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

DISCUSSION PAPER SERIES

IZA DP No. 13838

What Is at Stake without High-Stakes Exams? Students' Evaluation and Admission to College at the Time of COVID-19

Andreu Arenas
Caterina Calsamiglia
Annalisa Loviglio

NOVEMBER 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13838

What Is at Stake without High-Stakes Exams? Students' Evaluation and Admission to College at the Time of COVID-19

Andreu Arenas
Universitat de Barcelona and Institut d'Economia de Barcelona

Caterina Calsamiglia
ICREA, IPEG and IZA

Annalisa Loviglio
University of Bologna

NOVEMBER 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

What Is at Stake without High-Stakes Exams? Students' Evaluation and Admission to College at the Time of COVID-19*

The outbreak of COVID-19 in 2020 inhibited face-to-face education and constrained exam taking. In many countries worldwide, high-stakes exams happening at the end of the school year determine college admissions. This paper investigates the impact of using historical data of school and high-stakes exams results to train a model to predict high-stakes exams given the available data in the Spring. The most transparent and accurate model turns out to be a linear regression model with high school GPA as the main predictor. Further analysis of the predictions reflect how high-stakes exams relate to GPA in high school for different subgroups in the population. Predicted scores slightly advantage females and low SES individuals, who perform relatively worse in high-stakes exams than in high school. Our preferred model accounts for about 50% of the out-of- sample variation in the high-stakes exam. On average, the student rank using predicted scores differs from the actual rank by almost 17 percentiles. This suggests that either high-stakes exams capture individual skills that are not measured by high school grades or that high-stakes exams are a noisy measure of the same skill.

JEL Classification: I23, I24, I28

Keywords: performance prediction, high-stakes exams, college allocation, COVID-19

Corresponding author:
Caterina Calsamiglia
IPEG
Ramon Trias Fargas 25-27
08005 Barcelona
Spain

E-mail: caterina.calsamiglia@barcelona-ipeg.eu

* We gratefully acknowledge the IZA financial support under the IZA Coronavirus Emergency Research Thrust. We thank the staff at the Departament d'Ensenyament and IDESCAT for their help in processing the data.

1 Introduction

The COVID-19 outbreak has lead to unprecedented crisis all over the world. Because vaccines or widespread high-quality testing are not yet available, social distancing is the main tool to fight the spread of the virus, and hence, face-to-face classes and examinations have been interrupted. This is important especially in cases in which standardized test scores are used to allocate scarce resources, namely students to college, or to scholarships. What can policy-makers do in the absence of such information?

The main alternatives to be considered are either to postpone the exam, to use alternative methods to define admissions or to use a predictive model to assign grades on the high-stakes exam using historical data. If the exam is postponed it still may require some adjustments. In Spain, for instance, exams were postponed and in autonomous communities such as Catalonia, they adapted its format to ensure that all students had questions which they could answer given what they had covered in school.[1] In the UK, a predictive model was used for secondary school examinations. Likewise, the International Baccalaureate announced by the end of March that exams were cancelled and that a predictive model would be used to assign scores. The aim of this paper is to analyse the consequences of cancelling high-stakes exams in a given year and predicting them by using high school grades and information about the school and track attended. We outline the promises and pitfalls of different methods and empirically assess them.

There are two alternative policies that would be important to study: using a predictive model to replace a high-stakes exam, or running anyway the high-stakes exam after a period of prolonged school disruption. Ideally, we would like to compare the outcomes of each policy with the counterfactual outcomes in a normal year without the COVID-19 outbreak and the subsequent school disruption. This would allow us to empirically assess which policy ensures that test scores are closer to the counterfactual. This paper mainly focuses on studying the effect of the first policy - using a predictive model to replace a high-stakes exam. We use data from past cohorts of high school graduates in the Spanish autonomous community of Catalonia, leveraging information on both their high school grades and their test scores in the high-stakes

[1] Normally, students are given the option to pick between version A and B of the exam, but this year even for a specific version, individuals were allowed to pick four out of the six problems to answer.

exams to access college. We estimate various models using the first cohorts of students in the sample, and test how well they can predict the high-stakes exams of the last cohort of students. To the extent that the relationship between high school grades and high-stakes exams is stable over time, this exercise allows us to quantify the effects of using predicted scores to allocate students to college or other opportunities when it is not possible or desirable to run high-stakes exam. While we cannot directly test the second policy - running anyway the high-stakes exam after a period of prolonged school disruption-, in the last part of the paper we perform simulations to explore the consequences of shocks to performance in the high-stakes exams for college allocation.[2]

We estimate different predictive models for college admission scores, mainly based on high school grades and school characteristics. Our aim is to implement models that combine transparency and fairness with predictive power. This is important because although in principle we are not interested in estimating the parameters of the prediction model -we would just care about the out-of-sample predictive power-, models that make clear the way the information is processed will have a much broader appeal for policymakers. We then use information on individual level characteristics to understand how different prediction models may harm or favor different subgroups in the population.

In many countries, comprehensive standardized tests are used to select students along with measures of continuous assessment such as high school grades or GPA. In Catalonia, the allocation of students to majors and college is centralized and solely based on an admission score that combines high school GPA with high-stakes standardizes exams, the so-called *Selectivitat*.[3]

We collected detailed administrative data on high school and high-stakes exams performance, together with information on socio-demographic characteristics, for the universe of students who graduate from public high schools in Catalonia in 2010-2012. In particular, we use data on college applicants for 2010 and 2011, estimate a model to predict Selectivitat scores,

[2]Hopefully further access to data in the near future shall allow us to properly study this question. However, it is noteworthy that this task will not be easy, because most countries either did not run high-stakes exams or implemented a different exam, making the comparison with past outcomes challenging.

[3]Other examples of national exams include the Bagrut (Israel), the SAT (US), the Abitur (Germany), the Baccalaureat (France), the Maturità (Italy), or A-levels (UK). In Spain, college admission is fully regulated: colleges have no say in determining the final assignment. Hence, college admissions are similar to a school choice problem. Gale-Shapley Deferred Acceptance is used, where priority for each degree is defined by a weighted average of the Selectivitat results, giving higher weight to subjects that are more relevant for the particular degree.

and study what would have happened in 2012 if high-stakes exams had been cancelled and a prediction model had been used to input Selectivitat scores instead.

There are two main takeaways from our results. The first is that high-stakes exams are mainly explained by high school grades. The predictive accuracy of the model slightly improves when school effects are added, but the change is not large. Even adding gender or family background as predictors, which would not be politically implementable, does not substantially improve the predictions' accuracy. Results are similar using either linear models or machine learning techniques. Our preferred specification, which includes high school grades, track and school fixed effects, performs as well as linear regressions with a higher number of controls, or a random forest, which allows for a rich number of implicit interactions between variables. We also study how prediction accuracy differs across demographic groups. In our data, as in many other settings, female and low socio-economic status (SES) students tend to do worse in high-stakes exams than in high school grades. Given that gender or SES indicators are not included in the prediction models, using predicted scores for college admission may favor them. Adding school effects to the prediction reduces the prediction accuracy differences across students of different SES groups, but it does not affect the one for females.

The second important lesson is that the unexplained variation is very large: errors made in predicting correspond to half of the variance in high-stakes exams.[4] Given that colleges have capacity constraints and students' rank ultimately determine college admissions, we also study differences between students' rank as implied by the predicted scores and their actual rank. On average, the predicted and the observed rank differ by almost 17 percentiles.

Standardized exams are usually thought of as an objective measure of capabilities. Our results suggest either that high-stakes exams measure some aspect about individuals that high school grades and other school characteristics do not, or that they are a very noise measure of these same characteristics. For instance, different schools may have different grading standards and so it seems appropriate to somehow correct for grading standards when defining college admissions. We have run specifications where school effects, peer quality and a rich set of controls are included. But still, prediction errors remain very large. There is, to the best of our knowledge, no study that directly studies the predictive power of high-stakes exams

[4]Exams and grades are normalized to have mean 0 and standard deviation 1. The out of sample mean absolute error of our preferred specification is 0.55 s.d. (the root mean squared error id 0.7)

on future outcomes, given that a particular college is attained. High-stakes exams correlate with future outcomes because they greatly affect college admissions, but less is known about its relationship with future outcomes beyond, almost mechanically, college admissions.[5] This is important in order to understand the impact of using predicted scores on the efficiency of the resulting college assignment. On the basis of the available evidence, we can only conclude that the most accurate prediction model that is transparent is an OLS model with a small set of predictors, and it would deliver rather inaccurate predictions. Predictions would on average favor some subgroups, in particular females, who otherwise do relatively worse in high-stakes exams, compared to their performance in high school. Further research should aim at understanding whether this prediction error captures a latent ability that no other measure is capturing, or whether it captures irrelevant noise that distorts optimal matching in college admissions.

Section 2 includes the data description and summary statistics. Section 3 presents the main specifications used to train the prediction model and the in-sample and out-of-sample goodness of fit results. Section 4 discusses the goodness-of-fit analysis for different subgroups of the population. Section 5 presents the results of alternative prediction approaches, such as LASSO regressions and random forest analysis. Section 6 discusses simulations to quantify the change in ranking due to perturbation of the high-stakes exams. Section 7 concludes and discusses our results in light of the current debate in the context of the pandemic.

2 Background and Data

2.1 Institutional Setting

In Catalonia, as in the rest of Spain, students are allocated to colleges and majors through a centralized algorithm based solely on their entry score, which is a weighted average of high school GPA and a comprehensive exam at the end of high school, the so-called "Selectivitat".[6]

[5] Indirectly relating to this aspect is Ors, Palomino and Peyrache (2013), which suggests that gender differences observed in high-stakes exams do not correspond to performance in college for a set of applicants in an elite college in France. Similarly Estevan, Gall and Morin (2020) for Brazil, analyze the impact of affirmative action in college admissions. Affirmative action takes the form of requiring lower grades in the high-stakes exam. They do not find that students in the subgroup benefited by the policy reform perform worse once in college.

[6] In this paper we focus on students who enrol in University right after completing two years of academic upper secondary education ("Batxillerat"). A minority of students (20% of those enrolling for the first time)

The high school GPA, in turn, averages the evaluations for all subjects undertaken in the last two grades of secondary education. The teacher of a given subject assigns the evaluation taking into account students performance in several tests that she administers and grades during the year. Evaluations range from 1 to 10; the GPA is computed with up to two decimal digits, and a 5 or more is required to graduate.

All students that sit Selectivitat have to take five written tests: (1) Catalan, (2) Spanish, (3) Foreign Language (English, French, Italian or German), (4) History or Philosophy, and (5) one subject specific to the high school track they attended (Science, Humanities and Social Sciences, Arts).[7] Tests are graded on a 1-10 scale, with two decimal digits, and the final score is their average. Scoring at least 4 out of 10 is necessary to pass and be eligible to enrol in College. Students may undertake up to two additional field-specific exams to increase their entry score ("Selectivitat Específica"). The baseline entry score for student i is given by

$$\text{Entry Score}_i = 0.6\text{GPA}_i + 0.4\text{SEL}_i, \tag{1}$$

and can be improved by at most 4 points taking field-specific exams. In this paper we mainly focus on the relationship between GPA and SEL, which are common to everyone. We provide more details on the regulation for non-compulsory exams and implement analyses that include them in Appendix B.

College admissions here can be modelled as a one-sided matching market, where students are allocated to capacity-constrained programs following the Deferred Acceptance algorithm, where priorities for each of the colleges and majors are uniquely determined a major-specific weighted average of test scores. Following Azevedo and Leshno (2016) the equilibrium in this large market can be characterized by a set of thresholds for each program that define the applicant with the lowest entry score. These thresholds are publicly presented together with the final assignment.[8] While the overwhelming majority of students pass Selectivitat, having a

follow a different path and enrol at an older age, either after completing tertiary vocational education ("Grau Superior") or after spending some years in the labor market. Their entry score is computed in a slightly different way, while the allocation mechanism is the same.

[7]From 2017, only History can be chosen as (4), this change does not impact our analysis because the last year in our data is 2012.

[8]Students have to choose their major before enrolling, thus a program is a major in a given University. Before taking Selectivitat, they submit a list of preferences. Conditional on having a high enough entry score, students can enrol in any program.

high entry score is fundamental to have a shot at the most competitive programs.[9]

2.2 Data and sample selection

We use administrative data on students who graduate from a public high school in Catalonia between 2010 and 2012, and pass Selectivitat in the same year. We exploit two main data sources. The first one contains the universe on students applying to Catalan universities from 2010 to 2012, their entry scores, the assigned program, the high school they come from, and information on their background. In particular, we observe their gender, date of birth, postal code of their home address, nationality, paternal and maternal education and occupation category.[10] We work with the sub-sample of students who attended a public high school.[11]

The second data source includes the universe of students enrolled in a public high school from school year 2009/2010 to school year 2011/2012. We observe in which track they are enrolled (Sciences, Humanities and Social Sciences, or Arts), their end-of-the-year evaluations and the final high school GPA for those who graduate. Moreover, we observe their gender, date of birth, postal code and nationality. This data source was previously merged with census and local registers to retrieve information on paternal and maternal education.[12] We select the students who graduate between 2010 and 2012 and proceed to merge the two data sources.[13]

We find a match for 95% of the observations in the Selectivitat data source.[14] For 96.3% of them, there exists a unique match based on high school code, year of graduation, date of birth, gender, nationality, and postal code. An additional 0.7% of matching is retrieved exploiting also the variables about parental education. For the remaining 3%, we perform a fuzzy merge

[9] 95% of students passed Selectivitat in Catalonia in the years of our sample.

