Arenas, Andreu; Calsamiglia, Caterina

Working Paper

Gender Differences in High-Stakes Performance and College Admission Policies

IZA Discussion Papers, No. 15550

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Arenas, Andreu; Calsamiglia, Caterina (2022) : Gender Differences in High-Stakes Performance and College Admission Policies, IZA Discussion Papers, No. 15550, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/265771

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

DISCUSSION PAPER SERIES

IZA DP No. 15550

Gender Differences in High-Stakes Performance and College Admission Policies

Andreu Arenas
Caterina Calsamiglia

SEPTEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15550

Gender Differences in High-Stakes Performance and College Admission Policies

Andreu Arenas
Princeton University and University of Barcelona (IEB and IPErG)

Caterina Calsamiglia
ICREA-IPEG and IZA

SEPTEMBER 2022

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

Gender Differences in High-Stakes Performance and College Admission Policies*

We investigate the effect of increasing the weight of standardized high-stakes exams at the expense of high school grades for college admissions. Studying a policy change in Spain, we find a negative effect of the reform on female college admission scores, driven by students expected to be at the top. The effect on admission scores does not affect enrolment, but the percentage of female students in the most selective degrees declines, along with their career prospects. Using data on college performance of pre-reform cohorts, we find that female students most likely to lose from the reform tend to do better in college than male students expected to benefit from the reform. The results show that rewarding high-stakes performance in selection processes may come along with gender differences unrelated to the determinants of subsequent performance.

JEL Classification: J16, I23, I24

Keywords: college admissions, high-stakes exams, gender gaps

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

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

* We are gratful to seminar participants at Bar-Ilan, Bologna, EIEF, Columbia, Florida, Kadir-Has, Maryland, Oxford, Toulouse, UAB, UB, UC3M, UPF; the Catalan Economics Society Conference, EUI Alumni Conference, IZA Workshop on Economics of Education, Nuremberg conference on Gender Economics and the Workplace, Simposio of the Spanish Economic Society, SOLE, Workshop on Public Policies: Inequality of Opportunity, ZEW Mannheim Workshop on the Economics of Higher Education, the 16th Matching in Practice Workshop; and to Sule Alan, Ghazala Azmat, Juan Dolado, Andrea Ichino, Raquel Fernández, Rosa Ferrer, Naomi Friedman-Sokuler, Libertad González, Nagore Iriberri, Hannes Mueller, Daniele Paserman, Perihan Saygin and Miguel Urquiola for helpful comments. Arenas acknowledges funding from the Spanish Ministry of Science PID2020-120359RA-I00.

1 Introduction

The number of students attending higher education has more than doubled in the last decades (UNESCO, 2017). The increase is largest in developing countries but in developed countries most young adults already attend college. The field and the institution of enrolment have been shown to have a large impact on life prospects such as earnings (Kirkebøen *et al.*, 2016), whom one marries (Kirkebøen *et al.*, 2021) and even the well-being of potential children (Kaufmann *et al.*, 2021), particularly so for women. Hence, the mechanisms determining who has access to higher education and where do have a large impact on society.

Around the world, college admission decisions are based mainly on two types of inputs about candidates. On the one hand, on test scores from standardized exams, such as the SAT in the US, the Vestibular in Brasil, the Gaokao in China, the Bagrut in Israel, or the OSS in Turkey.[1] On the other hand, on measures of continuous assessment over a longer time horizon, such as high-school grades or extra-curricular activities.

In many countries, colleges and majors are allocated through a centralized procedure. Applicants submit a rank-ordered list of preferences and college-majors order students by some weighted average of their GPA in high school and in a standardized exam. Such procedures are in place in China, Korea, Chile, Norway, Brasil, Israel or Spain, to name a few. Hence, standardized exams often determine what and where individuals can study and therefore can be considered the highest stakes exams that affect individuals in a society. In Korea, for instance, Seoul closes shops, banks, and even the stock market opens late the day that the Suneung (the standardized college admission exam) takes place.

A large literature has documented gender differences in performance in high-stakes and competitive environments, even after controlling for ability. In lab experiments, men's performance tends to be more elastic to the competitiveness of the environment than women's.

