Economics of Education Review 73 (2019) 101916

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Grading on a curve: When having good peers is not good[☆]

Caterina Calsamiglia[*,a], Annalisa Loviglio[b]

[a] ICREA and IPEG, Spain
[b] Department of Economics, University of Bologna, Italy

ARTICLE INFO

Keywords:
Grading on a curve
Negative peer effects
School choice

JEL classification:
I21
I28
H75

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. This puts forth a source of distortion that may arise in any system that uses internal grades to compare students across schools and classes. We find suggestive evidence that school choice is impacted only the year when internal grades matter for future prospects.

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, such as 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 out 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.

Internal evaluations capture human capital accumulation (cognitive skills), as external evaluations do, but may also capture teachers' bias. Lavy (2008), Lavy and Sand (2015), Rangvid (2015), and Terrier (2016) provide empirical evidence that teachers exhibit a gender bias, often providing differential grades to females or minorities, 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.[1]

This paper provides empirical evidence of an additional source of disparity between internal and external grades and a channel through which having better peers *need not* be beneficial. 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 these facts presents large challenges,

[☆] 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 European Research Council Starting Grant 638893. Annalisa Loviglio acknowledges financial support of La Caixa Foundation (La Caixa-Severo Ochoa International Doctoral Fellowship).

[*] Corresponding author.
E-mail addresses: caterina.calsamiglia@barcelona-ipeg.eu (C. Calsamiglia), annalisa.loviglio@gmail.com (A. Loviglio).

[1] As shown in Betts and Grogger (2003) and Figlio and Lucas (2004) grading standards by teachers may also impact student performance and attainment.

https://doi.org/10.1016/j.econedurev.2019.101916
Received 15 October 2018; Received in revised form 7 July 2019; Accepted 28 July 2019
Available online 09 August 2019

both in terms of identification and data requirements. 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] We show that internal scores given by teachers are decreasing in the class average of external evaluations. That is, the internal evaluations are smaller with respect to the external evaluations as peer performance in external evaluations increases. This suggests that teachers value students less in a classroom with better peers. We control for school fixed effects to address selection of students into schools and 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 run a set of robustness checks to confirm the persistence of our results throughout primary and secondary school. Moreover we implement various robustness checks to test the validity of our assumption that internal and external evaluations are capturing similarly individual skills.

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 most 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 potentially 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 students can be affected by the quality of their 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.

What are evaluations in school important for? First, they provide feedback to students and their family on their ability and proficiency in a given subject. The informational content of grades and the attitudes of teachers towards individuals in class can 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.[4]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; Tincani (2015) shows that rank concerns may generate heterogeneous peer effects; Murphy and Weinhardt (2018) shows that the rank of a student in a class in primary school impacts performance in secondary school, when peers and teachers have changed – they provide survey data consistent with the reason being that students who are high in the rank improve self-confidence and therefore performance in the future. Similarly, Elsner and Isphording (2017) show that rank in the classroom has long run effects on achievement. Hence, grades in the classroom may affect students' perceived ability, future expectations and performance.[5] Our work speaks to this literature providing evidence that some of the signals that students receive on their ability may be distorted by peers' composition and school standards. This may affect their self-perception and educational choices even if their assessment is unrelated to their position in the class.

But evaluations in school 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.[6] 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 using our data in Catalonia. We find that the 25% top performing students are very different if selected through grades in internal or external evaluations. In particular, more than 30% of those selected through internal grades do not get selected through external grades; vice-versa more than 30% of those selected through external grades do not get selected through internal grades. Of these initial differences, about one third (10 p.p.) is due to 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.[7] 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 (2018) 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, Gall, Legros, and Newman (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 theoretical framework

[2] Catalonia is one of the most prosperous autonomous communities in Spain with more than seven million citizens. The Catalan government has the power to legislate in matters such as health or education, among others.

[3] Burke and Sass (2013), Carrell, Sacerdote, and West (2013), and Feld and Zölitz (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] Ahn, Arcidiacono, Hopson, and Thomas (2016) show that grading policies in college may affect major choice.

[5] Mayer and Jencks (1989) reviews the sociology literature and states that living in an advantageous neighborhood may be disadvantageous, because a given student will rank worse if in an advantageous neighborhood, which may affect his or her expectations.

[6] 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.

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

to describe how external and internal grades are generated. Section 3 describes the data. Section 4 contains the empirical strategy and the results. Section 5 discusses robusntess checks. Section 6 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 7 discusses strategic change of school at the end of low secondary education. Section 8 concludes.

2. A simple model for internal and external evaluations

In this illustrative framework 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:

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

where $\text{Cov}(\varepsilon_E, H) = 0$.

We assume that internal evaluations capture the same cognitive skills, but may also be affected by biases 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).[8] Moreover we allow teachers to consider both absolute and relative performance when they assign evaluations.

$$\begin{aligned} \text{int} &= (1-\xi)H + \xi(H - \overline{H}) + \delta F + \varepsilon_I \\ &= H - \xi\overline{H} + \delta F + \varepsilon_I \end{aligned} \tag{2}$$

where the error term ε_I is uncorrelated with both H and F. $\xi \in [0, 1]$ and $1 - \xi$ are weights given to relative and absolute performances respectively. Ignoring for now the contribution of δF, if $\xi = 0$, 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 $\xi = 1$, the internal evaluation is based only on relative performance, as measured by the distance from the mean in the class. $\xi \in (0, 1)$ means that both absolute and relative performance contribute to the final grade. In other words, 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 ξ in Eq. (2) tells us the relevance of grading on a curve in the school system under analysis.

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.

The core of our empirical analysis follows from Eqs. (1) and (2). Human capital is not observed, while the evaluations are. Hence, we can derive H from Eq. (1) and substitue H and $\overline{H}$ into (2):

$$\text{int} = \text{ext} - \xi\overline{\text{ext}} + \delta F + \varepsilon_I - \varepsilon_E + \xi\bar{\varepsilon}_E = \text{ext} - \xi\overline{\text{ext}} + \delta F + \varepsilon \tag{3}$$

$$\text{int} - \text{ext} = -\xi\overline{\text{ext}} + \delta F + \varepsilon \tag{4}$$

The baseline specification that we brings to the data in the empirical analysis in Section 4.1 is based on Eq. (4).[9] While using Eq. (4) rather than Eq. (3) solves the issue of the correlation between ext and the error term ε, any correlation between $\overline{\text{ext}}$ and the error may bias the estimation of ξ. In Appendix A we extensively discusses size and direction of potential biases, concluding the bias is negligible as far as external evaluations accurately measure the underlying human capital, and the errors in the class are not too correlated. In Appendix C we propose an instrument for $\overline{\text{ext}}$ to take care of any correlation between $\overline{\text{ext}}$ and the error and test the validity of our baseline approach. In Appendixes A.2 and Appendix C we also formulate and test an alternative specification to relax the assumption that internal and external evaluations measure the same cognitive skills, allowing them to capture human capital differently.

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.[10] 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.

Within school students are allocated to classes; students in a given class spend almost all the school time together, take all the core subjects together and therefore are exposed to the same set of teachers and teaching methods.[11] In our sample mean class size is 22.2 for primary school and 25.1 for middle school (medians are 23 and 26 respectively).

The core of our analysis (Sections 4–6) 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 in public primary schools in Catalonia from school year 2009/2010 to school year 2013/2014, and students enrolled in fourth grade in public middle schools in Catalonia from school year 2011/2012 to school year 2013/2014.[12] In Section 7 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.[13] The *Departament d'Ensenyament* (regional ministry of education in Catalonia)

[8] 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.

[9] It is worth stressing that although ext is part of the dependent variable and its mean $\overline{\text{ext}}$ is a regressor, the model does not suffer from the so-called "reflection problem" (Manski, 1993). As opposed to the case in which a variable y is regressed on its mean $\bar{y}$, here the expected value of ext is not a linear function of the expected value of the other regressors. There is no mechanical relation between the dependent variable and the expected value of $\overline{\text{ext}}$. To see this, imagine that $\xi = 0$, i.e. internal evaluations measure absolute performances. If human capitals in the class are correlated, increasing the expected value of average external in the class implies that it is more likely to have a higher human capital at the individual level. Under our assumptions this would affect exactly in the same way the expected internal and the expected external evaluation, leaving their expected difference (i.e. the dependent variable) unchanged.

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

[11] Mathematics, Spanish, Catalan, and English, which are the core subjects exploited in this work are always attended together by all students in the class. In middle school a small number of elective subjects may be attended by only a subset of students in the class, but their evaluations are not part of this study.

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

[13] All data sources have been anonymized by the Institut Català d'Estadistica (IDESCAT). They provided us with unique identifiers to merge them.

Table 1
Descriptive statistics .

School year	Parents' education				Female	Immigrant	*N* students	*N* schools
	low	*middle*	*high*	*missing*				
Primary school – 6th 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 – 4th 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

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.[14]

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, nationality, and 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.[15] 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 6th grade of primary school and 4th grade of middle school.[16] 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 are low stakes. They do not have a direct impact on student evaluations or progress to the next grades but they are transmitted to the principal of the school, who forwards them to the teachers, families and students. 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 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.[17]

Finally we collect information on the student's family background, more specifically on parental education from the 2001 Population and Housing Census and local register data (*Padró*).[18]

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

4. Empirical analysis

4.1. Baseline specification

Our analysis follows from the following empirical specification:

$$(\text{int}-\text{ext})_i = \overbrace{-\xi\overline{\text{ext}}_{c_i} + \sigma_{s_i}}^{\text{grading standards}} + \overbrace{\delta_F F_i + \mathbf{P}_i\delta_P + \delta_M M_i}^{\text{individual-level biases}} \overbrace{+\overline{\mathbf{X}}_{c_i}\beta}^{\text{class characteristics}} + \overbrace{\tau_i}^{\text{year}} + \varepsilon_i \tag{5}$$

where student i attends public school s_i in class c_i, and receives internal evaluations int_i and external evaluations ext_i. The dependent variable is the difference between internal and external evaluations $(\text{int}-\text{ext})_i$. The right hand side of Eq. (5) includes average external evaluations in the class ($\overline{\text{ext}}_{c_i}$), school fixed effects (σ_{s_i}), dummies for gender (F_i), foreign born status (M_i), a vector of dummies for parental education ($\mathbf{P}_i$, level of education is low, medium or high, or missing), the average of those characteristics at the class level (vector $\overline{\mathbf{X}}_{c_i}$), and year fixed effects (τ_i).