[10] Education and occupation of the parents are self reported when submitting the application. We group education levels in three categories: "low" for those who attained at most lower secondary education or vocational training, "average" for upper secondary education and two years of vocational tertiary education, "high" for college or higher qualifications. Occupation categories are "high skilled workers/managers", "employee in services", "employee in other sectors", "does not work or missing".

[11] About 60% of Catalan high school students taking Selectivitat attended a public high school, while 40% attended a private or semi-private high school. https://es.slideshare.net/coneixementcat/pau-dossier (in Catalan)

[12] See Calsamiglia and Loviglio (2019) for additional details.

[13] We keep only observations that are uniquely identified by high school code, year of graduation, date of birth, gender, nationality, postal code, mother education and father education. About 1.3% of observations are dropped in both data sources; they probably are same-sex twins enrolled in the same school.

[14] About 75% of graduates in our sample are matched with an observations in the Selectivitat data source. Usually, less than 80% of graduates undertake and pass Selectivitat. In fact only students who plan to enrol in a Spanish public University have a reason to take the exam.

that allows for discrepancy in postal code, nationality, and parental education.[15] When these variables are different, we keep the postal code in Selectivitat, the Spanish nationality, and the highest level of education.

We infer the Selectivitat score from entry score and high school GPA, using the formula in equation (1).[16] While we do not observe separately the results for the five exams in Selectivitat, we observe evaluations for the subject in school that are more likely to have affected performance in the exams. To mimic the first three exams, we retrieve evaluations in the second grade of high school for Catalan, Spanish, and the foreign language learned in school (English for 98% of students). For the fourth exam, we take the highest evaluations at the end of second grade between History and Philosophy. For the fifth exam, we take the average of the track-specific subjects taught in second grade.[17] When a students retake a subject twice because she repeats second grade, we keep the most recent evaluation. For a small subset of students we could not retrieve one or more of these variables. This issue affects less than 1% of students in 2011 and 2012, and 8% of them in 2010. In fact, in some cases retained students may be exempted from retaking subjects that they passed the previous year. For students in the first cohort of data, this means that we cannot observe the relevant evaluation. We standardize high school GPA, Selectivitat, and all other evaluations to have mean 0 and standard deviation 1 in a given year.[18]

For consistency, we run all the analyses on the subsample of students without missing evaluations. The final sample has 31079 observations; we use the 20425 from students who graduate in 2010 or 2011 for in-sample predictions and the 10654 observations from graduates in 2012 for out-of-sample predictions. As a robustness check, we replicate some of the analyses on the larger sample of 32154 observations.

[15] If the students moved after the enrollment in high school, the postal code in the enrollment data set may not be updated. Students may have double nationality and report only one of them, or acquiring the Spanish nationality when they turn 18. Information on parental education are self reported in the first data set while they come from administrative sources in the second one, thus some discrepancy is not unexpected. To be conservative, in the fuzzy merge we always require the education level of at least one parent to be the same.

[16] We drop 26 observations for which the retrieved score is smaller than 4 or larger than 10

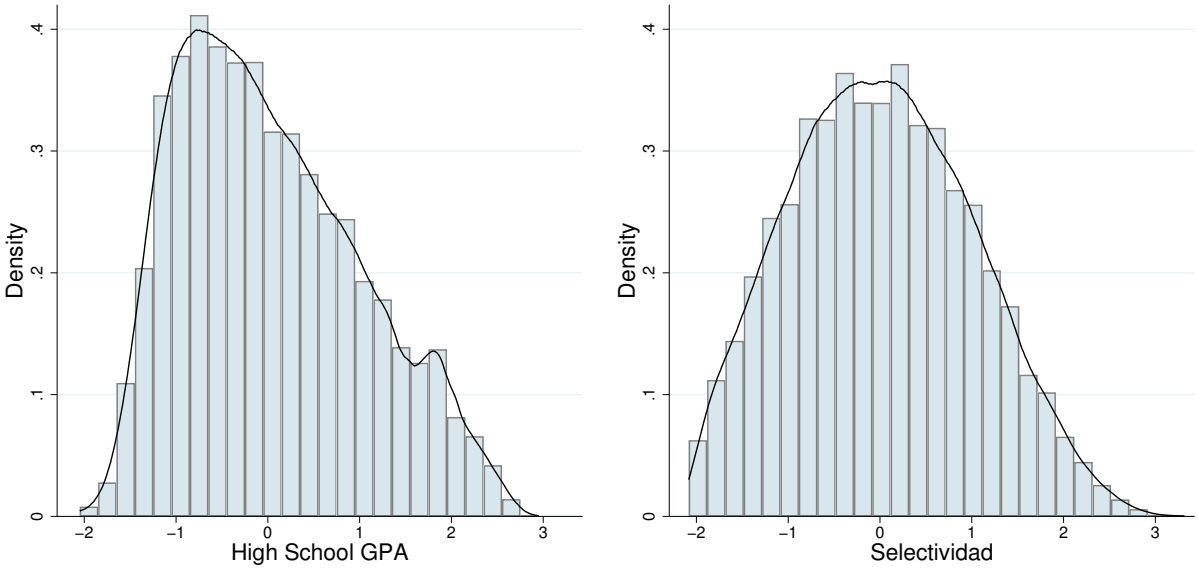
[17] Usually students undertake between 3 and 5 track-specific subjects. In the final sample, only 0.8% have less than 3 and 0.2% has more than 5.

[18] GPA and Selectivitat are standardized using the larger sample, to avoid deflating evaluations in 2010. The main purpose of the standardization is to offsets small differences over time in the difficulty of Selectivitat exams. Moreover, it makes the coefficients of the regressions easier to interpret. We replicate all the analyses using raw evaluations and find fully aligned results.

2.3 Descriptive Statistics

Figure 1 plots the distributions of high school GPA (left) and Selectivitat (right). Both have positive skewness, with a longer tail to the right, which is particularly thin for Selectivitat. In fact, very few students obtain evaluations close to the maximum in all five tests. Having a high GPA is relatively more common, and the density has a small pick at around +2 s.d. Moreover the distribution of GPA is somewhat more asymmetric, with a lot of mass between -1 s.d. and the mean, and a smaller left tail.[19] In fact, most students who barely graduate from high school choose to not pursue further academic studies and do not undertake Selectivitat.[20] While the truncation of the left tail is slightly more evident for Selectivitat, it is clear that few students score the bare minimum to pass. This is not surprising given that mostly only those who have graduated from high school and are planning on going to college take the test.

Figure 1: Distributions



Note. Distribution of evaluations for students who graduate from 2010 to 2012 and pass Selectivitat. The continuous black lines are kernel density plots.

Table 1 shows average GPA and Selectivitat by high school track, gender, and parental

[19]The local maximum in the right tail corresponds to a final GPA of 9 out of 10 in raw values. The interval $[-1, 0]$ corresponds approximately to raw scores from 6 to 7.

[20]In the sample used in this paper, only 31% of students whose high school GPA is lower than 5.5 undertake and pass Selectividad. For comparison, 61% of students with GPA between 5.5 and 6.5 do so. The share raises to 86% among those with a higher GPA.

education. For each category, share in the sample is reported in the last line. In the remaining of the paper we use parental education as main proxy for socioeconomic status (SES).[21]

About 46% of students in the sample attended the Humanities and Social Sciences track (Humanities from now on), about 50% attended the Sciences track, and the remaining 4%, the Arts track. The average GPA for students who did Sciences is almost 0.3 s.d. larger than for those in Humanities. They also do better in Selectivitat, but the gap decreases by about 0.1 s.d. Arts students have on average the lowest GPA, and the disadvantage is similar for Selectivitat. We include dummies for high school tracks in some of the specifications implemented in next Section, to account for systematic differences in the outcomes of students who come from different tracks.

Table 1: Descriptive Statistics by subgroups

	High School Track			Gender		Parental education		
	Arts	Sciences	Humanities	Male	Female	Low	Average	High
GPA	-0.233	0.184	-0.105	-0.092	0.123	-0.100	-0.039	0.218
	(0.025)	(0.008)	(0.008)	(0.009)	(0.007)	(0.011)	(0.009)	(0.010)
Selectividad	-0.263	0.121	-0.064	0.028	0.016	-0.227	-0.046	0.276
	(0.026)	(0.008)	(0.008)	(0.009)	(0.007)	(0.011)	(0.009)	(0.010)
% in sample	3.925	50.291	45.738	40.716	59.284	24.924	40.043	34.805

Note. The table shows average high school GPA and Selectivitat by high school track, gender, and parental education. Parental education is low if both parents have at most lower secondary education, high if one has tertiary education and the other at least upper secondary education, average in the other cases.

Females, who account for almost 60% of the students in the sample, on average have a GPA higher than males by 0.2 s.d. Conversely, they do not do better in Selectivitat (Azmat *et al.*, 2016). This rather striking difference may have several explanations. Females may do better in tasks or subjects that are captured by the GPA but are not tested by the high-stakes exams. Some teachers may have a positive bias for girls, or rewards non-cognitive traits (such as good behavior in class) that perhaps are more common among females (Lavy and Sand, 2018). Females may be more likely to choke on high-stakes tests and under perform. In this

[21]The three categories of parental education can be thought as a "mean" of father and mother education. Parental education is low if they both have low education, it is high if at least one is college educated and the other graduated high school, it is average in the other cases. When one of the parents is missing, the category of the other is used. Parental education is missing for 71 students in the sample. In Section 4, we also exploit additional characteristics, such as parental broad category of occupation.

paper we do not aim to provide any causal explanation about this difference in performance, but we will discuss how using predicted scores would affect students differently depending on their gender.

On average students with high educated parents perform substantially better than those with low educated parents. The gap is about 0.3 s.d. for high school GPA, and it widens to 0.5 s.d. for Selectivitat. Again, various reasons may be behind this difference. On one hand, students with low SES may be more likely to attend schools that set somewhat lower grading standards (e.g., for some grading on the curve effects, if performance are below the average in the population). In this case, their GPA would be somewhat inflated compared with their Selectivitat. On the other hand, they may be less likely to receive private tutoring and specific training for high-stakes exams. In this case high SES students may have an advantage in taking the test that does not reflect a higher level of skills. Finally, low SES students may also be vulnerable to high-stakes exams, for similar reasons that females do. As for gender, we will study how predictions based on high school evaluations would affect differently students with differing parental background.

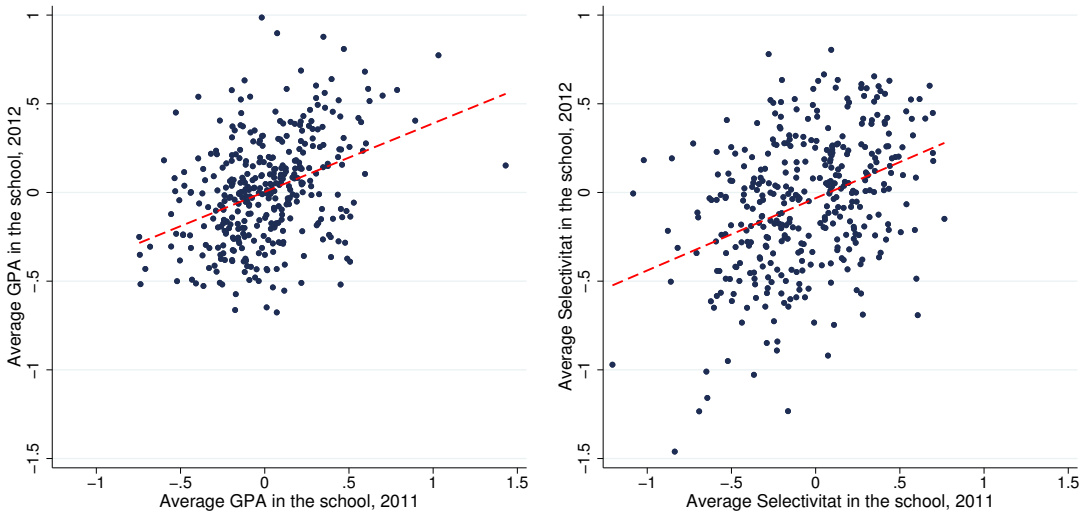
Public schools in Catalonia vary widely in size and share of students who take Selectivitat. In the models that require the estimation of school fixed effects, we include dummies only for schools with at least 6 students in each of the three years in the sample, and pool together the other schools in the baseline category. These schools are 369 out of 442, and they account for 95% of students in the sample.[22]

Figure 2 displays performance at the school level in two consecutive years. The left panel plots GPA in 2012 on the y-axis and GPA in 2011 on the x-axis. The right panel do the same for Selectivitat. In both years there is more dispersion in school average Selectivitat than in school average GPA. For both GPA and Selectivitat, the correlation between current and past average performance is positive and significant, but it is always below 0.5: there is clearly a lot of yearly variation that could not be anticipated looking at average performance in the previous year.

The correlation between average GPA and Selectivitat at the school/year level is about 0.4.