[1] Other examples include the A-levels (UK), the Maturità (Italy), Bac (France), Abitur (Germany), Suneung (Korea) or Selectividad - PAU (Spain).

Examples include solving mazes in tournaments (Gneezy *et al.*, 2003) or running in a physical education class (Gneezy and Rustichini, 2004). Females are also less likely to self-select into competitive tournaments, even after controlling for performance, confidence and risk aversion (Niederle, 2015; Niederle and Vesterlund, 2011, 2007). These gender differences in competitiveness do not necessarily relate to relevant differences in qualifications or subsequent performance: for instance, experiments by Balafoutas and Sutter (2012) and Niederle *et al.* (2013) find that affirmative action interventions encourage women to enter competitions more often, and performance is at least equally good, both during and after the competition.

Gender differences in high-stakes exams in educational competitive settings have been found in various countries (Jurajda and Münich, 2011; Saygin, 2018; Montolio and Taberner, 2018; Iriberri and Rey-Biel, 2019; Arenas *et al.*, 2021). For instance, Schlosser *et al.* (2019) and Cai *et al.* (2018) find significant gender differences in performance between mock and actual GRE and Gaokao (the Chinese college admission exam) tests. Azmat *et al.* (2016) find that throughout secondary and high school, girls always outperform boys, but especially in lower-stakes exams. Ors *et al.* (2013) find that male students outperform female students in admission exams of the most selective French Business School, but not in first-year courses nor in high school. Morin (2015) finds that male average grades and the proportion of male students graduating on time in college increased relative to females within a cohort of students in Canada which was exceptionally large, which increased competition for grades.

Overall, these findings suggest that selection processes emphasizing high-stakes performance may be consequential for gender differences in outcomes. However, the importance of such policies for gender differences in the field is still an open question. First and foremost, students' effort may react to changes in admission policies, possibly offsetting or amplifying the effect of re-weighting baseline differences in test scores across high and low-stakes exams. And secondly, the consequences of any effects on admissions depend on who are the compliers and whether they tend to compete for the same academic programmes.

A second open and important question concerns the matching or efficiency implications of such policy changes. In particular: are the male students who react positively to high-stakes better prospective college students than the female students who instead react negatively to higher stakes? In other words, are gender differences in high-stakes performance related to gender differences in college performance potential? The answer to this question is important to understand whether designing admission policies with different weights on high-stake exams entails a trade-off between gender inequality and match quality.

In this paper, we study the effect of a policy change which increased the weight of the high-stakes standardized exam for (centralized) college admissions in Spain from 40% to 57%, using administrative data on college applications and college performance in the region of Catalonia, which hosts some of the best universities of Spain. First, we study the effect of the reform on gender differences in admission scores. Second, we quantify the effect of the reform on gender differences in college enrolment, college selectivity and career prospects. Last but not least, we study the relationship between gender differences in high-stakes performance and college performance skills, by studying what type of students (based on their potential for college performance) are most affected by the reform.

The three main results of the paper are the following. First, we find a negative effect of increasing the weight of the high-stakes exam on female admission scores. The size of the effect is similar to the date of birth effect in our sample (i.e., the effect of being born in January rather than in December); to 15% of the parental college education gradient in admission scores in our sample; or to the effect of taking an exam in a day with high pollution (Ebenstein *et al.*, 2016). The effect is slightly larger than the effect of re-weighting high school grades (where females largely outperform males) and high-stakes grades (where there are smaller differences in performance) differently. This suggests that students' reaction to the policy did not attenuate, and instead slightly amplified, the consequences of the policy.

Second, we study the effect of the reform on students' allocation to college. This effect depends on who are the most affected students and whether they are competing for the same academic programmes. We find no effect of the reform on college enrolment, because the effect on admission scores is driven by students expected to be top performers. This is consistent with previous evidence finding that performance gaps at high percentiles are related to the differential manner in which men and women respond to competitive test-taking environments (Niederle and Vesterlund, 2010). However, we do find that female students become significantly less likely to attend the most selective programmes. Enrolment in programs above the median level of selectivity declines by 3pp, compared to enrolment in programs below the median. We also estimate that this change in the allocation to college leads to worse career prospects for female students: a 2pp points increase in the expected gender wage gap, on top of a pre-reform gender wage gap of 20% four years after graduation.