We study separately students in 6th grade of primary school and 4th grade of middle school. Both int_i and ext_i are computed as average of four subjects: Maths, Catalan, Spanish and English ("GPA" from now on). We use z-scores for each year. Using GPA rather than running separate analysis by subject is particularly convenient for two reasons. On

[14] 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.

[15] 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.

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

[17] We chose these two thresholds in order to keep approximately 80% of the observations for both levels of school. We replicate all the analyses 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.

[18] 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).

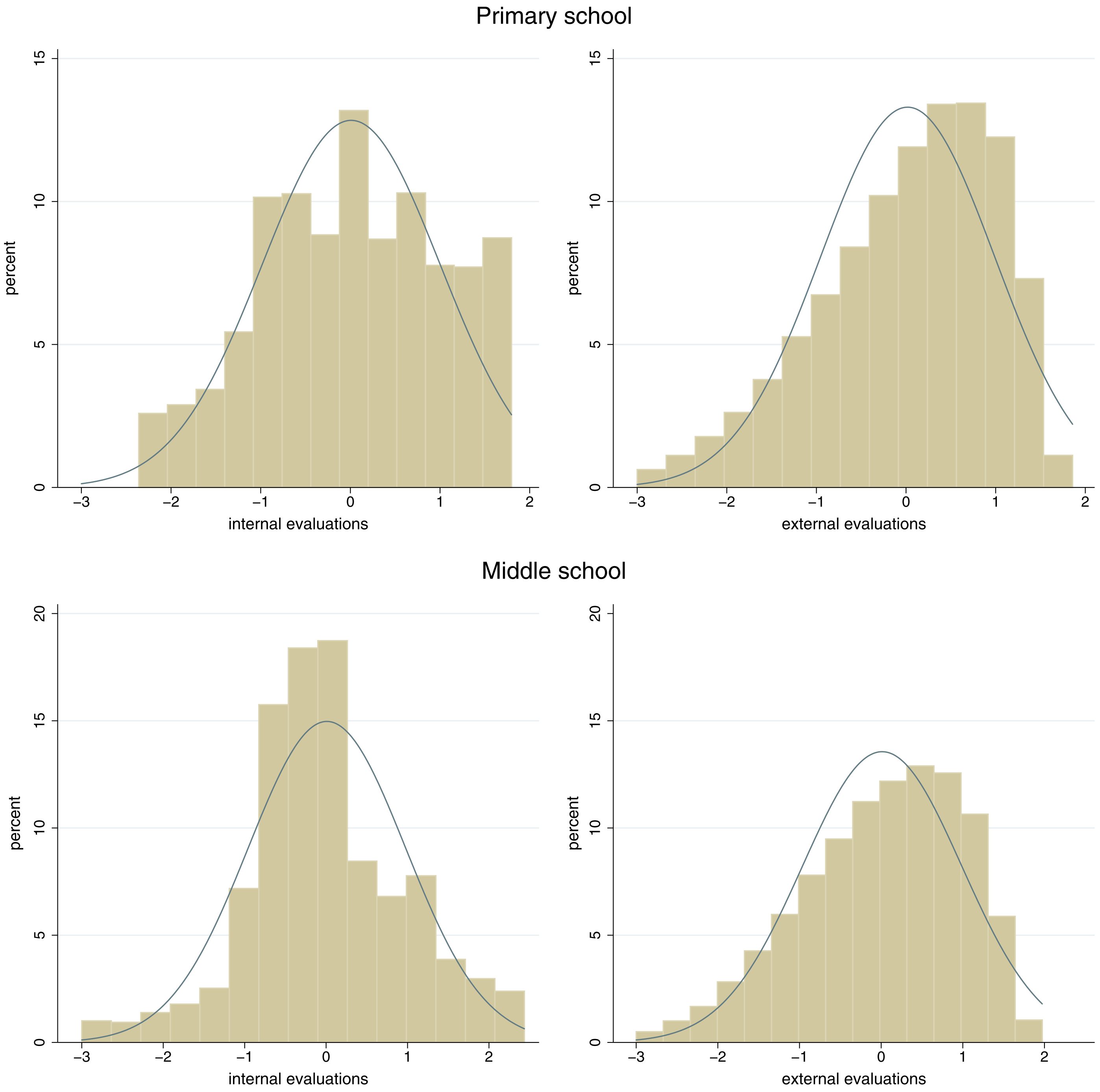

Fig. 1. Distribution of evaluations. Empirical distribution of internal and external evaluations at the end of primary school (grade 6th) and at the end of middle school (grade 4th). Continuous lines are normal fits. Evaluations are average of four subjects (Maths, Spanish, Catalan, English), standardized at the year level (mean 0, sd 1 among the observations in a given year).

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.[19] We also discuss results when analyses are performed by subjects. Main findings are unchanged.

The coefficient of $\overline{\text{ext}}_{cj}$, the average external evaluations in the class, will

[19] 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

(footnote continued)
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 are replicated using this second approach results are extremely similar.

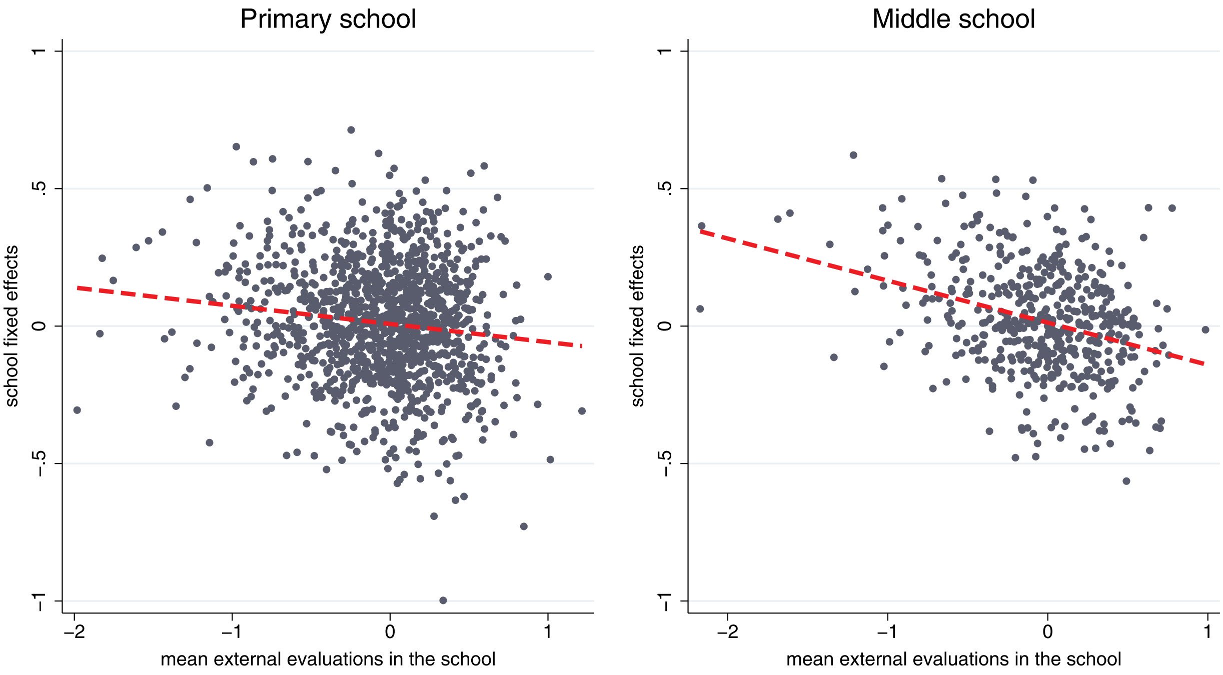

Fig. 2. Estimated school effects. School fixed effects estimated by the regressions shown in Table 2, columns (2).

allow us to estimate the rate ξ of grading on the curve. School fixed effects, σ_{Si}, 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, and depending on the requirements or objectives that are set at the school level, 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*.

Eq. (5) includes a few individual characteristics: gender, foreign born status, parental education; 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 $\overline{\mathbf{X}}_{ci}$). Including regressors in $\overline{\mathbf{X}}_{ci}$ 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.

Eq. (5) includes also year fixed effects (τ_i). Standard errors are clustered at the class level to allow for unobserved correlation of errors of students attending the same class.[21]

We prefer to use class-level average variables, rather than mean among peers in the class, although this might produce a small error in the estimation of ξ, as detailed in Appendix A. In fact it appears sensible to assume that teachers have a unique reference point (the "average performance") and compare each child with it, rather than changing reference point for every student. We replicated all the analysis described in the paper using both mean variables and leave-out mean variables. Given that results are extremely similar, we show and discuss here estimates from specifications with mean variables, which are closer to our theoretical framework.[22]

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 if students are not randomly matched to their peers. The same issue applies to $\overline{\text{ext}}_{ci}$. In Section 5.1 we extensively discuss children allocation within classes of the same school and potential issues related to sorting of students across classes in secondary education. We run several robustness checks that confirm our finding.

4.2. Results

Columns (2) of Table 2 present the results of the estimation of Eq. (5) for primary and middle school. For comparison in columns (1) only average external evaluations and school dummies are used as regressors.

Coefficients of average external evalutions are −0.61 for primary school and −0.57 for middle school, both significant at the 1%. Estimates are very similar, just slightly smaller in magnitude in columns (1) where we do not control for other regressors. Having in mind the simple model introduced in Section 2, we can deduce from the coefficients of average external evaluations that the estimated rate of grading on a curve $\hat{\xi}$ in the Catalan school system is more than 50%.

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 standard deviations

[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.