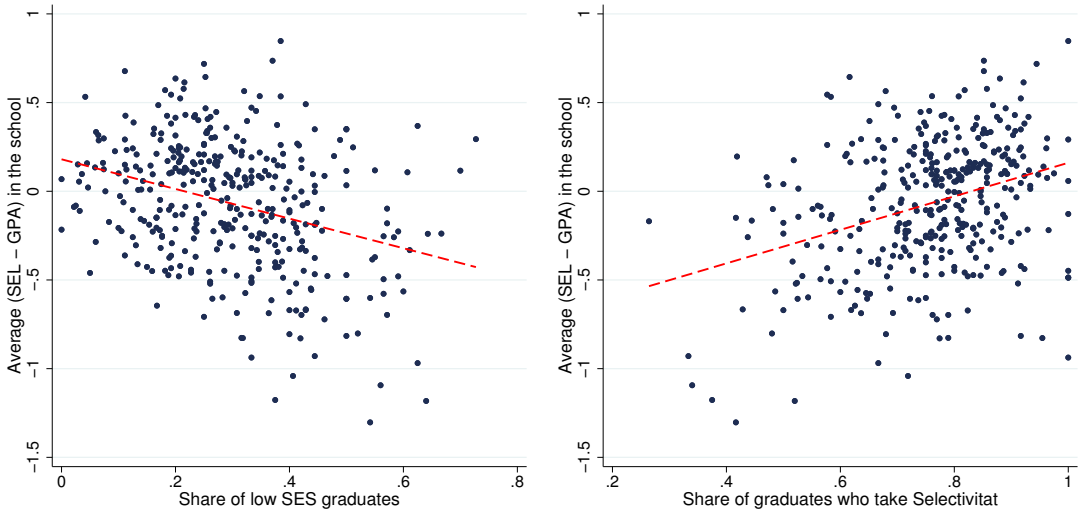
[22] As shown in Figure 9 in the Appendix, the number of graduates in a given year vary from 10 to more than 100, with most schools between 20 and 60. From 20% to 100% of them undertake and pass Selectivitat (the median is 75%). The correlation between size and share of Selectivitat takers is almost 0.

Figure 2: Average performance at the school level



Note. The graph plots avereage performance for schools with at least 6 students in the sample for each year. The dotted lines are a linear fit.

Figure 3: Average performance at the school level



Note. The graph plots data in 2012 for schools with at least 6 students in the sample for each year. The dotted lines are a linear fit.

As shown in Figure 3, in some schools students do much better in Selectivitat than in the internal grades, while in other schools the opposite happens. The left panel of the figure shows that schools with a higher share of low SES students tend to have worse outcomes for Selectivitat than for high school grades. This is not surprising given that, as discussed above, on average low SES perform relatively worse in Selectivitat than in high school. This descriptive evidence is also compatible with the grading on the curve documented for primary and lower secondary education in Catalonia by Calsamiglia and Loviglio (2019): teachers tend to inflate grades in classes where performance are below average, and to deflate them when average performance are above average. The right panel of the figure documents a positive correlation between the share of graduate who take (and pass) Selectivitat and the average gap between Selectivitat and high school GPA. It is also possible that schools in which many students aim to pursue tertiary education are more likely to provide specific training to maximize the score of the high-stakes exams.[23]

3 Empirical analysis

In this section we discuss how Selectivitat, the region-wide exam to access University, can be predicted using students' performance in high school. It is worth noting that even though covariates such as gender or socioeconomics may have an impact on the predicted grade, we do not use them in the estimation, as the objective is to use the estimated parameters for prediction. Considering covariates as gender or socioeconomics would imply that the policy used would treat different students differently because of characteristics that they should not be held responsible. In next Section 4 we test differences in the predictions across gender and SES. Moreover, in Section 5 we discuss whether a model that includes individual characteristics would deliver substantially better prediction, abstracting from any consideration about the desirability of such model. Hence, for now we study various linear models of Selectivitat as function of high school GPA, and school and track variables. Estimation is done by OLS, exploiting observations from students who graduated in 2010 and in 2011. The estimated

[23]The correlations in the figure are -0.36 (left) and 0.35 (right). Low SES students are slightly less likely to attend schools with a high share of college enrollment, but differences are small in size: the average low SES student who enroll in college attended a high school in which 76% of graduates took Selectivitat. This share slightly increases to 78% for the average high SES student.

coefficients are then used to perform in-sample predictions, and out-of-sample predictions for students who graduated in 2012.

3.1 Empirical specifications

Our analysis departs from the following empirical specification:

$$\text{SEL}_i = \alpha \text{GPA} + x_i'\beta + \epsilon_i, \tag{2}$$

$$\widehat{\text{SEL}}_i = \widehat{\alpha} \text{GPA} + x_i'\widehat{\beta} \tag{3}$$

where the dependent variable is student's i score in Selectivitat, and the right hand side includes high school GPA, a vector of other covariates x_i and the error term, ϵ_i.[24]

The specifications presented in Table 2 differ for the set of covariates x_i used. We start with a simple univariate regression of Selectivitat on high school GPA in column (1). For the purpose of students' allocation to majors and colleges, this model is fully equivalent to cancelling Selectivitat and relying only on high school GPA. In the following columns, we augment this specification with other covariates, to give more weight to subjects that are directly tested in Selectivitat, and to account for systematic differences due to track or high school attended. We add dummies for the track attended (from col(2)), peer GPA (col(2)), school fixed effects (cols (3) and (5)). Moreover, in column (4) and (5), we include high school evaluations for the subjects that are more relevant for Selectivitat (Catalan, Spanish, English, History or Philosophy, and the mean of track-specific subjects, as described in Section 2).

Equation (2) is estimated using the first two cohorts of students. Estimated parameters are used to predict SEL_i for all students. This approach relies on the assumption that the parameters in equation (2) are time-invariant, and therefore that previous cohorts can be used to infer the parameters and predict the outcomes of new cohorts. This is necessary for the prediction to be correct, namely to ensure that on average the predicted Selectivitat equals the observed one, but it is not enough to ensure that it is accurate, namely that the error in the prediction is reasonably small for all individuals. In fact, the larger the unexplained variation in the outcome (i.e. the larger the variance of the error ϵ_i in equation (2)), the less precise

[24] In the estimation, errors are clustered at the school level. The structure of the errors does not matter for the prediction, but it is relevant to comment on the significance of the coefficients.

the prediction is. Hence, for many students the predicted Selectivitat could be substantially different from the actual one. For each specification, we compute several measures of in-sample and out-of-sample goodness of fit to test the accuracy of the prediction: the coefficient of determination (R^2), the root mean square error (RMSE), the mean absolute error (AME), and the absolute rank deviation. While the first three are pretty standard, the last one deserve some additional explanations. The absolute rank deviation (ARD) gives us a measure of how different two ranks are. To compute it, we rank students by year using the observed Selectivitat, SEL_i, and the predicted, $\widehat{\text{SEL}}_i$. Ranks are rescaled to be between 0 and 1, thus the difference between ranks corresponds to the percentiles lost or gained by a given students if $\widehat{\text{SEL}}_i$ were used rather than SEL_i.

$$\text{abs rank dev} = \frac{1}{N}\sum_{i}^{N} |\text{rank}(\text{SEL})_i - \text{rank}(\widehat{\text{SEL}})_i| \tag{4}$$

We believe that this statistic is particularly relevant to evaluate the goodness of fit in the current setting. In fact, when high-stakes exams are used to allocate students to opportunities, a student's position in the rank can be more salient than the cardinal value of the test score. This is particularly true in our case, where student assignment to majors is fully determined by their rank and their stated preferences.[25] The absolute rank deviation of two independent variables is $\frac{1}{3}$, therefore in our analysis we expect this statistic to lie between 0 and 0.33.[26]

3.2 Results

As shown in column (1) of Table 2, more than 40% of the variance in Selectivitat is explained by high school GPA alone. On average, the absolute value of the difference between actual and

[25] As explained in Section 2, the final rank used to allocate students relies on an entry score that is a weighted average of Selectivitat and GPA. For the purpose of assessing the quality of our predictions, we believe that it is more relevant to look at Selectivitat alone, given that the entry score is mechanically correlated with GPA, the main regressor used to predict Selectivitat.

[26] Let $(a_i,\tilde{a}_i)$ be N realizations of two continuous random variables a and $\tilde{a}$, and r_i and $\tilde{r}_i$ their respective ranks. If $\text{Cov}(a,\tilde{a}) = 0$, then $E(|r_i - \tilde{r}_i|) = \frac{1}{3}$. In fact:

$$E(|r_i - \tilde{r}_i|) = \int_0^1 \int_0^1 |t-s|\,dt\,ds = \int_0^1 \left(\int_0^t t - s\,ds + \int_t^1 s - t\,ds \right) dt = \tag{5}$$

$$= \int_0^1 t^2 - t + \frac{1}{2}\,dt = \left[\frac{1}{3}t^2 - \frac{1}{2}t^2 + \frac{1}{2}t \right]_0^1 = \frac{1}{3} \tag{6}$$

predict Selectivitat is 0.6 s.d., and the absolute rank deviation is almost 19 percentiles. These statistics are remarkably similar for the in-sample and the out-of-sample prediction. On one hand, this confirms that the relationship between the high-stakes exam and the high school GPA is stable over time, therefore it is appropriate to use past cohorts to retrieve the relevant parameters. On the other hand, there is a lot left unexplained and therefore predictions are not very precise.

The remaining columns of the table show that adding covariates improves the out-of-sample predictive power, but not dramatically. Column (2) shows that being in the Scientific or Art tracks is associated with a small negative effects on Selectivitat (significant in some of the specifications), but do not improve the fit of the model. Peer GPA (on top of own GPA) is not correlated with the outcome.[27] As shown in column (3) adding school dummies has a sizable effects on the in-sample goodness of fit. In particular the R^2 improves from 0.44 to 0.51, the MAE decreases by almost 0.05 s.d. and the ARD decrease by about 1.6 percentiles. However, the improvements are lower for the out-of-sample predictions, for instance the R^2 goes from 0.42 to 0.45. Given that we use only two cohorts of data, school effects are probably capturing also short term differences across schools, thus their contribution to the out-of-sample prediction is positive but relatively low. We cannot rule out that exploiting a longer horizon to estimate the model would bring a larger improvement to the predictive power.

As displayed in column (4), high school subjects specific to Selectivitat have a large effect on the outcome and capture most of the effect of the overall GPA. Including these covariates raises goodness of fit both in-sample and out-of-sample in a similar way. Adding school effects, as in column (5), further improve the statistics. Figure 10 in the Appendix shows that the estimated school effects are sizeable (for instance the interquantile range is about 0.3 s.d.), but their correlation with the average difference between Selectivitat and GPA in the school is noticeably smaller in the out-of-sample year than in the previous year.

In this final specification, out-of-sample R^2 is 0.51, MAE is 0.55 s.d., and ARD is 16.7 percentiles. In other words, the richest model cannot account for almost half of the variation in the outcome. Consequently, prediction errors can be pretty large and this may have a sizable effect on ranks and therefore on students' allocation to majors.

[27] Peer GPA is the average GPA of students who graduate in the same school and year, including those who did not take Selectivitat. Given that the coefficient is virtually 0 we do not include it in further analyses.

Table 2: Selectivitat. Regressions and predictions

(a) In sample (cohorts 2010 and 2011)

	(1)	(2)	(3)	(4)	(5)
High School GPA	0.670**	0.669**	0.681**	0.174**	0.0995**
	(0.0099)	(0.0090)	(0.0111)	(0.0235)	(0.0192)
Art Track		-0.106*	-0.0794	-0.0959*	-0.0335
		(0.0495)	(0.0684)	(0.0488)	(0.0568)
Scientific Track		0.00536	-0.0109	-0.0396**	-0.0599**
		(0.0149)	(0.0152)	(0.0140)	(0.0140)
Peer average GPA		-0.00101			
		(0.0522)			
Catalan				0.0645**	0.108**
				(0.0150)	(0.0112)
Spanish				0.0632**	0.0858**
				(0.0154)	(0.0109)
History/Philosophy				0.0952**	0.118**
				(0.0132)	(0.0112)
Foreign Language				0.351**	0.354**
				(0.0109)	(0.0094)
Track subjects				0.0773**	0.0911**
				(0.0132)	(0.0106)
School FE	No	No	Yes	No	Yes
N	20425	20425	20425	20425	20425
R^2	0.442	0.442	0.518	0.507	0.585
RMSE	0.747	0.746	0.694	0.701	0.644
MAE	0.598	0.597	0.552	0.559	0.510
abs rank deviation	0.187	0.187	0.169	0.171	0.154

(b) Out of sample (cohort 2012)

	(1)	(2)	(3)	(4)	(5)
N	10654	10654	10654	10654	10654
R^2	0.424	0.425	0.448	0.493	0.512
RMSE	0.758	0.758	0.742	0.712	0.698
MAE	0.607	0.607	0.592	0.567	0.553
abs rank deviation	0.189	0.189	0.183	0.173	0.167

Note. The dependent variable is the score in the high-stakes exam Selectivitat. Selectivitat, high school GPA, and other evaluations used as regressors have been standardized at the year level to have mean 0 and standard deviation 1. The sample for panel (a) includes students who graduated from a public high school in Catalonia in 2010 and 2011. The sample for panel (b) includes students who graduated from a public high school in Catalonia in 2012. Errors are clustered at the school level. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

3.3 Robustness checks

As explained in Section 2, we drop from the main analysis students for whom we could not retrieve all the evaluations used in columns (4) and (5) of Table 2. As a robustness checks, we estimate the models in columns (1) to (3) on this larger sample. Columns (1) to (3) of Table 5 in Appendix A show that results are virtually unchanged.