Third, we study the correlation between gender differences in high-stakes performance and college performance skills. Using machine learning techniques, we identify the types of students who are most likely to benefit from the reform, based on a large set of pre-determined covariates. Focusing on pre-treatment cohorts, we compare the college performance of predicted winners and predicted losers from the reform. Within gender, we find that students expected to win from the reform tend to perform better in college than comparable students which had the same admission grade and were enrolled in the same college-major and pre-reform cohort. This suggests a positive relationship between high-stakes and college-performance skills. However, across genders, the sign of this relationship flips. We find that females predicted to lose from the reform are better college performers than comparable male students who are predicted to win from the reform. Hence, this suggests that the gender difference in high-stakes performance is not related to gender differences in college performance skills (if anything, they are negatively related).

We make three novel contributions to the literature. First, we study the effect of gender differences in high-stakes performance on admission grades, under different admission policies that give more or less weight to high and low-stakes performance, keeping competition constant. This is important because while the literature shows a gender gap in performance due to competition and high stakes, it remains an open question whether these would change under alternative policies. This case study deals directly with this policy question, and our estimates capture any equilibrium effects that may arise such as effort shifting across exams.

Second, we study and quantify the consequences of such gender differences for college allocation and career prospects, which depend on subtle interactions between the response to high-stakes and students' preferences (i.e., on whether the most affected students are competing between them for the same slots, and on what are the next-best options of students losing from the reform).

Third, we characterise the compliers' profile and relate it to a relevant trait that policymakers would like to select for (in this case, college performance skills). This is important because it allows us to jointly evaluate the distributional and efficiency implications of policies that put more or less weight on high-stakes performance.

Finally, our results directly speak to a large number of countries which make use of very similar centralized college allocation mechanisms, but which differ in the weights given to high school and high-stakes GPAs, such as Chile, China, Croatia, France, Germany, Hungary, Ireland, Sweden, Portugal, Russia, Turkey, or Ukraine.[2]

[2]Source: matching-in-practice.eu

2 Background and policy change

The college allocation process starts with students listing their preferences in an application form. Then, they are allocated to academic programmes (i.e., pairs major × university) based solely on their admission grades, which are a weighted average of high school grades and grades in a comprehensive high-stakes exam at the end of high school, namely the PAU (Proves d'Accés a la Universitat), which covers the contents of high school. High school lasts for two years, and students specialize in one of five possible specialities: arts, humanities, social science, science, or technology. The high-stakes exam (PAU) includes exams on core subjects common for all high school students (namely Catalan, Spanish, English, and Philosophy or History), and on three field subjects corresponding to the students' specialization in high school. Students are then allocated into academic programmes (i.e. pairs college-major), which are capacity constrained, using a Gale-Shapley mechanism (Gale and Shapley, 1962). Every year, the admission grade of the last student admitted into an academic programme becomes public and it is known as the *threshold grade*. The allocation is managed by regions, and it follows the same standard Gale-Shapley mechanism for the slots of public universities in every region.[3]

Before and after the 2010 reform, the admission grade was computed as follows:

- Before 2010:

$$\text{Admission Grade} = \frac{60 \times \text{High School GPA} + 40 \times \text{high-stakes GPA}}{100}$$

[3]Every student has to fill out an application form for every region where she is applying to college.

■ After 2010:

$$\text{Admission Grade} = \frac{60 \times (\text{High School GPA}) + 40 \times (\text{high-stakes GPA, Core + Field Subject A})}{140}$$

$$+ \quad \frac{W_B \times (\text{high-stakes GPA, Field Subject B}) + W_C \times (\text{high-stakes GPA, Field Subject C})}{140}$$

Where W_B, W_C can be 10 or 20 depending on the subject relevance for the degree where the student is applying (and could be zero if the student does not take the exam).[4] This means that the post-reform high-stakes exam amounts to up to $\frac{80}{140} \approx 57\%$ of the admission grade, a substantial increase from the pre-reform weight (40%).