[21] We prefer to cluster at the class level because this is the fundamental unit of our analysis. One might worry about remaining correlation of errors across classes of the same school; in practice we estimate the standard errors clustering at the school level and the significance level never change.

[22] Results with leave-out mean variables are available upon request to the authors.

Table 2
Results.

	Primary school		Middle school	
	(1)	(2)	(1)	(2)
Avg external ev.	−0.5789 (0.0139)**	−0.6097 (0.0140)**	−0.5352 (0.0087)**	−0.5662 (0.0118)**
Female		0.1562 (0.0034)**		0.3605 (0.0060)**
Immigrant		−0.0058 (0.0057)		0.1986 (0.0103)**
Parents M		0.0581 (0.0042)**		−0.0215 (0.0068)**
Parents H		0.1522 (0.0047)**		0.0616 (0.0079)**
Additional regressors	no	yes	no	yes
School fixed effects	yes	yes	yes	yes
N	127,082	127,082	73,899	73,899

Note. Dependent variable is difference between internal and external evaluations at the end of primary school and middle school respectively (“GPA”, i.e. average of scores in Mathematics, Spanish, Catalan, English). Sample for primary school spans from 2009 to 2013; sample for middle school spans from 2011 to 2013. Regressors shown in the table are average external GPA in the class (“avg external ev.”), and individual characteristics (gender, immigrant, control for parental education (high, middle or low) – parents L, i.e. low educated, is the baseline category). Regressions in columns (2) include also a dummy for observations without info on parental background, average characteristics of the class (share female, share immigrant, average parental education), and year dummies. All regressions include school fixed effects. Standard errors are clustered at the class level.

Table 3
Estimates by subject.

	Primary school				Middle school			
	Maths	Spanish	Catalan	English	Maths	Spanish	Catalan	English
Avg external ev.	−0.653 (0.012)**	−0.771 (0.012)**	−0.748 (0.013)**	−0.670 (0.011)**	−0.604 (0.014)**	−0.742 (0.013)**	−0.718 (0.014)**	−0.567 (0.012)**
Female	0.221 (0.004)**	0.161 (0.005)**	0.147 (0.005)**	0.077 (0.004)**	0.489 (0.007)**	0.245 (0.008)**	0.257 (0.007)**	0.191 (0.006)**
Immigrant	−0.010 (0.007)	0.029 (0.008)**	0.017 (0.008)*	−0.099 (0.007)**	0.163 (0.011)**	0.247 (0.014)**	0.309 (0.013)**	0.031 (0.010)**
Parents M	0.091 (0.006)**	0.052 (0.006)**	0.051 (0.005)**	0.086 (0.005)**	−0.004 (0.008)	−0.008 (0.009)	−0.010 (0.008)	−0.008 (0.007)
Parents H	0.213 (0.006)**	0.154 (0.006)**	0.148 (0.006)**	0.173 (0.006)**	0.074 (0.009)**	0.108 (0.010)**	0.089 (0.010)**	0.060 (0.008)**
N	127,082	127,082	127,082	127,082	73,899	73,899	73,899	73,899

Note. Dependent variable is difference between internal and external evaluation in the subject reported above the column; “avg external ev.” is the average at the class level in the same subject. Other regressors, school and year fixed effects are as in Table 2. Standard errors are clustered at the class level

respectively.[23] 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 0.15 higher than their external), while differences in middle school are smaller in size. There is no effect associated with being foreign born in primary school, while there is a positive premium in middle school.

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.[24]

Fig. 2 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 3.[25] Overall the qualitative results discussed in previous paragraphs are unchanged; all coefficients are significant at 1% and similar in size to the estimate for the GPA. However both for primary and middle school the estimated rate of grading on a curve is slightly higher for Catalan and Spanish, suggesting that the languages leave more room for subjective 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 highly consistent with the main specification, we will focus on GPA from now on.

5. Robustness checks

In this section we discuss threats to identification and propose robustness checks and an alternative specification to test the validity of the empirical strategy discussed in the previous section. In Section 5.1, we focus on issues related to sorting of students, while in Section 5.2 we focus on differences between internal and external evaluations in how they capture the underlying human capital.

[23] Lavy (2008) and Terrier (2016) also find a positive bias for female in non-blind test scores.

[24] 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.

[25] The limitations of studying subjects separately have already been discussed in Section 4.1.

5.1. Allocation of students and teachers to 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 most schools have 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 rules or in the demography of the region that may suggest changes 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.[26] 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 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 D 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 D – 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.[27] 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.[28] 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 paragraphs we will discuss how these potential issues may affect our estimates and perform robustness checks. Appendix E contains a more formal illustration of the challenges to identification.

5.1.1. Sorting of students across classes

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.[29] 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 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 is different because the dependent variable is the difference between two types of evaluations: if the model in Section 4.1 is correctly specified, the coefficients of the regressors measure the differential effect that the regressors have on internal versus external evaluations. The fact that regressors at the class level are correlated with individual human capital would not be problematic, precisely because human capital is not a determinant of the dependent variable. 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.[30] 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 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. $-\xi$) to be negative. In practice $-\hat{\xi}$ would also capture a spurious positive effect on internal evaluations of being with

[26] 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.

[27] We performed the same battery of tests described in Appendix D 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.

[28] 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.

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

[30] 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.

high performing peers. Thus the estimated coefficient may be smaller in magnitude than the true value, leading to an underestimation of ξ.

5.1.2. *Matching of teachers to classes*

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 $-\hat{\xi}$ 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.

5.1.3. *Proposed robustness checks and their results*

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 twofold strategy to ensure identification when working with middle school data. First, we replicate analysis for middle school using teacher fixed effects rather than school fixed effects. For middle school we can identify Maths and Spanish teachers for each class. 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 or reduce it substantially. For this robustness check we work with internal and external grades in Spanish and Mathematics. Results are shown in Table 4.[31]

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 distinguish the main teacher from substitutes. Thus some of the teachers may have spent only few days with the class, for instance while the main teacher was on sick leave, 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)). Results are qualitatively similar to the analysis by subjects discussed in the previous section (Table 3). Estimated coefficients are somewhat smaller in magnitude, although the difference is at most 0.08: while there might be some correlation between classes of high performers and strict teachers, we can surely rule out the concern that grading on the curve is only apparent and spuriously capture the assignment of teachers to class.

Table 4
With teachers' dummies.

	Middle school			
	Spanish (1)	Spanish (2)	Math (1)	Math (2)
Avg external ev.	−0.691 (0.014)**	−0.698 (0.016)**	−0.519 (0.015)**	−0.525 (0.017)**
N	72,056	54,627	72,044	51,259

Note. Dependent variable is difference between internal and external evaluation in either Spanish or Mathematics. 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. Standard errors are clustered at the class level.

Table 5
With school-level mean.

	Primary school	Middle school
Avg external ev.	−0.647 (0.014)**	−0.662 (0.040)**
N	122,951	70,195

Note. Dependent variable is difference between internal and external evaluations (GPA). Average external evaluations is the mean of students' evaluation at the school level in a given year (rather than at the class level). Other regressors and school fixed effects are as in Table 2 (average are computed at the school level). Standard errors are clustered at the school level.

In the second robustness check we replicate the analysis using school-level rather than class-level regressors. In other words, we broaden the definition of peers, including all the schoolmates enrolled in the same level, rather than just classmates, when computing average regressors. Average performance at the school level is obviously correlated with average performance at the class level, but it may be a less precise measure of the references group that teachers have in mind when grading children. Moreover this measure varies only over time (five years for primary school, three years for middle school) 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.

As shown in columns (2) of Table 5 results are quite close to the baseline model; if anything the coefficient for middle school is slightly larger in magnitude. Standard errors are larger for middle schools, but results are still significant at 1%.[32]

5.2. *Internal and external evaluations*

Even if the allocation of students across classes is as good as random for our purposes, any misalignment between internal and external

[31] 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.

[32] 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 5 is slightly smaller than in Table 2.

evaluations in capturing human capital that is not controlled for in specification (5) may be a source of concern.

In Appendix B we discusses additional robustness checks that provide evidence that our results are not driven by unobserved difference between internal and external evaluations across the distribution of human capital. More specifically, we replicate the analysis on the subsample of classes whose rank correlation between internal and external grades is high, we exclude outlier schools, we verify that results are not driven by particular subgroups of the population. Results always support our initial findings.

In Appendix C we develop an alternative specification which allow us to show that internal and external evaluations on average capture human capital in a similar manner. This alternative model relax the previous assumption that external and internal evaluations measure the same cognitive skills, and bring to the data the following specification:

$$\mathrm{int}_i = \gamma \mathrm{ext}_i - \xi \overline{\mathrm{ext}}_{c_i} + \sigma_{s_i} + \delta_F F_i + \mathbf{P}_i \delta_P + \delta_M M_i + \overline{\mathbf{X}}_{c_i} \beta + \tau_i + \varepsilon_i, \tag{6}$$

in which we estimate the coefficient γ rather than assuming that it takes value 1. As detailed in Appendix A.2, ext_i is correlated with the error term ε_i. Our identification relies on the use of A_i, student's age at enrollment in primary school, as instrument for external evaluation. In a further specification, we also instrument for $\overline{\mathrm{ext}}_i$, using as additional instrument $\overline{A}_i$, the average age at enrollment in primary school in the class. The instrumental variable approach is correct if A_i affects the human capital accumulation, but does not impact differently external and internal evaluations. In the appendix we describe in details our strategy and the results. The estimated coefficients using the alternative specifications are fully aligned to the results in Section 4.2, confirming the validity of our baseline approach.