Another potential concern is that most of the observations dropped belong to students who repeated the last year of high school twice and graduated in 2009. Thus to some extent the 2009 cohort is positively selected because it has less retained students than the following cohorts 2010 and 2011. To show that this does not affect the results, we replicate the analysis in columns (1), (4), and (5) using only the 2010 cohort to estimate the models. Again, Table 5 shows that results are quite similar to the finding in Table 2. Out-of-sample statistics are virtually identical for the models without school fixed effects, and only slightly worse for the model with school FE (e.g. the R^2 is 0.50 rather than 0.53, the MAE is 0.57 rather than 0.55). This is not surprising given that the estimation relies only one cohort of data, thus it cannot average out idiosyncratic shocks at the school/year level.

Moreover, we estimate models (1) and (5) on a subsample of "big" schools, namely the schools in which at least 20 students each year undertake Selectivitat (about 60% of the sample). The purpose of this exercise is to rule out that including school effects do not dramatically improve predictions because they are not precisely estimate for smaller schools. Results in column (7) and (8) of Table 5 show that the goodness of fit for this subsample is very similar to the baseline one.

Finally, we replicate the analysis using only the subsample of observations for which a perfect match was retrieved when merging high school and Selectivitat data (as described in Section 2.2). Column (9) replicates model (5) in Table 2. Results are almost identical, if anything the fit of the model is marginally better.

4 Analysis by subgroups

Any prediction of students' grades should aim at being both precise and fair, i.e., being equally precise across demographic groups. This does not imply that any differences across groups

between high school and Selectivitat test scores are fair, but simply that the prediction does not affect those differences.

Figure 4 displays the density of the prediction error ($\widehat{\text{SEL}}_i - \text{SEL}_i$) for the simplest prediction model (model (1), with only High School GPA), and the most saturated model (model (5), including school effects and other academic variables). The figure shows that the predictions are not very precise, and suggests that male test scores are relatively under-predicted. This is consistent with the descriptive evidence in Section 2.3, and it is aligned with previous findings in the literature that females tend to outperform males in lower stakes more often than in high-stakes exams.[28]

To quantify these differences, Table 3 reports the estimates of regressions of different measures of prediction error on a female dummy. In the first two columns, the dependent variable is the prediction error in test scores, which is the difference between predicted and actual scores: $\widehat{\text{SEL}}_i - \text{SEL}_i$. In columns 3 and 4, the dependent variable is the absolute value of this prediction error, which is a measure of its variability. In columns 5 and 6, it is the difference between the student rank implied by the predicted test score and the actual rank: $\text{Rank}(\widehat{\text{SEL}}_i) - \text{Rank}(\text{SEL}_i)$. In columns 7 and 8, the absolute value of this rank difference. Rank differences are important because the allocation of students to academic programmes with capacity constraints ultimately depends on their rank. For instance, a high prediction error at a part of the distribution with a low density of students may result in large differences in test scores prediction errors but in modest differences in rank prediction errors, and hence, be less consequential.

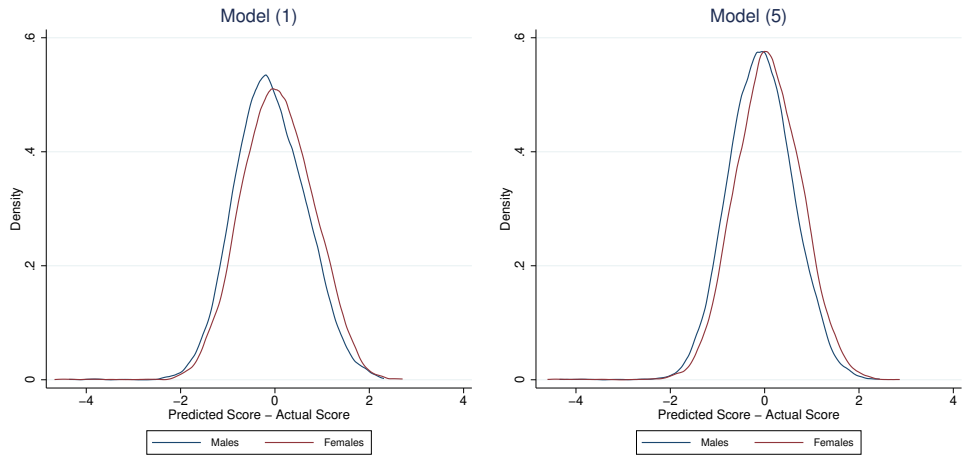
In Table 3, the intercept shows the average prediction error for males (the omitted category), and the female coefficient measures the gender difference in the prediction error. As suggested by Figure 4, the estimates show that male test scores are under-predicted by the model (.095 s.d.), while female test scores are over-predicted. The gender difference in the prediction error is around 0.156 standard deviations (s.d.), or between 6 and 7 percentiles.[29] These differences are large, because descriptive statistics in Table 1 show that the average gender difference in Selectivitat test scores is 0.012 s.d. In absolute value, however, there are no significant gender differences in the prediction error, which means that the dispersion of the prediction error is about the same for male and female students. Importantly, models (1) and (5) perform

[28] Azmat *et al.* (2016); Cai *et al.* (2019); Arenas and Calsamiglia (2019); Schlosser *et al.* (2019).

[29] Figure 11 in the the Appendix shows the distribution of the prediction error in terms of students' rank.

rather similarly across all prediction quality measures, which means that school effects and other academic variables are rather orthogonal to the gender differences in prediction quality. Despite the significant difference in prediction errors by gender, the R^2 of all the models in table 3 is very low, indicating that gender differences in prediction quality explain very little (2% at most) of the overall variation in prediction quality across students.

Figure 4: By gender



Note. Differences between predicted and actual rank, cohort 2012

Table 3: By gender

	Predicted - Actual Score		| Predicted - Actual Score |		Predicted - Actual Rank		| Predicted - Actual Rank |	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.156***	0.159***	0.00686	0.0102	0.0722***	0.0609***	-0.00295	0.000705
	(10.58)	(11.69)	(0.77)	(1.22)	(15.06)	(14.28)	(-0.95)	(0.25)
Constant	-0.0947***	-0.0797***	0.603***	0.548***	-0.0433***	-0.0367***	0.191***	0.167***
	(-8.47)	(-7.76)	(89.91)	(87.30)	(-11.78)	(-11.32)	(80.16)	(78.22)
Model	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)
R^2	0.0103	0.0126	0.0000554	0.000138	0.0209	0.0187	0.0000851	0.00000594
N	10654	10654	10654	10654	10654	10654	10654	10654

Header indicates the dependent variable. Omitted category: males. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In a similar vein, Figure 5 displays the density of the prediction error for the simplest and the most saturated prediction models, across parental Socio Economic Status (SES) categories, which are defined according to parental education. The figure suggests that high SES test scores are the most under-predicted, especially in the simplest model, while low SES test scores are over-predicted. Again, this is not surprising, given that, as discussed in Section 2.3, high SES students tend to outperform other students especially in higher stakes exams.

Table 4 reports the estimates of regressions of different measures of prediction error on a

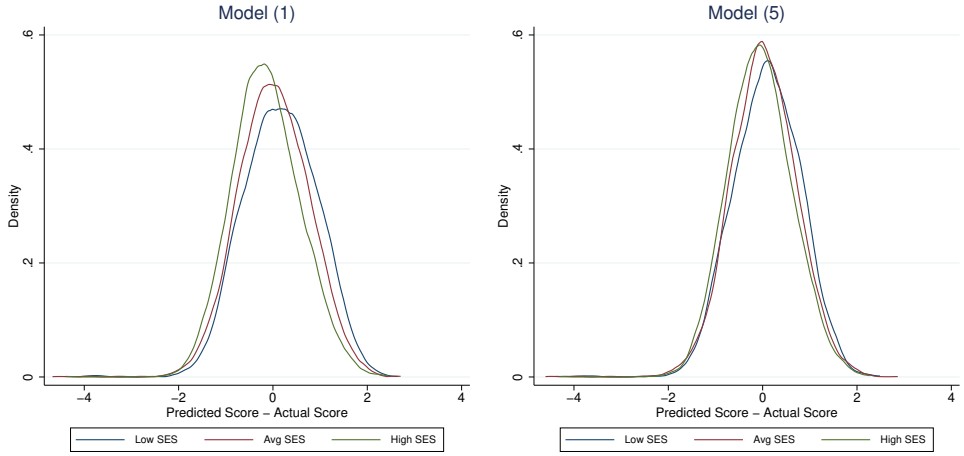
dummy for being high SES and a dummy for being average SES, the base category being low SES. As suggested by Figure 5, the estimates show that the test scores of high SES students are under-predicted by the model and those of low SES students are over-predicted, while the predicted test scores of average SES students are well predicted, on average. This is consistent with the descriptives in Table 1, which show that the magnitude of the high school differences in GPA across SES groups is amplified in Selectivitat test scores. The estimates in Table 4 show that when compared to the baseline category of low SES students, high SES students' scores are under-predicted by between 0.15 and 0.3 s.d. while average SES students' scores are under-predicted by between 0.7 and 0.14 s.d. on average. The magnitude of these estimates is large, considering that the average unconditional differences in Selectivitat test scores between high and low SES students are 0.5 s.d.; and between average and low SES students, around 0.2 s.d. At the same time, the results in columns 3 and 4, where the dependent variable is the absolute value of the prediction error, show that the prediction error for lower SES students has a higher dispersion.[30]

However, in contrast with gender differences, a more saturated model with school fixed effects (i.e., model (5)) considerably reduces the differences in prediction errors across groups, both on average and in dispersion, even if significant differences remain. This means that school effects and other academic covariates pick up some of the differences in prediction quality across SES categories, but very few gender differences. A similar pattern emerges in columns 5 to 8, which report regressions where the dependent variables are the prediction errors in terms of actual and predicted student ranks. Compared to the base category of low SES students, the rank of high SES students is underpredicted by between 6 and 1.5 percentiles, and the rank of average SES students is underpredicted by between 4 and 1 percentiles. Again, despite the significant differences in prediction errors across groups, the R^2 of all the models in Table 4 is very low, indicating that SES differences in prediction quality explain very little (2.4% at most) of the overall variation in prediction quality across students.

The low R^2 of the regressions of prediction errors on individual characteristics suggest that differences across gender and SES explain only a small part of the differences between high school grades and Selectivitat at the individual level. In next Section we confirm this intuition augmenting the prediction models with a vector of individual characteristics.

[30] Figure 12 in the the Appendix shows the distribution of the prediction error in terms of students' rank.

Figure 5: By SES



Note. Differences between predicted and actual rank, cohort 2012

Table 4: By SES

	Predicted - Actual Score		| Predicted - Actual Score |		Predicted - Actual Rank		| Predicted - Actual Rank |	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
High SES	-0.303***	-0.147***	-0.0471***	-0.0283**	-0.0600***	-0.0145**	-0.0267***	-0.0132***
	(-15.74)	(-8.25)	(-4.03)	(-2.61)	(-9.52)	(-2.58)	(-6.62)	(-3.66)
Avg SES	-0.144***	-0.0678***	-0.0389***	-0.0225*	-0.0398***	-0.0123*	-0.0108**	-0.00507
	(-7.54)	(-3.84)	(-3.35)	(-2.07)	(-6.28)	(-2.20)	(-2.71)	(-1.41)
Constant	0.162***	0.0928***	0.639***	0.573***	0.0362***	0.00902*	0.203***	0.174***
	(10.64)	(6.62)	(69.15)	(67.08)	(7.17)	(2.04)	(63.96)	(61.46)
Model	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)
R^2	0.0237	0.00660	0.00169	0.000679	0.00866	0.000695	0.00440	0.00133
N	10638	10638	10638	10638	10638	10638	10638	10638

Header indicates the dependent variable. Omitted category: low SES. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We report additional results in the Appendix, splitting the sample using other parental background covariates, with similar conclusions. Concretely, in Figure 14 and Table 15 in the Appendix the sample is split according to students' expected Selectivitat scores based on a more precise set of parental background variables (rather than the high-average-low SES classification based on education). These are estimated with a regression of Selectivitat scores on the full vector of dummies for mother and father high, average, low, or missing education, and for maternal and paternal occupational categories (services, manager, or unknown), for the prediction sample. The predicted performance is split in 3 quantiles, to make it easier to visualize. The prediction is done solely with parental background variables, which are not used in the prediction of grades, to avoid any mechanical correlation between the prediction error and the actual prediction. The results show, in a similar vein, that students that tend to do well based on parental variables are especially good at Selectivitat (compared to high school), and hence the prediction models underestimates their performance, although the variance of the prediction is lower for them.

5 Alternative prediction approaches

The baseline results are based on simple linear models. These models have the advantage of being transparent and easy to understand by policy-makers and the public. However, other more sophisticated prediction methods could outperform the linear model's predictions. In this section, we study the performance of two machine learning methods, namely Lasso Regressions and Random Forests.

