (1) is student's internal evaluations, as in table 2; external evaluations is instrumented with entry age, while average external evaluations in the class is *not* instrumented. Dependent variable in regressions (2) and (3) is the difference between internal and external evaluations. In (3) average external evaluations in the class is instrumented with average entry age. Other regressors, school and year fixed effects are as in table 2.

Table A-5: Robustness check (I & III)

	primary school		middle school	
	(1)	(2)	(1)	(2)
external ev.	1.032	1.030	1.265	0.890
	(0.017)**	(0.018)**	(0.107)**	(0.059)**
avg external ev.	-0.587	-0.313	-0.538	-0.592
	(0.114)**	(0.128)*	(0.297)+	(0.309)+
N	106,949	122,951	26,097	70,195

Note. Dependent variable is student's internal evaluations (GPA). Columns (1) show results of the baseline specification performed on the subset of classes in which the rank correlation between external and internal evaluations is larger than 0.75 (other regressors, school and year fixed effects are as in table 2).
In columns (2) average external evaluations is the mean of students' evaluation at the school level in a given year (rather than at the class level). The instrument is average entry age at the school level. Other regressors and school fixed effects are as in table 2 (average are computed at the school level).

Table A-6: Robustness check (II)

		middle school		
	Spanish	Spanish	Math	Math
	(1)	(2)	(1)	(2)
external ev.	1.117	1.195	1.100	1.081
	(0.110)**	(0.134)**	(0.157)**	(0.182)**
avg external ev.	-0.494	-0.571	-0.499	-0.446
	(0.151)**	(0.171)**	(0.181)**	(0.207)*
N	72,056	54,627	72,044	51,259

Note. Dependent variable is student's internal evaluations in either Spanish or Mathematics. Students' evaluation (in either Spanish or Mathematics) and average external evaluations in the class are instrumented with entry age and average entry age. Regressions (1) include dummies for Spanish or Maths teachers. Regressions (2) works similarly, but the sample is restricted to classes that have a unique Spanish or Math teacher associated. Other regressors, school and year fixed effects are as in table 2).

Table A-7: Analysis by gender

	primary school		middle school	
	girls	boys	girls	boys
external ev.	1.043	1.016	1.309	0.953
	(0.023)**	(0.026)**	(0.146)**	(0.129)**
avg external ev.	-0.372	-0.340	-0.694	-0.476
	(0.176)*	(0.158)*	(0.172)**	(0.134)**
N	62,940	64,142	37,581	36,318

Note. Regressions are performed separately on the two subsamples of female and male students. Regressors, school and year fixed effects are as in table 2.

Table A-8: Analysis by predicted performances

	primary school			middle school		
	low	middle	high	low	middle	high
external ev.	0.918	1.029	1.118	1.069	0.942	1.368
	(0.035)**	(0.038)**	(0.034)**	(0.374)**	(0.146)**	(0.172)**
avg external ev.	-0.336	-0.476	-0.333	-0.631	-0.298	-0.669
	(0.192)+	(0.215)*	(0.203)	(0.352)+	(0.174)+	(0.190)**
N	41,949	41,943	43,190	24,397	24,407	25,095

Note. Regressions are performed separately on three subsamples: students with low, middle, and high predicted external evaluations. Regressors, school and year fixed effects are as in table 2.

Table A-9: Variance decomposition

Variable	**Between**	**Within**	**Total**
GPA external	0.184	0.076	0.259
	70.8%	29.2%	
GPA internal	0.071	0.056	0.127
	56.2%	43.8%	
A	0.001	0.003	0.004
	18.1%	81.9%	
parents	0.377	0.074	0.450
	83.7%	16.3%	
female	0.002	0.007	0.009
	21.0%	79.0%	
migrant	0.027	0.006	0.033
	81.1%	18.9%	