6. 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 performance 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".[33]

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 (32.5%) 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.[34] 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, biases, and residual errors. More specifically, given the estimates reported in Table 2, columns (2), the individual internal evaluations can be rewritten as

$$\mathrm{int}_i = \widehat{\mathrm{int}}_i + \hat{\varepsilon}_i = \mathrm{ext}_i - \hat{\xi}\overline{\mathrm{ext}}_{c_i} + \hat{\sigma}_{s_i} + \hat{\delta}_F F_i + \mathbf{P}_i \hat{\delta}_P + \hat{\delta}_M M_i + \overline{\mathbf{X}}_{c_i} \hat{\beta} + \hat{\tau}_i + \hat{\varepsilon}_i \tag{7}$$

where $\widehat{\mathrm{int}}$ is the sum of ext and the predicted difference between internal and external evaluations, estimated following the model discussed in Section 4.1. The residual $\hat{\varepsilon}_i$ includes all the differences between internal and external evaluations that are not explained by our model. In particular it contains the difference between the random component of internal and external evaluations $\varepsilon_I - \varepsilon_E$, as it is clear from Eq. (3). 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 $\widehat{\mathrm{int}}_i$ to rank students allows us to get rid of differences between internal and external evaluations due to that randomness.[35] The selection of the top quartile of students performed using external evaluations ext_i and predicted internal $\widehat{\mathrm{int}}_i$ differs for 20.6% of the selected students in primary school and 22.25% of the selected students in middle school. Thus removing the unexplained residual closes about one third of the gap, while the remaining two thirds depends on bias and grading standards, as detailed in Eq. (7).

We can now "shut off" the various components of the RHS in Eq. (7), by simply subtracting the estimated coefficients multiplied by the variables of interest. For instance to get rid off the bias for female we sum $-\hat{\delta}_F F_i$. Then we 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 6 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 more than 7 p.p. (from 20.6% to 13.5%); eliminating school fixed effects reduces the gap by more than 5 p.p. (from 20.6% to 15.5%). Eliminating both of them simultaneously, namely removing the effect of grading standards from internal evaluations, decreases the gap by more than 13 p.p., explaining 63.8% of the differences between the selection with ext_i and the selection with $\widehat{\mathrm{int}}_i$.

The second part of Table 6 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 2.5 p.p., while removing the effect of parental education *increases* the gap by about 3 p.p.[36]

For middle school removing the effect of grading on a curve reduces the gap by about 7 p.p. (from 22.3% to 15.3%), 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 about 9.5

[33] 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.

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

[35] 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 $\widehat{\mathrm{int}}_i$ as a lower bound of the true gap if we could remove only the differences in the random errors ε_I and ε_E.

[36] 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 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.

Table 6
Selection of top quartile of students.

Primary school			
Selection based on:		Diff. with external	Improvement
predicted	$(\hat{int})$	20.62%	
w/o GOC	$(\hat{int} + \hat{\xi}\overline{ext})$	13.48%	34.63%
w/o school FE	$(\hat{int} - \hat{\sigma})$	15.49%	24.88%
w/o school FE & GOC	$(\hat{int} + \hat{\xi}\overline{ext} - \hat{\sigma})$	7.46%	63.84%
w/o female bias	$(\hat{int} - \hat{\delta}_F F)$	20.11%	2.49%
w/o parents bias	$(\hat{int} - P\hat{\delta}_P)$	21.24%	-3.01%
w/o all individual bias	$(\hat{int} - \hat{\delta}_F F - P\hat{\delta}_P - \hat{\delta}_M M)$	20.61%	0.06%
Middle school			
Selection based on:		**Diff. with external**	**Improvement**
predicted	$(\hat{int})$	22.25%	
w/o GOC	$(\hat{int} + \hat{\xi}\overline{ext})$	15.33%	31.11%
w/o school FE	$(\hat{int} - \hat{\sigma})$	17.99%	19.17%
w/o school FE & GOC	$(\hat{int} + \hat{\xi}\overline{ext} - \hat{\sigma})$	12.92%	41.96%
w/o female bias	$(\hat{int} - \hat{\delta}_F F)$	18.97%	14.74%
w/o parents bias	$(\hat{int} - P\hat{\delta}_P)$	22.69%	-1.97%
w/o all individual bias	$(\hat{int} - \hat{\delta}_F F - P\hat{\delta}_P - \hat{\delta}_M M)$	18.78%	15.62%

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.

p.p.; thus grading standards explains 42% of the differences between the selection with ext_i and the selection with $\hat{int}_i$.

Biases are slightly more relevant for middle school, explaining 15.2% of the gap – with bias for female having the larger effect. Overall 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.[37]

6.1. 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

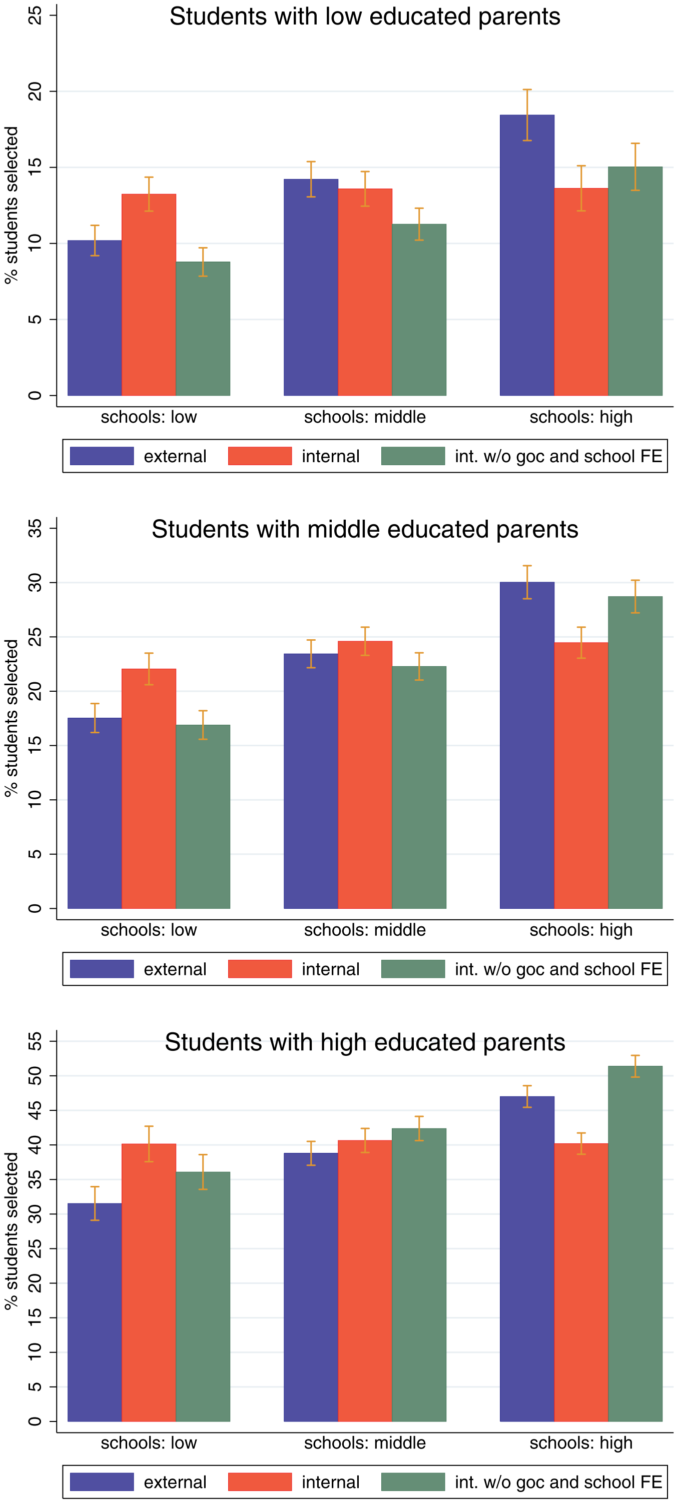

Fig. 3. Simulated selection of top quartile of students at the end of primary school. The graphs plot share of admitted students under different selection process in 2013. School quality is defined using school average outcomes in external evaluations from 2009 to 2012: low quality schools are in bottom 33% for at least two years, and never above 66 percentile; high quality schools belong to the best one third for at least two years, and they never belong to the bottom 33%. The top graph concerns students with low educated parents (both attended at most middle school), the bottom graph focuses on students with high educated parents (at least one with tertiary education and the other with high school).

[37] We replicated the simulation using the alternative specification described in Appendix C, with both student's and class' external evaluations instrumented. Results are qualitatively very similar. The total improvement of removing grading standards is slightly larger when the alternative specification is used (60.5% in primary school and 44.4% in middle school). In primary school, school fixed effects are more relevant than grading on the curve (removing them reduces the gap of about 9 and 5 p.p. respectively); this appears consistent with the lower estimated value for $\hat{\xi}$ using the instrumental variable approach.