Lasso (least absolute shrinkage and selection operator) regressions are a form of penalized regressions, typically estimated on a large number of candidate predictors, with a penalty for each non-zero coefficient. Lasso's $\hat{\beta}$ are the solution to: $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i \boldsymbol{\beta}'\right)^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\}$, where $\lambda > 0$ is the Lasso penalty parameter, and ω_j are the penalty loadings. Lasso regressions are estimated via cross-validation: slicing the sample into different parts, a training sample and a testing sample, the estimates maximize the predictive power of the training samples on the testing samples (Athey and Imbens, 2019).[31] In principle, we would expect LASSO to outperform OLS, because it keeps the

[31] In this case, the prediction sample is sliced into 10 folds, as suggested by Kuhn and Johnson (2013) or

same functional form, and reduces the prediction's noise by setting the coefficients of irrelevant predictors to zero.

Random forests are estimated by randomly drawing bootstrap samples and estimating regression trees over them, and then averaging out the results. A regression tree is estimated in a step-wise fashion. At each step, the algorithm splits the data into subsets, and among all the possible splits, the algorithm selects the one that minimizes the prediction error, jointly defined by choosing an independent variable and a splitting value. Whether the random forest model outperform OLS is a matter of relative functional form misspecification. If the relationship between high school GPA and Selectivitat is quite non-linear, we would expect random forests to improve on linear models, but not necessarily otherwise. See (Mullainathan and Spiess, 2017) and (Athey and Imbens, 2019) for further details.

Figure 6 displays the density of the prediction error from different methods, using the predictors of model (5) (school dummies and other academic variables). The results show that in this context, LASSO does slightly better but almost identically, and Random Forests do a little worse.

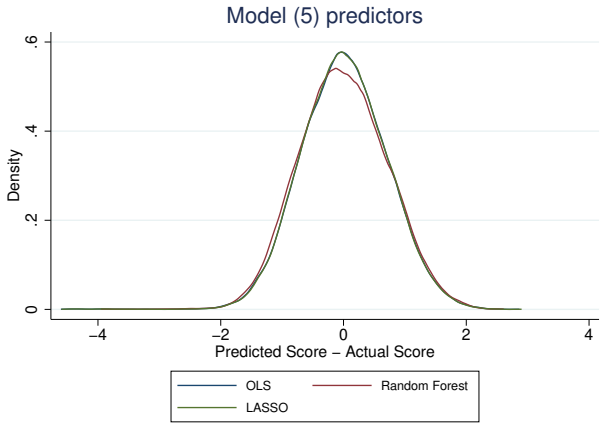
Finally, one may wonder whether including individual characteristics as predictors could improve the quality of the prediction. This is often seen as problematic for a variety of reasons which we do not aim at discussing in this paper, but it is nonetheless interesting to see how far individual characteristics go in improving test scores' predictions. To this aim, we replicate the analysis in Section 3 and in this section including also gender and socio-economic characteristics of the students. More specifically, Table 6 augments our preferred linear specification with gender (column (1)), parental background (column (2)), gender and parental background together (column (3)), postal code effects (column (4)), all of the above (column (5)). Adding gender and parental background characteristics slightly improve the out-of-sample predictions, but the difference is almost negligible. For instance the RMSE decreases from 0.7 to 0.6 and the absolute rank deviation decreases from 16.7 to 16.5 percentiles. Including the postal code fixed effects probably causes an over fit of the model in-sample, because the fit of the model out-of-sample is worse. The estimated coefficients for gender and parental education are significant and sizeable, but, as already stressed in Section 4, average differences across gender or SES explain only a small part of the overall variation between high school grades and Selectivitat.

Kohavi (1995), but other methods of selecting λ (3 folds, or the one-standard-error rule) give similar results.

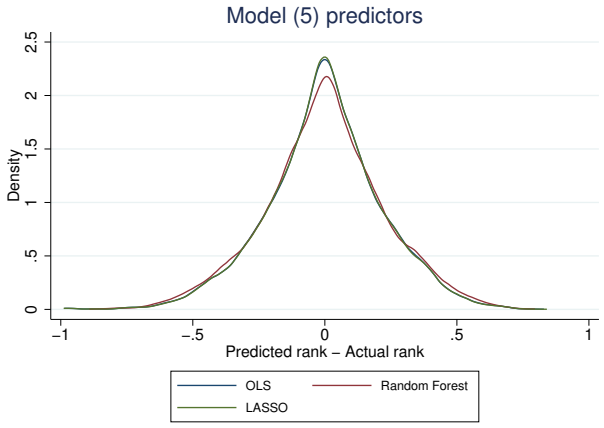
Most of it appears to be idiosyncratic. It is also worth noting the school effects estimated when socio-economic characteristics are added are extremely similar to the one in our preferred specification.[32] This suggests that at least part of the school effects measure across school variation in grading policy: if they were only capturing systematic variation in the demographics of the schools, they would shrink when socio-economic characteristics are added as covariates.

We also augment the LASSO and RF models discussed in this Section with individual characteristics. The density of the prediction error from the different errors are displayed in Figure 16 in the Appendix. This provides further confirmation that in this setting there appears to be no major gain from including individual pre-determined characteristics in the prediction model.

Figure 6: OLS vs. LASSO vs. RF



Note: Differences between predicted and actual score, cohort 2012.



Note: Differences between predicted and actual rank, cohort 2012.

6 Simulation: perturbed Selectivitat

Previous sections discuss to what extent predictions based on high school evaluations are a viable replacements for high-stakes exams to access college. A different but related question is how much school disruption would affect high-stakes exams and therefore students' allocation to college. Ideally, one would like to know the counterfactual score in the absence of disruption and estimate the expected loss (or gain) among individuals with given characteristics. Results

[32] The correlation of school effects from the baseline specification with the school effects from column (3) in Table 6 is 0.97. The two distributions are also very similar.

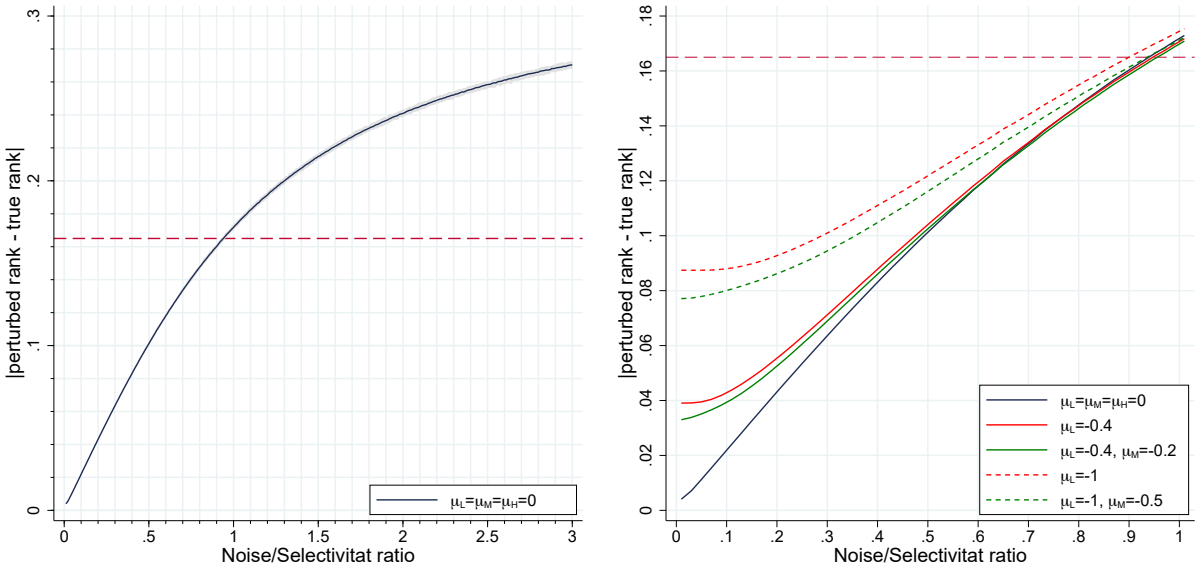
could be compared with the loss or gain that would take place if predicted scores were used instead. While a precise quantification of the effects of school disruption on students' allocation to colleges is beyond the scope of this paper, in this section we perform some simulation exercises to explore the consequences of perturbing Selectivitat on student rank.

Let $G_1, ..., G_g$ be a partition of the population into g subgroups (for instance a partition by low, average, and high SES). Let SEL_i be the true performance of student i who belongs to group G_j. The perturbed performance is

$$\widetilde{\text{SEL}}_i = \text{SEL}_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(\mu_{G_j}, \sigma_{G_j}) \tag{7}$$

where the error term is a random draw from a normal distribution with group-specific mean μ_{G_j} and standard deviation σ_{G_j}. In this section we denote r_i and $\tilde{r}_i$ the rank of student i computed using SEL_i and $\widetilde{\text{SEL}}_i$ respectively.[33]

Figure 7: Perturbed Selectivitat (simulation)



Note. The dashed red line plot the absolute rank deviation if predicted Selectivitat is used (model (5))

In the first exercise, we assume that $\epsilon_i \sim \mathcal{N}(0, \sigma)$, i.e. the noise generating process, is the same for everyone. We perform the same simulation for 300 values of σ, from 0.01 to 3: we

[33]Note that the perturbed score should be rescaled or truncated so that it belongs to the appropriate interval for Selectivitat. Focusing on the rank rather than on the cardinal values allow us to ignore this technicality.

draw random shocks for every individual in the sample, generate the perturbed Selectivitat, rank students, and finally compute the absolute difference between $\tilde{r}_i$ and r_i. This is replicated 100 times and results are averaged out to compute the absolute rank deviation for each σ. Results are shown in the left panel of Figure 7: the y-axis plots the ARD and the x-axis plots the noise to Selectivitat ratio. Given that Selectivitat is normalized to have mean 0 and s.d. 1, the noise to Selectivitat ratio corresponds to σ. Not surprisingly, the ARD is increasing in σ.[34] For comparison, the ARD for our preferred predictive model using high school evaluations is 16.7 percentiles. To obtain the same value, σ should be about 0.95. In other words, the variance of the noise should be almost as large as the variance of the true measure of performance to produce the same ARD as the prediction based on internal evaluations.

In the right panel of Figure 7, we replicate the analysis allowing μ_G to vary by parental education: more specifically, we assume that low SES are hit more negatively than high SES, and average SES are somewhere in between, i.e. $\mu_L \leq \mu_M \leq \mu_H$. In fact, not all students are affected in the same way by a prolonged period of school closure: those who do not have at home the necessary resources to access distance-learning or study on their own are for sure more severely hit. Kuhfeld *et al.* (2020) revise the literature on school absenteeism: averaging the finding of several papers they suggest that missing a school day decreases test scores by almost 0.007 s.d.[35] During the COVID-19 outbreak, schools were closed for almost 3 months; taking at face value the estimate in Kuhfeld *et al.* (2020), this would correspond to a loss of about 0.4 s.d. for those who did not have access to education in that period.[36] In one of the simulation in Figure 7, we set $\mu_L = -0.4$, while μ_M and μ_H stay at 0 (continuous red line). For small values of σ, the ARD is sizeably larger than in the baseline case (it is 4 percentiles when σ is close to 0); the gap decreases with σ and the two lines in the graph eventually overlap. In another simulation, we assume that also students with average SES are negatively hit, although less

[34] A graph with a longer x-axis would show that it slowly converges to 33 p.p. ARD is above 30 p.p. for $\sigma > 10$.

[35] Maldonado and De Witte (2020) is the only study in our knowledge that analyze the drop in test scores after the COVID-19 outbreak, using data from Flemish schools in Belgium. They find that after 7 weeks of school closures, test scores decreases by 0.19-0.29 s.d. These estimates are remarkably similar to the figures suggested by the literature on school absenteeism. Both this study and Kuhfeld *et al.* (2020) consider students in primary or lower secondary education, and focus only on Mathematics and reading or literature. Admittedly, their external validity for students near the end of upper secondary education is unclear. Given the illustrative purpose of our exercises, we believe that their estimates provide a nice benchmark nonetheless.

[36] Schools were closed from March 13, 2020 to June 19, 2020; assuming a school week of 5 days and not counting holidays, this corresponds to 62 days.

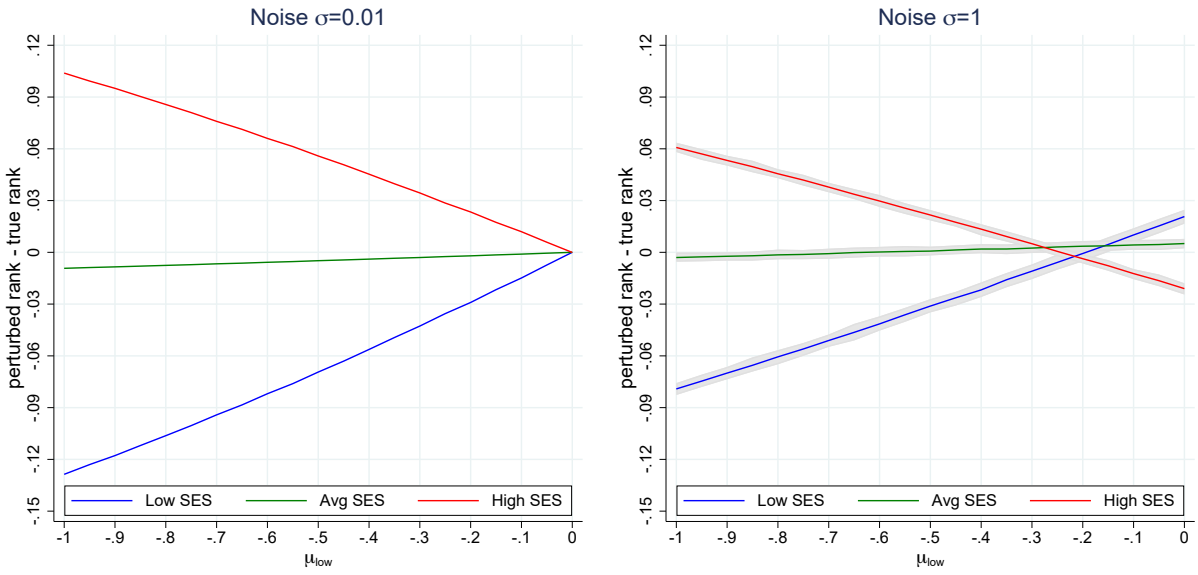
than those with low SES, and set $\mu_M = -0.2$ (continuous green line in the figure). For all values of σ, the ARD is slightly smaller than the one in the previous simulation: if more students are negatively affected by school disruption, overall performance decrease but the ranking changes relatively less, because the gap between low and average SES is reduced. Figure 7 also shows the outcomes of simulations that set much larger penalties for low and average SES ($\mu_L = -1$ and $\mu_M = -0.5$, dotted red and green lines in the figure). While for small values of σ the ARD is much larger than in the baseline exercise, it is still lower than the ARD of our preferred predictive model for $\sigma < 0.9$.