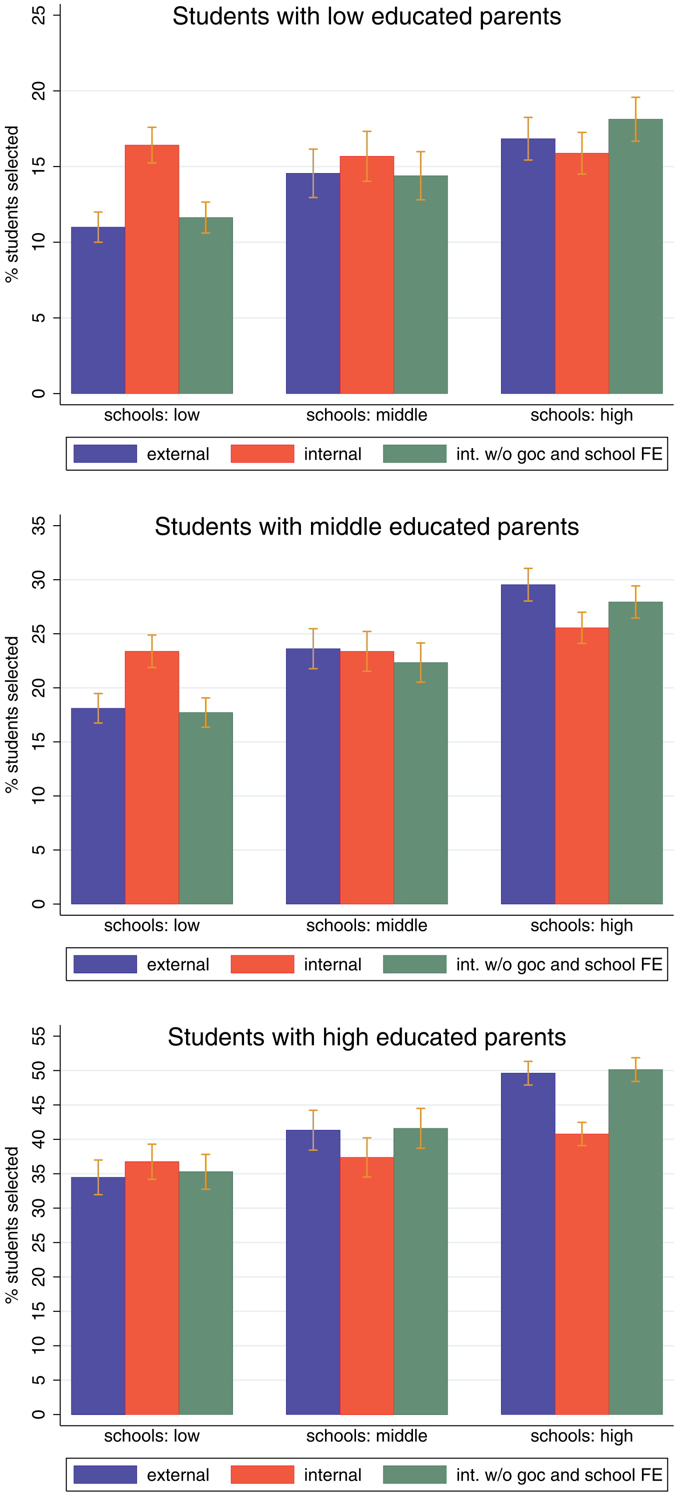

Fig. 4. Simulated selection of top quartile of students at the end of middle school. The graphs plot share of admitted students under different selection process in 2013. School quality is defined using school average outcomes in external evaluations from 2011 to 2012: low quality schools are in bottom 33% for at least one year, and never above 66 percentile; high quality schools belong to the best 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 attended at most middle school), the bottom graph focuses on students with high educated parents (at least one with tertiary education and the other with high school).

to attend "low performing" schools, that inflate internal evaluations more. In fact Figs. 3 and 4 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 performance 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 Figs. 3 and 4 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" by removing residuals and grading standards, as explained in the previous section. The evidence is similar 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 almost flat 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 schools.

Looking only at the aggregate statistics one 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.

7. 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. Therefore it is 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 major. This outcome may be suboptimal in terms of human capital accumulation, if they select lower quality schools and end up with less cognitive skills 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 seat. 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.[38]

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.[39]

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.[40]

Hence, these results should be taken as suggestive evidence that families understand that there may be grading on a curve and re-optimize their choice in a way that does not necessarily lead to optimal human capital accumulation, but to improved college assignment given that selection depends on internal grades.[41]

8. Conclusions

A large challenge when designing school admission procedures is not only to understand what the efficient assignment of resources is, but also to collect the relevant information about individuals to determine that assignment. The main problem is that skills are mostly unobservable. Children spend a large number of years in school, with teachers who spend a large number of hours with them. Hence teachers are the natural candidates to transmit information about their students, their learning capabilities and achievement. But teachers may have biases, may respond to stereotypes or have particular preferences which make them less reliable to be the only determinants of a measure of individual skills. That is why in a large number of countries access to tracks, colleges or resources in general is determined using an objective and comparable measure, that is, a nationwide exam. Such countries include South Korea, Brasil, China or India.[42] The problem though is that that these exams are very high stakes exams that occur in a small amount of time. Future prospects depeding on a realization of a one day exam does not seem ideal. In addition to the problems that measuring skills in one day exam may presents, there a are a number of papers describing how high pressure can lead to worse performance (Ariely, Gneezy, Loewenstein, & Mazar, 2009), how there are heterogenous effects in reaction to such large stakes by gender or race (Azmat, Calsamiglia, & Iriberri, 2016 or Attali, Neeman, & Schlosser, 2011), how the levels of pollution can affect the performance (Ebenstein, Lavy, & Roth, 2016). Hence, such large dependence on external exams does not seem efficient nor fair. Other countries have a combination of centralized exams and college-specific selection criteria, such as the US or the UK. Countries such as Spain, Israel, Australia, Denmark, Norway or Sweden have a mixed system, where internal and external grades mechanically produce a rank that determines college assignment.[43]

This paper provides empirical evidence on a novel source of distortion that may arise in any system that uses internal grades to compare students across schools. In particular it puts forth a novel form of grading bias that results when having better peers harms the grades assigned by teachers in school. Internal grades result from teachers following students and evaluating them in a more continuous basis, which seems to be a better evaluating procedure.

On the one hand distortions due to individual or peer characteristics seem unjustified, especially when they may affect the future allocation of students into further career paths. In countries like Israel access to college depends on both internal and external grades but deviations between internal and external evaluations are monitored and penalized at the school level. This may allow for the optimal use of information gathered by teachers, but avoiding systematic and unjustified biases. The literature in artificial intelligence has taken a different approach on the matter. College admissions could use all the data available about the student, such as place of birth, neighborhood of residence, school or family attributes and train an algorithm to predict college success. Papers such as Wightman (1998) or Bastedo and Bowman (2017) show some groups are unnecessarily discriminated when using grades to evaluate their entrance to college. In particular grades for certain school and certain subgroups of the populations underpredict performance in college. Mathioudakis, Castillo, Barnabò, and Celis (2019) propose training algorithms that will optimize both efficiency and fairness of the algorithm used for college admissions. Such algorithms would partially solve the problems we are pointing out if sufficient information about individuals could be used to train the college admissions algorithm.

On the other hand, grading on the curve indirectly induces effects similar to policies such as the Top Ten Percent Law in Texas, by imposing that rank within class counts towards access to college; or such as other Affirmative Action laws implemented worldwide, that favor children of disadvantages background when accessing further studies. The fact that the impact is achieved through a somehow indirect or unconscious channel can have the added benefit of limiting stigmatizing certain individuals. However, the fact that the impact depends on the school grading policies in general is not ideal if assignment to college, academic tracks or jobs are at stake.

What the optimal balance between internal and external

[38] 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.

[39] 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.

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

[41] A rigorous analysis on this would involve estimating preferences by parents under a school choice mechanism that does not provide incentives to tell the truth, which is beyond the scope of the present study. See Calsamiglia, Fu, and Güell (2019) for details on the mechanism used and the challenges that such estimation would entail.

[42] See portal.mec.gov.br for Brazil; en.moe.gov.cn for China; mhrd.gov.in/higher_education for India; www.kice.re.kr/sub/info.do?m = 0205&s = english for South Korea.

[43] See https://webgate.ec.europa.eu/fpfis/mwikis/eurydice for European countries; www.aqf.edu.au for Australia. Lavy (2008) describes school system in Israel.

examinations may be requires further understanding of the consequences that such distortions may have. For instance, are those accessing college because of the bias in internal grades benefiting from increased access, are they dropping out more often when in college? Do they benefit more from college than those who were kept out would have? More generally, what are the consequences of these distortions? Similar challenges are faced by the research studying the misallocation costs of Affirmative Action policies and shall be explored in future research.

Appendix A. More on the simple model for evaluations

A1. Correlation between $\overline{\text{ext}}$ *and* ε

In this section we depart from Eq. (4) in Section 2, assuming for simplicity that $\delta = 0$, i.e.

$$\text{int} - \text{ext} = -\xi\overline{\text{ext}} + \varepsilon = -\xi\overline{\text{ext}} + \varepsilon_I - \varepsilon_E + \xi\bar{\varepsilon}_E \tag{A.1}$$

and we study $\text{Cov}(\overline{\text{ext}}, \varepsilon)$. Given that the coefficient of $\overline{\text{ext}}$ (namely $-\xi$) is smaller or equal than 0, a negative bias leads to overestimating the magnitude of ξ, while a positive bias leads to underestimating the magnitude of ξ.

Let $(\varepsilon_{E,i})_{i=1,\dots,N}$ be the errors for N students in a class. To simplify the exposition, we pose

$$\text{Cov}(\varepsilon_{E,i}, \varepsilon_{E,j}) = \begin{cases} w & \text{if } i = j \\ c \geq 0 & \text{otherwise} \end{cases} \tag{A.2}$$

This formulation allows the error to have a class-level component that affect similarly all the students in the class. If the errors in the class are uncorrelated, $c = 0$. Thus, $\text{Var}(\bar{\varepsilon}_E) = \frac{1}{N^2}(Nw + 2(N-1)c)$.

Note that we can assume without loss of generality that ε_I and ε_E are independent, and thus $\text{Cov}(\overline{\text{ext}}, \varepsilon_I) = 0$. In fact if ε_I and ε_E share a common component, it would be netted out when taking their difference.

In the current framework, we define $\overline{\text{ext}}$ as the mean at the class level, rather than as a leave-out mean, i.e. the average of peer evaluations in the class. In fact it appears sensible to assume that teachers have a unique reference point (the "average performance") and compare each child with it, rather than changing reference point for every student. However, this introduces a further bias because $\overline{\text{ext}}$ is correlated with the individual error ε_E even when errors are independent across individuals, given that for each i ext_i is part of the mean.

$$\text{Cov}(\overline{\text{ext}}, \varepsilon) = \xi\text{Var}(\bar{\varepsilon}_E) - \text{Cov}(\bar{\varepsilon}_E, \varepsilon_E)= \tag{A.3}$$

$$= \xi\frac{1}{N^2}(Nw + N(N-1)c) - \frac{1}{N}w - \frac{N-1}{N}c \tag{A.4}$$

$$= -\frac{1-\xi}{N}w - \frac{(1-\xi)(N-1)}{N}c \tag{A.5}$$

Rewriting the covariance as in (A.4) highlights three sources of biases: (1). positive bias due to the correlation between $\overline{\text{ext}}$ and the average error $\bar{\varepsilon}_E$; (2). negative bias due to the fact that the average includes the individual observation; (3). negative bias due to a positive correlation among errors of the students.[44] Because of (2.), the total covariance is negative even if $c = 0$, as shown by (A.5). If $c > 0$, the total covariance could be negative even in the absence of the second bias (e.g. if leave-out means are used).

To have a sense of the magnitude of the bias we can consider the OLS estimator of an univariate regression (int − ext) on $\overline{\text{ext}}$:

$$-\hat{\xi} = \frac{\text{Cov}(\overline{\text{ext}}, \text{int} - \text{ext})}{\text{Var}(\overline{\text{ext}})} = -\xi + \xi\frac{\text{Var}(\bar{\varepsilon}_E)}{\text{Var}(\overline{\text{ext}})} - \frac{\frac{1}{N}w}{\text{Var}(\overline{\text{ext}})} - \frac{\frac{N-1}{N}c}{\text{Var}(\overline{\text{ext}})} \tag{A.6}$$

First, notice that biases due to 1. and 2. are small if external evaluations accurately measure the underlying human capital. In fact, $\frac{w}{N} \leq \text{Var}(\bar{\varepsilon}_E)$, and they are equal if $c = 0$, thus

$$\xi\frac{\text{Var}(\bar{\varepsilon}_E)}{\text{Var}(\overline{\text{ext}})} - \frac{\frac{1}{N}w}{\text{Var}(\overline{\text{ext}})} \geq -(1-\xi)\frac{\text{Var}(\bar{\varepsilon}_E)}{\text{Var}(\overline{\text{ext}})} \geq -\frac{\text{Var}(\bar{\varepsilon}_E)}{\text{Var}(\overline{H}) + \text{Var}(\bar{\varepsilon}_E)} \tag{A.7}$$

In other words, this bias is small if the "signal to noise" ratio is high, i.e. the errors are small perturbations and the average evaluations in the class is close to the average level of human capital. In practice, as detailed in Section 4, we perform all the empirical analysis using both average at the class level and among peers. The difference among the two specifications is negligible, providing evidence that the small bias due to the correlation of $\overline{\text{ext}}$ with the individual error is not a concern.

Second, if c is large the estimated $\hat{\xi}$ might be substantially larger than the true ξ. For concreteness, consider the case in which the entire error in the external evaluations is at the class-level, i.e. $\varepsilon_{E,i} = \varepsilon_{E,j} = \bar{\varepsilon}_E$ for all i, j, and thus $c = w$. Then, if $\text{Var}(\overline{H}) \sim \frac{\text{Var}(H)}{N}$

$$\frac{\frac{N-1}{N}c}{\text{Var}(\overline{\text{ext}})} \sim \frac{Nw}{Nw + \text{Var}(H)} \tag{A.8}$$

In other words, we are assuming that there is a shift of the distribution of external evaluations within a class, but teachers are able to unravel the common shock in their evaluations. For instance, this might happen if the test asks several questions on a topic which was not covered well in class: all the external evaluations in the class would be negatively affected ($\bar{\varepsilon}_E < 0$), and the expected difference between internal and external would be positive even if $\xi = 0$. Conversely, if the teachers accidentally went through an example that was exactly the same as an exam question, a positive

[44] More precisely, if leave-out mean are used, $\text{Cov}(\overline{\text{ext}}_{-i}, \varepsilon) = \xi\frac{1}{(N-1)}(w + (N-2)c) - c$

measurement error would affect all the external evaluations within the class, and $E(\text{int} - \text{ext}) < 0$.[45]

Instrumenting for $\overline{\text{ext}}$ would address this concern and also the issue 1. described above. In fact, by definition an instrument is correlated with $\overline{\text{ext}}$ but not with $\bar{\varepsilon}_E$, thus when a 2SLS approach is implemented, the estimate of ξ is not affected by the class level measurement error. In the next subsection we introduce a more general framework, and we propose an instrumental variable approach in Appendix C.

A2. A more general specification

The simple framework described in Section 2 is a special case of the following, more general, specification:

$$\text{ext} = \mu_{\text{ext}} H + \varepsilon_E \tag{A.9}$$

$$\text{int} = (1-\xi)\mu_{\text{int}} H + \xi\mu_{\text{int}}(H - \overline{H}) + \delta F + \varepsilon_I \tag{A.10}$$

Here we allow internal and external evaluations to differ in how they capture the human capital. Deriving H from Eq. (A.9) and replacing in (A.10), we obtain

$$\text{int} = \frac{\mu_{\text{int}}}{\mu_{\text{ext}}}\text{ext} + \delta F - \xi\frac{\mu_{\text{int}}}{\mu_{\text{ext}}}\overline{\text{ext}} + \varepsilon \tag{A.11}$$

This is equivalent to the model in Section 2 (up to "rescaling" the human capital) only if $\mu_{\text{int}} = \mu_{\text{ext}}$, namely if the initial assumption that internal and external evaluations capture human capital in the same way holds. If $\mu_{\text{int}} \neq \mu_{\text{ext}}$, it is not correct to use int − ext as dependent variable, and a different strategy should be implemented to estimate ξ and $\frac{\mu_{\text{int}}}{\mu_{\text{ext}}}$.

In Appendix C we discuss an alternative approach that allows us to estimate (A.10): we instrument for individual and average external evaluations in order to address the measurement error issues. This also allows us to directly verify that the coefficient of ext is about 1, namely that internal and external evaluations measure the same cognitive skills. Instrumenting for $\overline{\text{ext}}$ addresses also the issue of the regressor being correlated with the error term we discussed in the previous subsection.

Appendix B. Additional robustness checks

Even if the allocation of students across classes is as good as random for our purposes, any misalignment between internal and external evaluations in capturing human capital that is not controlled for in specification (5) may be a source of concern. Appendix C discusses an alternative specification which allow us to show that internal and external evaluations on average capture human capital in a similar fashion. The robustness checks on the baseline specification we discuss here provide evidence that our results are not driven by unobserved difference between internal and external evaluations across the distribution of human capital.

B1. Classes with high rank correlation

We replicate the analysis described in Section 4 on the sample of classes whose rank correlation between internal and external grades is high.[46] In our model 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: 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 teacher's biases for given individual characteristics and random errors may change the ranking. Moreover, and more problematic for our analysis, if internal evaluations take into account (non cognitive) skills that are not measured by external evaluations, the order might change. Hence, for the subsample of classes with high rank correlation, the two evaluations are truly aligned measure of human capital; if our results are spurious they should not remain when analysis are performed on such subsample.

Within class rank correlation among internal and external grades is generally large in primary school: the median value is 0.85 (mean is 0.83), 75% of classes have rank correlation higher than 0.79, and 25% are above 0.89. In middle school rank correlation is not as large as in primary school, but it is still sizable: median value is 0.69 (mean is 0.64), 75% of classes have a value higher than 0.55 and 25% are above 0.79.

In the empirical analysis in Table 7 we include only classes for which rank correlation of internal and external evaluations is higher than the median. Estimated coefficients are fully aligned with the baseline specification in columns (2) of Table 2.[47] This is a further confirmation that previous results are not driven by spurious correlation with unobserved variables.

B2. Schools without outliers

We exclude from the analysis schools in which, at least once, a class has particularly high (or particularly low) mean external evaluation; this robustness check acknowledges that ceiling effects may kick in at different points of the underlying distribution of human capital for internal and external evaluations. To clarify this point suppose that the test administered externally is "easy" so that most students above a given level of human capital achieve a top score, while teachers are able to discriminate among them when assigning internal evaluations. Suppose further that human capitals in the class are positively correlated (not necessarily because of sorting of students, positive peer effects on human capital may generate such correlation). Then high achievers would be more likely to be in a class with extremely high external evaluations and they would also be more likely to have a negative difference between internal and external evaluations. This spurious negative correlation might cause the estimated coefficient of mean external evaluations to be negative. Running the analysis only on "average quality" schools would mitigate or solve this issue: if the effect of

[45] We thanks an anonymous referee who suggested this example.

[46] 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.

[47] Setting alternative thresholds (e.g. rank correlation larger than the 60 percentile, or rank correlation larger than 0.75) deliver very similar results.

Table 7
Classes with high rank correlation.

	Primary school	Middle school
Avg external ev.	−0.616	−0.543
	(0.015)**	(0.020)**
N	63,564	36,960

Note. Dependent variable is difference between internal and external evaluations (GPA). This table shows results of the baseline specification performed on the subset of classes in which the rank correlation between external and internal evaluations is larger than the median (other regressors, school and year fixed effects are as in Table 2).

Table 8
Schools without outliers.

	Primary school			Middle school		
	(1)	(2)	(3)	(1)	(2)	(3)
Avg external ev.	−0.6175	−0.5829	−0.5882	−0.5746	−0.4647	−0.5688
	(0.0217)**	(0.0147)**	(0.0284)**	(0.0183)**	(0.0233)**	(0.0368)**
N	70,202	78,029	28,690	34,478	30,598	13,613

Note. Dependent variable is difference between internal and external evaluations (GPA). This table shows results of the baseline specification performed on subsamples of schools. In columns (1) we exclude all schools with at least one class in the top 15% of mean external evaluations in a given year. In columns (2) we exclude schools with at least one class in the bottom 15% in a given year. In columns (3) we exclude those with a class in one of the two tails. Other regressors, school and year fixed effects are as in Table 2.

grading on the curve is not real, it should not remain in this specification.

Table 8 show results of this robustness check. More specifically, we identify top and bottom 15% of classes for each year.[48] In columns (1) of Table 8 we exclude all schools with at least one class in the top 15% in a given year. In columns (2) we exclude schools with at least one class in the bottom 15%. In columns (3) we exclude those with a class in one of the two tails. Estimations on these subsamples confirm that previous results are not driven by classes who did particularly well or particularly badly in their external evaluations. Overall estimated coefficients are quite similar to baseline results in Table 2. They are somewhat smaller in magnitude when middle schools with classes in the bottom tail are removed (0.47 rather than 0.57), while removing top-performing classes do not affect the estimates neither in primary nor in middle school.