Low performing students are generally favored by a larger noise. At the extremes, perturbing the score can only improve the rank of the worst student, while it can only worsen the rank of the best student. More in general, as previously shown in Figure 1, the density of Selectivitat is much higher around the mean than in the tails. If someone in the left tail is hit by a positive shock, he typically gains more positions in the rank that what he would lose if hit by a negative shock of the same size. Vice-versa, a student in the right tail gains fewer positions if she receives a positive shock than what she loses for a negative shock of the same size. This intuition is confirmed by panel (a) of Table 7 in the Appendix, which displays the average difference $\tilde{r}_i - r_i$ by quantile of Selectivitat, for various values of σ, when $\mu_G = 0$ for all groups G. On average, for students in the bottom quartile the rank $\tilde{r}_i$ is higher than the baseline rank r_i, and the gain is increasing in σ. Conversely, on average $\tilde{r}_i$ is lower than r_i for those in the top quartile. Given that students with low SES are over-represented in the left tail of the distribution, and those with high SES are over-represented in the right tail, on average for low SES $\tilde{r}_i$ is larger than r_i on average. As shown in panel (b), the gains for low SES are sizable only for large values of σ. Therefore it seems likely that they would be completely offset by a decrease in μ_L, which would negatively affect all low SES, while an increase in noise would only benefit the low performers among them.

We explore the effects of varying the mean shock in the simulations plotted in Figure 8. μ_L ranges from -1 s.d. to 0, and $\mu_M = 0.5\mu_L$. For simplicity, we set $\mu_H = 0$. In the left panel, the variance of the shock is negligible ($\sigma^2 = 0.0001$), while in the right panel it is as large as the variance of performance ($\sigma^2 = 1$). The rank of students with low parental background decreases linearly with $-\mu_L$, while the rank of students with high parental background increases linearly. Almost nothing changes for those with average parental background: in fact, they do

relatively worse than the high SES due to the negative shock, but they do relatively better than the low SES who are hit by a larger shock. When σ is small, shocks have a sizable effects on the expected rank. For instance, for $\mu_L = -0.4$, the rank for low SES decreases by 5.6 percentiles, while the rank for high SES increases by 4.5 percentiles. For $\mu_L = -1$, the rank decreases by 12.9 and increases by 10.4 respectively. Increasing the variance of the shock reduces the gap, and reverses it if $|\mu_L|$ is small relatively to σ: this happens for $|\mu_L| < 0.2$ in the right panel of the figure. In fact, as we discussed above, adding noise favors students in the left tail of the distribution of performance, who are over represented among students with low SES.

Figure 8: Perturbed Selectivitat by SES



Note. The gray bands are 95% confidence intervals for the simulation.

7 Discussion and conclusions

College admissions match students to colleges using a combination of high-stakes exams, high school grades, and potentially other characteristics. The COVID-19 outbreak has forced social distancing worldwide, and this has made it harder to run large scale exams and hence to decide on college admissions. Some countries have postponed admission exams, and others have used historical data to predict students' performance in the high-stakes exams which would have taken place in absence of the pandemic. In this paper, we analyse how using the optimal

predictive model would affect college admissions. That is, we use historical data to train a model and evaluate its accuracy in a year where the high-stakes exam was run normally. We find that the optimal prediction is rather simple: a linear model with high school grades and school characteristics does similarly than more complex and less transparent prediction models. However, predictions are rather inaccurate in general, and more so for particular subgroups. Hence, if one aims at constructing a policy that perfectly substitutes high-stakes exams, using prediction models is questionable.

This paper studies the effect of using a prediction model a relatively short time after an unexpected shock occurs. More specifically, it is reasonable to assume that in 2020 high school GPA was not affected by COVID-19: in March most of the grades have already been assigned, the majority of the material is covered, and mostly the preparation for the high-stakes exams is relevant. Therefore, differences in access to online teaching, tutoring, and other educational resources when schools are closed would mainly distort the outcomes of high-stakes exams. The models presented here would require some adjustments if the outbreak continues to be inhibiting face-to-face exam taking next year, as the GPA would be itself distorted by the pandemic. In fact, access to online teaching and educational resources has been extremely unequal: many students with low SES faces more severe challenges their high SES peers, and may perform worse than in normal times.[37] Understanding these effects is not only crucial for eventual future predictions of high-stakes exams, but also for evaluating the consequences of an alternative policy of running high-stakes exams after months of school closure.

In what remains, we will go over some of the solutions adopted by different countries and discuss the problems that they seem to have generated. In Catalonia, the authorities announced that the exam was easier, giving students the option to choose a subset of questions to answer. This ensured that not having covered particular material in school would not harm students too much. But it also meant that the exam was less difficult than in previous years. Hence we expect student scores to exhibit more of a bimodal distribution: for those more heavily hit by the outbreak, the exam was most likely still inaccessible given that they had been locked at home for four months. For those with a relative stability around them and access to online

[37] A large number of surveys have evidenced these inequalities worldwide. See, for example, the policy brief run by the European Comission "Educational inequalities in Europe and physical school closures during COVID-19": https://ec.europa.eu/jrc/sites/jrcsh/files/fairness_pb2020_wave04_covid_education_jrc_i1_19jun2020.pdf

schooling or tutoring, we expect their grades to be higher than in the past (given that the exam is substantially easier). Data about the distribution of grades are not available yet, but we do know that the fraction of students with a grade higher than 9 over 10 has increased and so has the fraction of students who failed.[38] Given that access is determined by the weighted average of Selectivitat and high school grades, it is likely that if a student was affected by the outbreak, Selectivitat reduced her capacity to compete for selective programs, while for those less affected by the outbreak, most variation in entry grades was most likely coming from high school grades, since variation coming from Selectivitat was smaller at the top. Future access to these data should help understand the implications of the chosen policy further.

In the UK a predictive model imputed external examinations' grades.[39] When the results came out there were a number of complaints. A first issue related to how the authorities dealt with schools with small class sizes or limited historical data (mostly private schools): in those cases, high school evaluations were used as the predictor. Instead, for other schools a prediction model similar to the one proposed in this paper was introduced and hence, inequities present in societies were mirrored by the prediction.[40] In particular, schools located in low SES neighborhoods found that their prediction was lower than the prediction of students with the same high school GPA but in more affluent neighborhoods. In Catalonia, similarly, in schools with high concentration of low SES students, on average Selectivitat is lower than high school GPA. Therefore, for some of them the prediction of a model with school fixed effects is lower than the prediction without school fixed effects.[41]

The International Baccalaureate has followed a different strategy that has also induced a

[38]Number of failures may also be affected by the fact that the fraction of students who graduated from high schools also increased. The Spanish press covered extensively the topic, see for instance https://www.elperiodico.com/es/sociedad/20200728/selectividad-catalunya-2020-notas-pau-8056577 or https://www.lavanguardia.com/vida/20200806/482688617975/notas-corte-selectividad-acceso-universidad-catalunya-2020.html

[39]England, Northern Ireland, Wales, and Scotland regulate separately their exams, but they adopted a similar approach and incurred in similar controversies. For a summary, see https://en.wikipedia.org/wiki/2020_UK_GCSE_and_A-Level_grading_controversy. The controversy was covered by The Guardian, the BBC and on twitter as #alevels2020.

[40]More specifically, teachers formulate a recommendation for the grades of the high-stakes exams, based on student performance in school and in mock exams administered during the year. For the first group of schools the final grades simply mirrored the teachers' proposals. For the other schools, proposals were adjusted using historical data.

[41]In the end, in the UK the administrations decided not to use any prediction model, but to use teacher evaluations for all students, because the prediction model was harming more vulnerable students.

large number of complaints.[42] In their case the problem is slightly different. They used historical data on grades provided by teachers and analysed how they related to final grades. However they also complemented grades provided by teachers this year with externally provided grades. The problem was that it was not clear how they used these external evaluations to correct their prediction models, since this external source of data was only available to them this year, not in historical data. This lack of transparency complicated the process and led to a subsequent adjustment in grades.

To sum up, COVID-19 has inhibited the capacity to run high-stakes exams and has forced countries to think how to resolve college admissions, usually heavily based on high-stakes exams. This paper highlights some of the challenges of using historical high school data to predict high-stakes exams in a given year, but also indirectly opens an important question: is the result of the high-stakes exams what we want to predict? High-stakes exams are used to measure college preparedness and originated in an era when college performance data were limited. Nowadays, it seems relevant to study whether high-stakes exams capture skills that other data do not capture, and whether they are better predictor of success in college than performance in high school. Also, in cases like the current pandemic, given that in most countries data on high school and college performance are available, we may want to contrast predicting high-stakes exams versus directly predicting college performance to resolve college admissions.

References

Arenas, A. and Calsamiglia, C. (2019). *Gender differences in high-stakes performance and college admission policies*. Mimeo IPEG.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, **11**, 685–725.

Azevedo, E. M. and Leshno, J. D. (2016). A Supply and Demand Framework for Two-Sided Matching Markets. *Journal of Political Economy*, **124** (5), 1235–1268.

[42] See the the official announcement at https://www.ibo.org/news/news-about-ib-schools/the-assessment-and-awarding-model-for-the-diploma-programme-may-2020-session. Also see #ibscandal on twitter.

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

Cai, X., Lu, Y., Pan, J. and Zhong, S. (2019). Gender Gap under Pressure: Evidence from China's National College Entrance Examination. *Review of Economics and Statistics*, **101** (2), 249–263.

Calsamiglia, C. and Loviglio, A. (2019). Grading on a curve: When having good peers is not good. *Economics of Education Review*, **73**, 101916.

Estevan, F., Gall, T. and Morin, L.-P. (2020). Redistribution without Distortion: Evidence from An Affirmative Action Programme At a Large Brazilian University. *The Economic Journal*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*, pp. 1137–1143.

Kuhfeld, M., Soland, J., Tarasawa, B., Johnson, A., Ruzek, E. and Liu, J. (2020). *Projecting the potential impacts of COVID-19 school closures on academic achievement*. Tech. Rep. 226, Annenberg Institute at Brown University.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, vol. 26. Springer.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics*, **167**, 263–279.

Maldonado, J. E. and De Witte, K. (2020). *The Effect of School Closures on Standardised Student Test Outcomes*. Discussion paper series 20.17, KU Leuven Department of Economics.

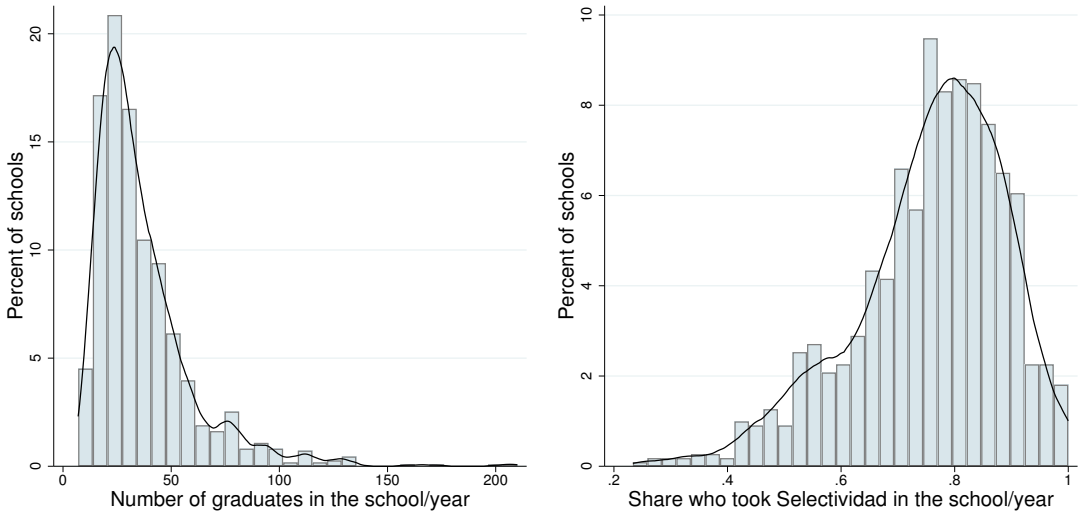
Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, **31** (2), 87–106.

Ors, E., Palomino, F. and Peyrache, E. (2013). Performance Gender Gap: Does Competition Matter? *Journal of Labor Economics*, **31** (3), 443–499.

Schlosser, A., Neeman, Z. and Attali, Y. (2019). Differential performance in high versus low stakes tests: evidence from the GRE test. *The Economic Journal*, **129** (623), 2916–2948.

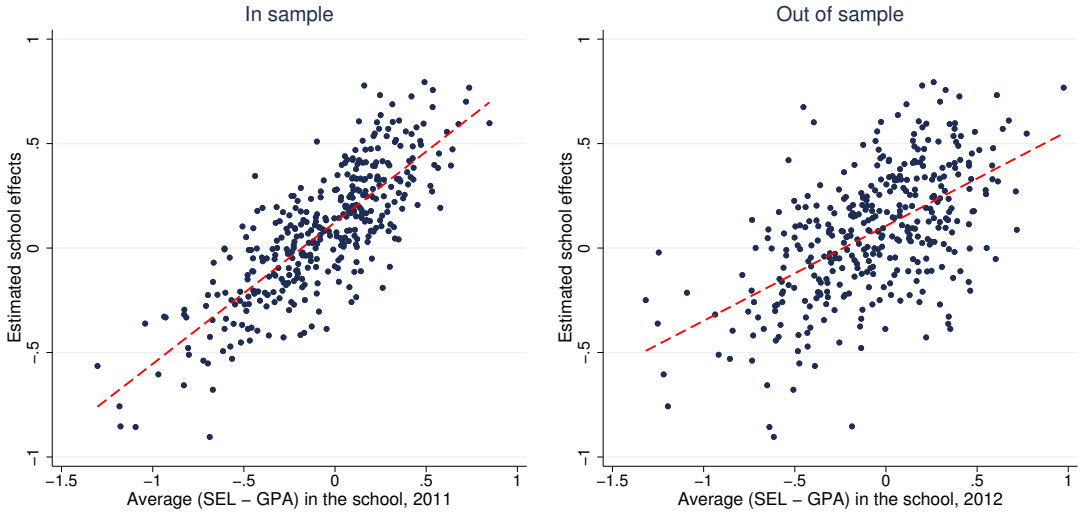
A Additional Tables and Figures

Figure 9: High schools



Note. Distributions of school size (i.e. number of graduates) and share of students who took Selectivitat, years 2010-2012. Only schools with at least 6 students who took Selectivitat in each year are included. The continuous black lines are kernel density plots.

Figure 10: Estimated school effects and raw differences in the school



Note. The graph plots data for schools with at least 6 students in the sample for each year. The dotted lines are a linear fit.

Table 5: Selectivitat. Robustness checks.

(a) In sample

	Cohorts 2010-2011			Cohorts 2011			Large schools		1-1 match
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High School GPA	0.661**	0.662**	0.671**	0.668**	0.157**	0.0389	0.685**	0.118**	0.0942**
	(0.0129)	(0.0102)	(0.0146)	(0.0151)	(0.0314)	(0.0247)	(0.0087)	(0.0268)	(0.0197)
Art Track		-0.107*	-0.0711		-0.0754	-0.0535		-0.0473	-0.0384
		(0.0466)	(0.0652)		(0.0672)	(0.0587)		(0.0647)	(0.0581)
Scientific Track		0.00422	-0.0115		-0.0703**	-0.0947**		-0.0534**	-0.0637**
		(0.0149)	(0.0155)		(0.0205)	(0.0214)		(0.0162)	(0.0143)
Peer average GPA		-0.0424							
		(0.0770)							
Catalan					0.0496*	0.124**		0.104**	0.112**
					(0.0194)	(0.0136)		(0.0139)	(0.0113)
Spanish					0.0638**	0.103**		0.0882**	0.0846**
					(0.0193)	(0.0130)		(0.0144)	(0.0110)
History/Philosophy					0.0977**	0.127**		0.116**	0.120**
					(0.0164)	(0.0157)		(0.0136)	(0.0117)
Foreign Language					0.353**	0.362**		0.351**	0.356**
					(0.0132)	(0.0116)		(0.0120)	(0.0095)
Track subjects					0.105**	0.117**		0.0901**	0.0934**
					(0.0179)	(0.0147)		(0.0139)	(0.0107)
School FE	No	No	Yes	No	No	Yes	No	Yes	Yes
N	21426	21426	21426	10367	10367	10367	12474	12474	19684
R^2	0.437	0.438	0.510	0.445	0.509	0.606	0.467	0.598	0.590
RMSE	0.750	0.750	0.700	0.744	0.700	0.627	0.728	0.633	0.639
MAE	0.600	0.600	0.556	0.591	0.553	0.490	0.585	0.504	0.508
Abs. rank deviation	0.188	0.188	0.171	0.185	0.169	0.146	0.184	0.153	0.153

(b) Out of sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
N	10728	10728	10728	10654	10654	10654	6381	6381	10260
R^2	0.423	0.424	0.447	0.424	0.493	0.487	0.448	0.527	0.519
RMSE	0.759	0.759	0.743	0.758	0.711	0.716	0.741	0.685	0.692
MAE	0.608	0.608	0.593	0.607	0.567	0.568	0.593	0.546	0.550
Abs. rank deviation	0.190	0.190	0.183	0.189	0.173	0.172	0.186	0.166	0.166

Note. Errors are clustered at the school level. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 6: Selectivitat. Regressions and predictions adding individual characteristics.

(a) In sample (cohorts 2010 and 2011)

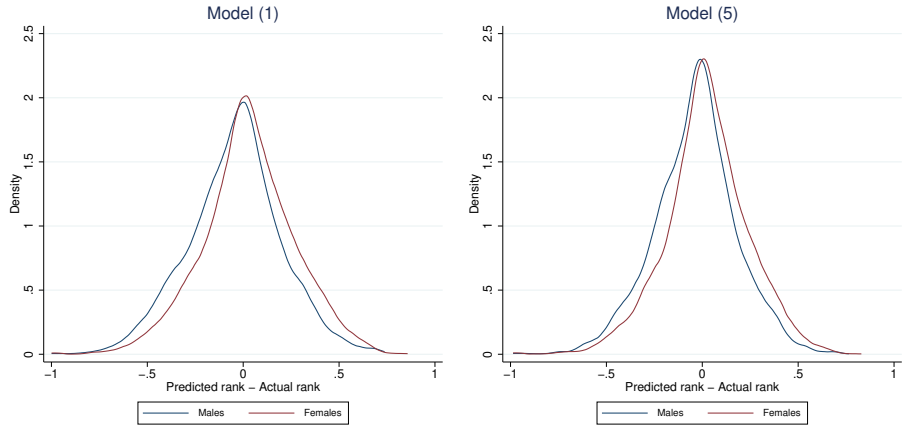
	(1)	(2)	(3)	(4)	(5)
Female	-0.198**		-0.193**		-0.194**
	(0.0107)		(0.0107)		(0.0108)
Mother high school		0.0224+	0.0220+		0.0216+
		(0.0125)	(0.0125)		(0.0124)
Mother college		0.0574**	0.0495**		0.0462**
		(0.0156)	(0.0155)		(0.0145)
Father high school		0.0503**	0.0447**		0.0438**
		(0.0124)	(0.0124)		(0.0131)
Father college		0.0897**	0.0790**		0.0773**
		(0.0128)	(0.0127)		(0.0128)
School FE	Yes	Yes	Yes	Yes	Yes
Other Parental background	No	Yes	Yes	No	Yes
Zip Code FE	No	No	No	Yes	Yes
N	20425	20425	20425	20425	20425
R^2	0.593	0.587	0.595	0.593	0.603
RMSE	0.637	0.642	0.636	0.638	0.630
MAE	0.504	0.508	0.503	0.506	0.499
abs rank deviation	0.151	0.153	0.151	0.152	0.150

(b) Out of sample (cohort 2012)

	(1)	(2)	(3)	(4)	(5)
N	10654	10654	10654	10654	10654
R^2	0.518	0.516	0.521	0.508	0.517
RMSE	0.693	0.695	0.691	0.701	0.695
MAE	0.549	0.551	0.547	0.556	0.552
abs rank deviation	0.165	0.166	0.165	0.168	0.166

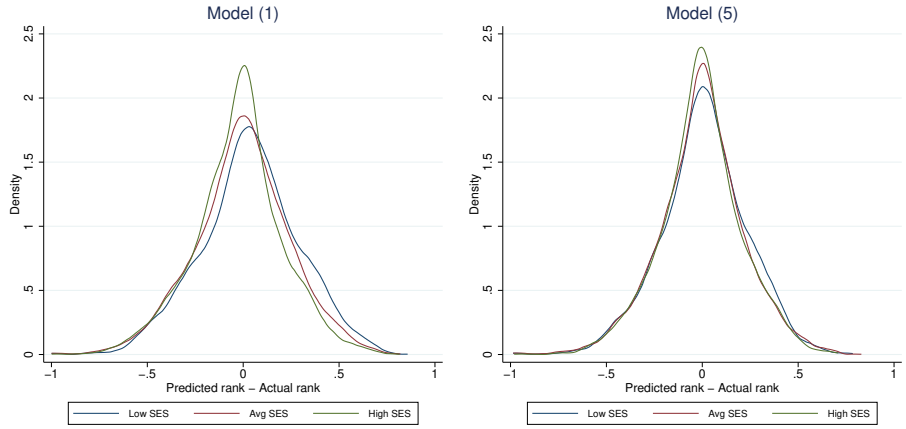
Note. "Male", "Mother primary/middle school", "Father primary/middle school" are the omitted categories when gender and/or parental education variables are used. Other parental background variables include in (2), (3), and (5) are, for both mother and father, dummies for missing info about education, and dummies for broad categories of occupation (Manager, Services, Blue collar job, Does not work or retired). Errors are clustered at the school level. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Figure 11: By gender



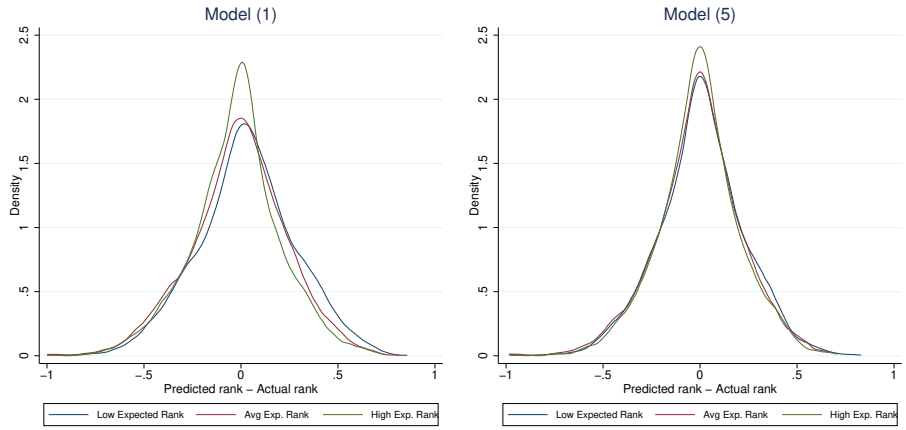
Note. Differences between predicted and actual rank, cohort 2012

Figure 12: By parental background



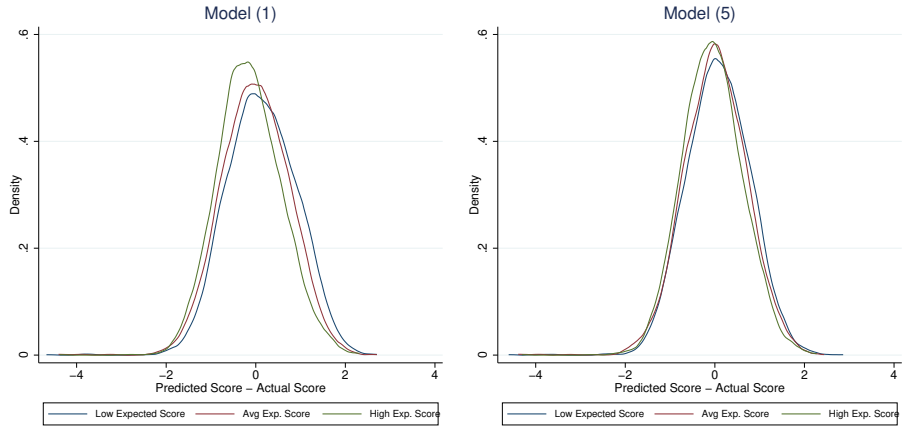
Note. Differences between predicted and actual rank, cohort 2012

Figure 13: By expected score



Note. Differences between predicted and actual rank, cohort 2012

Figure 14: By expected score



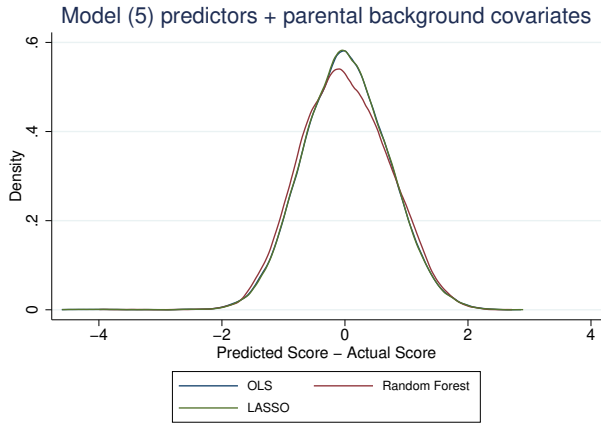
Note. Differences between predicted and actual rank, cohort 2012

Figure 15: By Expected Score

	Predicted - Actual Score		| Predicted - Actual Score |		Predicted - Actual Rank		| Predicted - Actual Rank |	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
High Expected Rank	-0.279***	-0.133***	-0.0417***	-0.0281**	-0.0513***	-0.0101	-0.0255***	-0.0135***
	(-15.67)	(-8.06)	(-3.86)	(-2.80)	(-8.85)	(-1.95)	(-6.84)	(-4.03)
Avg Expected Rank	-0.144***	-0.0682***	-0.0266*	-0.0107	-0.0391***	-0.0120*	-0.00808*	-0.00178
	(-7.96)	(-4.08)	(-2.44)	(-1.04)	(-6.57)	(-2.26)	(-2.14)	(-0.52)
Constant	0.136***	0.0796***	0.629***	0.566***	0.0290***	0.00629	0.201***	0.172***
	(10.53)	(6.70)	(80.19)	(78.00)	(6.76)	(1.67)	(74.19)	(71.07)
Model	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)
R^2	0.0223	0.00596	0.00143	0.000729	0.00786	0.000576	0.00453	0.00174
N	10654	10654	10654	10654	10654	10654	10654	10654

Header indicates the dependent variable. Omitted category: low expected rank. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Expected rank based on a regression of test scores on parental characteristics.