B3. Analysis by subgroups

As last robustness check we verify that results are not driven by particular subgroups of the population. We estimate the baseline model on subsamples of the population, such as males or females, and students with low, medium, or high predicted external evaluations (on the basis of their predetermined characteristics). Estimates are consistent across groups and always significant, confirming that the comparison with peers affect all types of students.[49]

Appendix C. Alternative specification and instrumental variable approach

Our baseline model relies on the assumption that external and internal evaluations measure the same cognitive skills. In this section, we bring to the data the more general formulation introduced in the previous Appendix A.2. More specifically, we estimate

$$\text{int}_i = \gamma \text{ext}_i - \xi \overline{\text{ext}}_{c_i} + \sigma_{s_i} + \delta_F F_i + \mathbf{P}_i \delta_P + \delta_M M_i + \overline{\mathbf{X}}_{c_i} \beta + \tau_i + \varepsilon_i, \tag{C.1}$$

using an instrumental variable approach. As already discussed in Section 2, ext_i is correlated with the error term ε_i. Our identification relies on the use of A_i, student *i*'s age at enrollment in primary school, as instrument for external evaluation. This approach is correct if A affects the human capital accumulation, but does not impact differently external and internal evaluations. In the next paragraphs we provide evidence in support of the validity of age at enrollment as instrument.

In $\overline{\text{ext}}_{c_i}$ several observations are averaged together, reducing measurement errors that come from idiosyncratic shocks. However, as discussed in Appendix A, $\overline{\text{ext}}_{c_i}$ might be correlated with class-level noise. Thus we will also estimate a two stage least square specification in which both A and its average $\overline{A}$ in the class are used as instruments for ext and $\overline{\text{ext}}$.

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. Therefore there is large 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 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.[50]

[48] We replicate the analysis with stricter or looser thresholds finding very similar results.

[49] The estimated effect is somewhat larger for boys and for students with low predicted internal, although differences are not large in magnitude. Results are available upon request.

[50] Fredriksson and Öckert (2014) for Sweden, (Puhani & Weber, 2007) for Germany, (Schneeweis & Zweimüller, 2014) for Austria, (Black, Devereux, & Salvanes, 2011) for Norway, (Crawford, Dearden, & Meghir, 2010) for England, (McEwan & Shapiro, 2008) for Chile, (Ponzo & Scoppa, 2014) for Italy, and (Elder & Lubotsky, 2009) for the US.

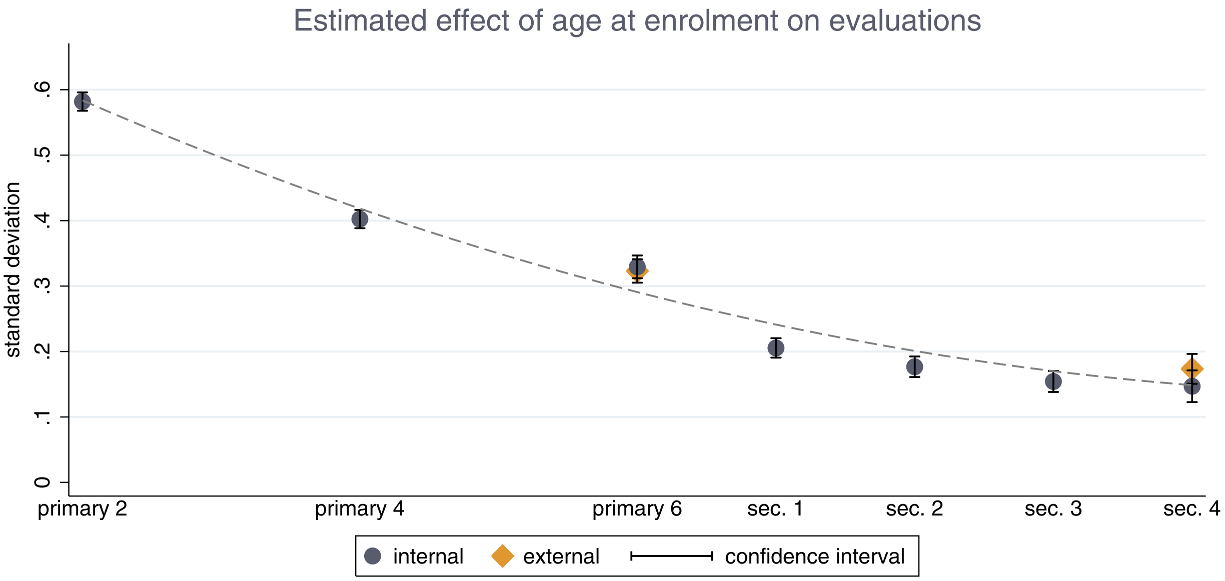

Fig. 5. Effect of age at enrollment over time. Each dot (diamond) is the estimated coefficient of a regression of internal (external) evaluations on age at enrollment and controls. The dotted line is a quadratic fit of all the estimates.

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.[51] We can verify using enrollment data for first grade of primary school that more than 99% of children are compliers.

Calsamiglia and Loviglio (2019), 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. Fig. 5 is based on a replication of 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 other individual characteristics, including year and school fixed effects; the figure plots the estimated beta coefficients.[52] 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, where it decreases at a slower pace.

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 closed at the end of lower secondary education.[53]

The finding summarized in Fig. 5 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 or later. In fact a potential concern is that if teachers are aware of the “disadvantage of being young”, they might want to correct for it when assigning the evaluation; in this case the effect of age on internal evaluations would be smaller than that on external evaluations. Conversely the estimated effect is extremely similar using internal or external evaluations: coefficients are virtually the same for primary school and the confidence intervals largely overlap for middle school.

We discuss now the plausibility of exclusion restriction for the average age in the class. It would be violated if the average age in the class has a direct effect on internal evaluations above and beyond its contribution to external evaluations. However the above results suggest that A does not have a different effect on internal and external evaluations, thus $\overline{A}$ should affect differently individual internal and external evaluations even if A does not. To be concrete suppose that students born in the first months of the years, having had an easy time at the beginning of their school career, grow up more enthusiastic about school. Teachers might not reward enthusiasm directly, but they might appreciate a class with more enthusiastic students because it makes a better work environment, thus they might inflate the grades of everyone in the class. In this case $\overline{A}$ would have a direct positive effect on internal evaluations and the estimated coefficient of $\overline{\text{ext}}$ would be smaller in size than the true one. For an example with the opposite bias, we may suppose that teachers do not inflate grades for younger students, but they somehow acknowledge a disadvantage of being in a class with younger students and inflate everyone’s evaluations, disregarding their month of birth. These illustrations sound quite unrealistic, but as a matter of fact we cannot completely exclude that $\overline{A}$ is correlated with some unobserved characteristic of the class that affect differently internal and

[51] 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.

[52] Sample for level 6 or primary school and level 4 of middle school is the same we used for the main analysis of this paper. In other levels we use all the students for which we could find internal evaluations in mathematics, Spanish, Catalan and English in one of the year in the time range from 2009/2010 to 2013/2014.

[53] 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.

Table 9
2SLS regressions.

	Primary school		Middle school	
	(1)	(2)	(1)	(2)
External ev.	1.0384	1.0296	1.1746	1.1217
	(0.0178)**	(0.0169)**	(0.0990)**	(0.0937)**
Avg external ev.	−0.6476	−0.3600	−0.7396	−0.5695
	(0.0223)**	(0.1820)*	(0.0986)**	(0.1141)**
Female	0.1509	0.1519	0.3570	0.3579
	(0.0041)**	(0.0041)**	(0.0066)**	(0.0064)**
Immigrant	0.0061	0.0053	0.2697	0.2490
	(0.0080)	(0.0079)	(0.0416)**	(0.0394)**
Parents M	0.0449	0.0460	−0.0559	−0.0482
	(0.0074)**	(0.0073)**	(0.0209)**	(0.0199)*
Parents H	0.1268	0.1314	−0.0187	0.0048
	(0.0126)**	(0.0121)**	(0.0461)	(0.0437)
N	127,082	127,082	73,899	73,899

Note. Dependent variable for all specifications is student's internal evaluations (GPA). In columns (1) student's external evaluations is instrumented with student's entry age; in columns (2) student's external evaluations and average external evaluations are instrumented with student's entry age and average entry age. Other regressors, school and year fixed effects are as in Table 2. First stage estimates are shown in Table 10. Standard errors are clustered at the class level.

Table 10
First stage estimates.

	Primary school			Middle school		
	(1) ext. ev.	(2) ext. ev.	(2) avg ext. ev.	(1) ext. ev.	(2) ext. ev.	(2) avg ext. ev.
Entry age	0.3238	0.3272	0.0002	0.1729	0.1068	0.0046
	(0.0086)**	(0.0087)**	(0.0001)	(0.0112)**	(0.0100)**	(0.0014)**
Avg entry age		−0.0830	0.2220		1.4972	1.7221
		(0.0544)	(0.0552)**		(0.1288)**	(0.1395)**
Female	0.1385	0.1385	0.0054	0.0524	0.0501	0.0348
	(0.0051)**	(0.0051)**	(0.0012)**	(0.0070)**	(0.0070)**	(0.0040)**
Immigrant	−0.3259	−0.3259	−0.0223	−0.5692	−0.5630	−0.1870
	(0.0096)**	(0.0096)**	(0.0024)**	(0.0134)**	(0.0133)**	(0.0094)**
Parents M	0.3258	0.3258	0.0177	0.2938	0.2882	0.1202
	(0.0064)**	(0.0064)**	(0.0016)**	(0.0084)**	(0.0083)**	(0.0055)**
Parents H	0.6416	0.6415	0.0308	0.6254	0.6192	0.2068
	(0.0071)**	(0.0071)**	(0.0019)**	(0.0104)**	(0.0103)**	(0.0073)**
N	127,082	127,082	127,082	73,899	73,899	73,899

Note. First stages of 2sls regressions shown in Table 9. "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. Standard errors are clustered at the class level.

external evaluations. We will first estimate (C.1) instrumenting only ext with A, and then instrumenting both ext and $\overline{\text{ext}}$ with A and $\overline{A}$. Finding results that are similar among them and aligned with the baseline specification in Section 4.2 would provide evidence that neither measurement errors nor class-level bias are a major concern in the current framework.[54]

Table 9 presents results of 2SLS regressions. First stage estimations are shown in Table 10. In columns (1) only individual external evaluations is instrumented with A, while in columns (2) both individual and average external evaluations are instrumented using A and $\overline{A}$.[55] As in the baseline specifications, we cluster standard errors at the class level. In all the specifications the estimated coefficient of ext is quite close to 1, although the difference is statistically significant for some of them.[56] Following Eq. (A.11) in Appendix A, ξ can be backed up from the empirical estimation dividing the coefficient of $\overline{\text{ext}}$ by the coefficient of ext. Thus the estimated rate of grading on a curve for middle school is 0.62 according to specification (1) and 0.51 according to specification (2). Both figures are quite close to the value of 0.57 found in our baseline specification. For primary school $\hat{\xi}$ is 0.62 (specification (1)) or 0.35 (specification (2)). The estimate in specification (1) is very similar to the value of 0.61 found in our baseline specification; the estimate of specification (2) is somewhat smaller in magnitude, however it also has a much larger confidence interval that contains the previous estimates.[57]

We replicated robustness checks described in Section 5 using specification (2), which is the most conservative of the two. All results are very

[54] In middle school $\overline{A}$ is most likely correlated with individual performances because some schools sort students across classes. The discussion in Section 5.1 and Appendix E apply to $\overline{A}$ as well.

[55] A is equivalent to the day of birth, rescaled in the interval [0,1], so that the value for a child born on January, 1 is 1, while it is 0 for a child born on December, 31.

[56] For middle schools p-values of a test that the coefficient is 1 are 0.08 and 0.19 for specifications in columns (1) and (2) respectively. P-values for primary school are 0.03 and 0.08.

[57] To compute formal test of the significance of the difference between the estimate of the baseline specification and each of the alternative specifications, we bootstrapped the statistics 1000 times (resampling classes, which are the cluster unit). We cannot reject the null hypothesis that the difference is 0 for specification (2) (both for primary and middle school). Specification (1) is more precisely estimated, and the null hypothesis is rejected at 5% (not at 1%), although the difference is small in size, especially for primary school.

consistent with the estimates in Table 9, although in few cases coefficients are less precisely estimated.

Overall results in this section confirm the robustness of the finding obtained with our main specification in Table 2, Section 4.2. We also replicated the analysis in Section 6 using the alternative approach described in this section, finding very similar results.[58]

Appendix D. 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 receives 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.[59] 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.[60]

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 11 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 of each variable, and then we decompose the total variance in these class averages into within school and between school variances.[61]

Table 11
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%	

Appendix E. 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 Eq. (5) as follows

$$\text{int} - \text{ext} = -\xi\overline{\text{ext}} + X\delta + \overline{\mathbf{X}}\beta + \varepsilon \tag{E.1}$$

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 $\overline{\mathbf{X}}$ 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 explained in Sections 2 and 4, the model in (E.1) relies on the assumptions that internal and external evaluations measure the same cognitive

[58] Results obtained with the alternative specification are available upon request. Footnote 37 briefly compares them with the finding described in the main text.

[59] 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.

[60] 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.

[61] The formula we use is

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

where x is the variable under analysis, $s = 1, \ldots, S$ is the school indicator, $c_s = 1, \ldots, C_s$ 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 $C_s = 6$.

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 two aspects in a simple way:

$$\text{int} - \text{ext} = -\xi\overline{\text{ext}} + X\delta + \overline{\mathbf{X}}\beta + \psi N + T_k + \eta \tag{E.2}$$

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 Eq. (E.1) $\varepsilon = \psi N + T_k + \eta$. If students are randomly allocated to classes, regressor are uncorrelated with ε and Eq. (E.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). Suppose that in class A the external evaluations average to e_A, while in class B their average is e_B, with $e_A > e_B$. Then $E(N|\overline{\text{ext}} = e_A) > E(N|\overline{\text{ext}} = e_B)$. $\text{cor}(\overline{\text{ext}}, \eta) \neq 0$. Abstracting from teacher effects, the correlation is surely positive, and would add a positive bias to the coefficient of $\overline{\text{ext}}$, decreasing its magnitude; thus $\hat{\xi}$ 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 $\text{cor}(\overline{\text{ext}}, T_k) < 0$, for instance because stricter teachers are assigned to high performing classes, then $\text{cor}(\overline{\text{ext}}, \eta)$ may be negative. Consequently the estimated coefficient of $\overline{\text{ext}}$ would be larger in magnitude than the real one, and $\hat{\xi}$ would overestimate the true effect of GOC.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.econedurev.2019.101916.

References

Ahn, T., Arcidiacono, P., Hopson, A., & Thomas, J. (2016). *Equilibrium grade inflation with implications for female interest in STEM majorsMimeo*.

Ammermueller, A., & Pischke, J. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics, 27*(3), 315–348.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies, 76*(2), 451.

Attali, Y., Neeman, Z., & Schlosser, A. (2011). *Rise to the challenge or not give a damn: Differential performance in high vs. low stakes testsIZA Discussion Papers 5693*. Institute for the Study of Labor (IZA).

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association, 14*(6), 1372–1400.

Azmat, G., & Iriberri, N. (2010). 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.

Bastedo, M. N., & Bowman, N. A. (2017). Improving admission of low-SES students at selective colleges: Results from an experimental simulation. *Educational Researcher, 46*(2), 67–77.

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

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review, 22*(4), 343–352.

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

Bobba, M., & Frisancho, V. (2014). *Learning about oneself: The effects of signaling academic ability on school choiceMimeo*.

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

Calsamiglia, C., Fu, C., & Güelll, M. (2019). Structural estimation of a model of school choices: The Boston mechanism vs. its alternatives. *Journal of Political Economy*. https://doi.org/10.1086/704573 forthcoming.

Calsamiglia, C., & Güell, M. (2018). Priorities in school choice: The case of the Boston mechanism in Barcelona. *Journal of Public Economics, 163*, 20–36.

Calsamiglia, C., & Loviglio, A. (2019). Maturity and school outcomes in an inflexible system: Evidence from Catalonia. *SERIEs - Journal of the Spanish Economic Association*. https://doi.org/10.1007/s13209-019-0196-6 forthcoming.

Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica, 81*(3), 855–882.

Crawford, C., Dearden, L., & Meghir, C. (2010). *When you are born matters: The impact of date of birth on educational outcomes in EnglandIFS Working Papers W10/06*. Institute for Fiscal Studies.

Cunha, F., Heckman, J. J., & Lochner, L. (2006). *Interpreting the evidence on life cycle skill formation. Vol. 1*, Elsevier697–812).

Diamond, R., & Persson, P. (2016). *The long-term consequences of teacher discretion in grading of high-stakes testsNBER Working Papers 22207*. National Bureau of Economic Research, Inc.

Ebenstein, A., Lavy, V., & Roth, S. (2016). The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution. *American Economic Journal: Applied Economics, 8*(4), 36–65.

Elder, T. E., & Lubotsky, D. H. (2009). Kindergarten entrance age and children's achievement: Impacts of state policies, family background, and peers. *Journal of Human Resources, 44*(3).

Elsner, B., & Isphording, I. E. (2017). A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics, 35*(3), 787–828.

Epple, D., & Romano, R. (2011). *Peer effects in education: A survey of the theory and evidence. Handbook of social economicsVol. 1. Handbook of social economics* 1053–1163.

Estevan, F., Gall, T., Legros, P., & Newman, A. F. (2014). *College admission and high school integrationWorking Papers, Department of Economics 2014 - 26*. University of Sao Paulo (FEA-USP).

Feld, J., & Zölitz, U. (2016). *Understanding peer effects - On the nature, estimation and channels of peer effectsResearch Memorandum 002*. Maastricht University, Graduate School of Business and Economics (GSBE).

Figlio, D., & Lucas, M. (2004). Do high grading standards affect student performance? *Journal of Public Economics, 88*(9–10), 1815–1834.

Fredriksson, P., & Öckert, B. (2014). Life cycle effects of age at school start. *Economic Journal, 124*(579), 977–1004.

Kinsler, J., Pavan, R., & DiSalvo, R. (2014). *Distorted beliefs and parental investment in childrenMimeo*.

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

Lavy, V., & Sand, E. (2015). *On the origins of gender human capital gaps: Short and long term consequences of teachers stereotypical biasesNBER Working Papers 20909*. National Bureau of Economic Research, Inc.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies, 60*(3), 531–542.

Mathioudakis, M., Castillo, C., Barnabò, G., & Celis, S. (2019). Affirmative action policies for top-k candidates selection, with an application to the design of policies for university admissions. *CoRR*. arXiv:1905.09947.

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

McEwan, P. J., & Shapiro, J. S. (2008). The benefits of delayed primary school enrollment: Discontinuity estimates using exact birth dates. *Journal of Human Resources, 43*(1).

Murphy, R., & Weinhardt, F. (2018). *Top of the class: The importance of ordinal rankNBER Working Papers 24958*. National Bureau of Economic Research, Inc.

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

Puhani, P., & Weber, A. (2007). Does the early bird catch the worm? *Empirical Economics, 32*(2), 359–386.

Rangvid, B. S. (2015). Systematic differences across evaluation schemes and educational choice. *Economics of Education Review, 48*, 41–55.

Sacerdote, B. (2011). *Peer effects in education: How might they work, how big are they and how much do we know thus far? Vol. 3*, Elsevier249–277).

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

Terrier, C. (2016). *Boys lag behind: How teachers' gender biases affect student achievementIZA Discussion Papers 10343*. Institute of Labor Economics (IZA).

Tincani, M. (2015). Heterogeneous peer effects and rank concerns: Theory and evidence. Mimeo.

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

Wightman, L. F. (1998). *LSAC national longitudinal bar passage studyLSAC Research Report Series*. Law School Admission Council.