Figure 16: OLS vs. LASSO vs. RF (w parental covariates)



Note. Differences between predicted and actual score, cohort 2012

Table 7: Simulation. Differences in rank by subgroups

(a) Quantiles of Selectivitat

	$\sigma = 0.1$		$\sigma = 0.5$		$\sigma = 1$		$\sigma = 2$	
	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$
1 (low SEL)	0.2	1.8	4.3	8.7	11.8	16.2	21.5	25.2
2	0.1	2.8	1.4	12.5	3.9	19.0	7.2	23.2
3	-0.1	2.7	-2.1	12.1	-4.9	18.9	-7.7	23.4
4 (high SEL)	-0.2	1.5	-3.6	7.4	-10.8	14.9	-21.0	24.6
All	0.0	2.2	0.0	10.2	0.0	17.2	0.0	24.1

(b) Parental education

	$\sigma = 0.1$		$\sigma = 0.5$		$\sigma = 1$		$\sigma = 2$	
	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$	$\tilde{r} - r$	$\lvert\tilde{r} - r\rvert$
Low	0.0	2.2	0.8	10.4	2.5	17.4	4.0	23.9
Average	0.0	2.2	0.3	10.3	0.3	17.3	0.8	24.2
High	-0.1	2.1	-0.9	9.9	-2.1	17.0	-3.8	24.2

B Selectivitat Total

In the main analysis, we have focused on the first and compulsory component of Selectivitat, the so-called Selectivitat General, because it has the advantage of having a common structure for all students in the sample. The second part, the Selectivitat Específica, consists of exams in the remaining high school track subjects, and is not compulsory. At most, two additional exams may count for admission, and the weight that these subjects carry in the admission score depends on how relevant they are for the program where the student is applying. In this section we provide the main results when predicting the Selectivitat Total score (sum of the general and specific parts), which is the relevant score for admission. This additional analysis leads to very similar conclusions.

The admission score is computed according to the following formula:

$$\begin{aligned}\text{Admission score} = 60 \times (\text{High School GPA}) + 40 \times (\text{Selectivitat General GPA}) \\ + \underbrace{W_A \times (\text{Field Subject A GPA}) + W_B \times (\text{Field Subject B GPA})}_{\text{Selectivitat Específica}}, \end{aligned} \tag{8}$$

where W_A, W_B can be 10 or 20 depending on their relevance for the degree where the student is applying. The initial score used in the admission process is computed using weights and exams that maximize the admission score for the preferred program in the application list. If the student cannot be allocated to the preferred choice, the score for the second best program is used, and so on moving down the list.

In June of each year, a first round of the application algorithm to match students and programs is performed. At this point, students have the option to accept the outcome of the algorithm: to do so they have to delete the programs in their application list above the assigned one, and they can then enrol in the accepted program. In our data, these students are recorded as having accepted their preferred choice, and we do not observe deleted programs. Alternatively, they can wait for additional re-allocation rounds performed during Summer: in case some other student initially allocated to one of their most preferred programs do not enrol, they can take their slot. For this second group of students we observe the full list of preferences.[43] Overall, 78% of students in our sample are recorded has having enrolled in their

[43] As a further alternative, in September they can submit a new application for the programs which have not

preferred choice; this suggests that probably many of them accepted the initial allocation after the first round without waiting for additional rounds.

In the analysis in this section we compute "Selectivitat Total" as a weighted sum of Selectivitat General and Selectivitat Específica. For the latter, we use exams and weights for the top program that we observe, with the caveat that it is either the actual preferred choice, or the program accepted after deleting the higher ranked options. In our sample, 85% of students take at least one additional exam, and 55% take two additional exams. The average unconditional weight in our sample is 13.2 (with zero weights for students who do not take the additional exams). The average weight conditional on taking an exam is 18.9, in fact the weight is 20 in 90% of the cases.

Columns (1) to (5) of Table 8 replicate Table 2. Columns (6) and (7) augment the regressors in (4) and (5) with the additional covariates "Best subject" and "Second best subject". To compute those variables we associate each observed elective subjects in high school with an exam that students can undertake for Selectivitat Específica. Then, we multiply the evaluations for the weights that the associated exams have for the student preferred choice and rank the resulting scores. Finally, we identify the best and second best subject according to the rank, and use their evaluation as regressors.[44] Overall results are extremely similar to what discussed in Section 3. Model (7) has the best out-of-sample performance; for instance the MAE is 0.55 s.d. and the ARD is 16.2 percentiles.

Figures 17 and 18 and Tables 9 and 10 are the analogous of Figures 4 and 5 and Tables 3 and 4. Again, the results are very similar, with only one exception, which is that when looking at Selectivitat Total, female test scores are still over-predicted by a similar magnitude, but also the prediction error is a little less disperse (columns 3 and 4, Table 9).

reached their capacity constraint, and the same allocation procedure takes place for these students and slots.

[44] If no observed subjects has positive weight, a 0 is imputed instead.

Table 8: Selectivitat total. Regressions and predictions.

(a) In sample (cohorts 2010 and 2011)

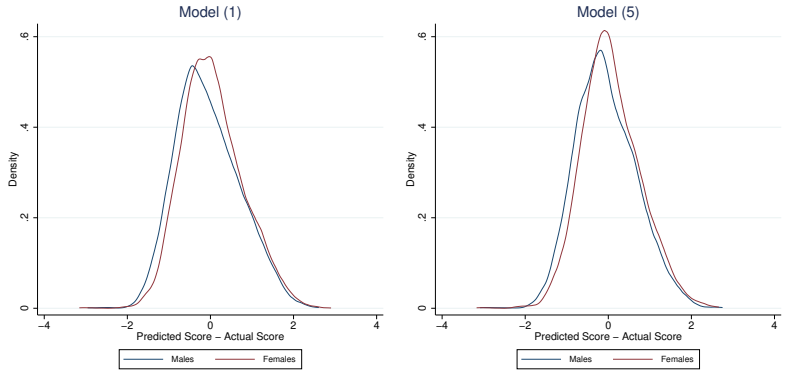
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High School GPA	0.663**	0.671**	0.686**	0.121**	0.0650**	0.132**	0.0757**
	(0.0091)	(0.0086)	(0.0100)	(0.0249)	(0.0215)	(0.0245)	(0.0212)
Art Track		0.128*	0.146+	0.105*	0.133	0.150**	0.178*
		(0.0520)	(0.0863)	(0.0509)	(0.0822)	(0.0469)	(0.0745)
Scientific Track		-0.108**	-0.123**	-0.102**	-0.117**	-0.0910**	-0.107**
		(0.0170)	(0.0165)	(0.0178)	(0.0172)	(0.0175)	(0.0169)
Peer average GPA		0.0304					
		(0.0456)					
Catalan				0.0191	0.0543**	0.0194	0.0546**
				(0.0145)	(0.0114)	(0.0141)	(0.0108)
Spanish				0.0661**	0.0749**	0.0644**	0.0736**
				(0.0155)	(0.0109)	(0.0151)	(0.0103)
History/Philosophy				0.0564**	0.0761**	0.0523**	0.0717**
				(0.0122)	(0.0093)	(0.0123)	(0.0092)
Foreign Language				0.194**	0.191**	0.194**	0.190**
				(0.0102)	(0.0078)	(0.0102)	(0.0076)
Track subjects				0.337**	0.357**	0.151**	0.180**
				(0.0147)	(0.0122)	(0.0173)	(0.0148)
Best subject						0.0223**	0.0230**
						(0.0070)	(0.0071)
Second best subject						0.106**	0.100**
						(0.0053)	(0.0056)
School FE	No	No	Yes	No	Yes	No	Yes
N	20425	20425	20425	20425	20425	20425	20425
R^2	0.438	0.442	0.522	0.479	0.560	0.497	0.576
RMSE	0.744	0.741	0.686	0.716	0.659	0.703	0.646
MAE	0.604	0.601	0.551	0.579	0.528	0.566	0.517
abs rank deviation	0.186	0.185	0.166	0.176	0.157	0.171	0.154

(b) Out of sample (cohort 2012)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
N	10654	10654	10654	10654	10654	10654	10654
R^2	0.429	0.427	0.447	0.462	0.480	0.503	0.520
RMSE	0.754	0.755	0.742	0.732	0.719	0.703	0.691
MAE	0.605	0.605	0.589	0.584	0.570	0.560	0.547
abs rank deviation	0.184	0.184	0.178	0.175	0.170	0.167	0.162

Note. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Figure 17: By gender (Sel. Total)



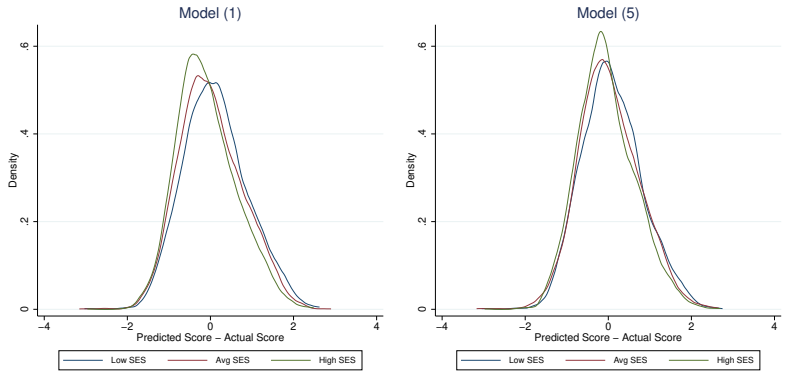
Note. Differences between predicted and actual rank, cohort 2012

Table 9: By gender (Sel. Total)

	Predicted - Actual Score		| Predicted - Actual Score |		Predicted - Actual Rank		| Predicted - Actual Rank |	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.119***	0.123***	-0.0442***	-0.0300***	0.0624***	0.0534***	-0.00600	-0.00219
	(7.97)	(8.72)	(-5.01)	(-3.49)	(13.24)	(12.18)	(-1.96)	(-0.76)
Constant	-0.0626***	-0.0520***	0.630***	0.588***	-0.0378***	-0.0326***	0.187***	0.171***
	(-5.41)	(-4.75)	(94.13)	(90.58)	(-10.39)	(-9.70)	(78.92)	(77.38)
Model	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)
R^2	0.00600	0.00714	0.00233	0.00113	0.0163	0.0137	0.000362	0.0000536
N	10654	10654	10654	10654	10654	10654	10654	10654

Header indicates the dependent variable. Omitted category: males. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 18: By SES (Sel. Total)



Note. Differences between predicted and actual rank, cohort 2012

Table 10: By SES (Sel. Total)

	Predicted - Actual Score		| Predicted - Actual Score |		Predicted - Actual Rank		| Predicted - Actual Rank |	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
High SES	-0.214***	-0.132***	-0.0303**	-0.0265*	-0.0349***	-0.0152**	-0.0186***	-0.0109**
	(-11.20)	(-7.22)	(-2.59)	(-2.37)	(-5.69)	(-2.66)	(-4.69)	(-2.94)
Avg SES	-0.107***	-0.0626***	-0.00665	-0.00148	-0.0294***	-0.0143*	-0.00360	0.0000534
	(-5.58)	(-3.44)	(-0.57)	(-0.13)	(-4.79)	(-2.53)	(-0.92)	(0.01)
Constant	0.126***	0.0920***	0.618***	0.581***	0.0229***	0.00980*	0.192***	0.174***
	(8.29)	(6.41)	(65.65)	(65.37)	(4.70)	(2.19)	(61.65)	(60.03)
Model	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)	Model (1)	Model (5)
R^2	0.0119	0.00497	0.000804	0.000782	0.00336	0.000794	0.00263	0.00128
N	10638	10638	10638	10638	10638	10638	10638	10638

Header indicates the dependent variable. Omitted category: low SES. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.