Besides increasing the weight of the high-stakes exam, the reform comes along with two additional relevant changes, which we will also study to understand whether they could be confounding any effects driven by the change in the weight of the high school versus the high-stakes exam. First, there are changes in the relative weights of subjects within the high-stakes exam GPA. Indeed, the main reason for the reform was to increase the weight of field subjects for college admissions. However, this was done in a way that led to a quite large change in the overall weight of the high-stakes exam compared to the high school GPA. Field subjects account for up to 60% of the high-stakes GPA after the reform, compared to 50% before the reform.[5] If there are systematic gender differences in performance in field vs. core subjects, this could have an effect on admission grades beyond the change in the weights of high school and high-stakes GPAs.

Second, after the reform, the weight of two field subjects may change depending on whether the student is being considered for enrolment in a related field. In practice, because students tend to apply and enrol into programmes related to their high school studies, W_B

[4] 60% take both, 25% only one, and 15% none.

[5] Before the reform, each core subject in the high-stakes exam counted for 12.5%; one field subject for 10%, and two field subjects for 20%. After the reform, each core subject counts for 10%, one field subject for 10%, and two field subjects for up to 25%.

and W_C are on average 19, conditional on taking the field exams. Table A1 in the Appendix shows that there are no significant gender differences neither in taking field subjects exams nor on the average weights. As a benchmark, we use the admission grade with W_B and W_C from students' program of enrolment, but as a robustness check, we will also provide estimates using $W_B = W_C = 19$ for all students, as well as controlling for them.

2.1 Data

The main data source for this paper consists of administrative records on enrolment applications to public universities in Catalonia, a large region of Spain with some of the best universities in the country (for instance, according to the 2018 Times Higher Education World University Ranking, five out of the seven best Spanish Universities are in Catalonia).[6] Cross-region student mobility for undergraduate studies in Spain is low, such that 85% of students stay in their region for undergraduate studies.[7] In the period of analysis, 90% of students in Catalonia attend public universities, where tuition fees are highly subsidized.[8] In 2018, Catalonia's GDP per capita in purchasing power standard (PPS) was €33200, slightly above both the Spanish (€28100) and the EU (€31000) averages (Eurostat, 2020).

We use administrative data on all applicants to Catalan universities, on the regular track (high school + PAU), who took the high-stakes exam and applied to college every year between 2006 and 2012. The main outcome variables in the sample are the students' Admission Grades and their Academic Programme (degree × university) of admission. The main predetermined covariates in the sample are parental and maternal education and occupation, postal code of residence, and high school. For every programme, we compute, every year, the threshold grade of admission, which is the lowest admission grade of a student that managed

[6] Universitat Pompeu Fabra (1st), Universitat Autònoma de Barcelona (2nd), Universitat de Barcelona (3rd), Universitat Politècnica de Catalunya (6th) and Universitat Rovira i Virgili (7th).

[7] Source: El Mundo. According to Eurostat, Spain is one of the EU countries where young people live with their parents for longer, leaving at age 29.5, compared to an EU average of 26.

[8] Source: https://www.idescat.cat/pub/?id=aec&n=753&t=2010

to enrol into that program, given the capacity constraints. For every programme, we observe the field of study, the faculty and the municipality where it is taught. We refer to these data as the *Selectivitat* dataset.

We combine these data with three additional datasets. First, with an administrative dataset of all students enrolled in public high schools for the post-reform period, including detailed information on their high-school grades.

Second, with a survey dataset of a sample of pre-reform students of Catalan universities, with information on their earnings four years after graduation, to compute the career prospects associated with each academic program.

Third, with an administrative dataset on college performance of pre-treatment cohorts (enrolling in college between 2006 and 2009) of the three main public universities in Catalonia (Universitat de Barcelona, Universitat Autònoma de Barcelona, and Universitat Pompeu Fabra), which enrol more than 60% of students in Catalan Public Universities.

3 Admission Grades

The top panel of figure 1 displays standardized admission grades by gender over time. It shows that before the reform, females' admission grades were around 0.14 standard deviations higher than males' admission grades and that this difference was stable over time. After the reform, this difference shrinks to around 0.08 standard deviations. Hence, the reform had a negative effect on females' admission grades. The bottom panel of figure 1 displays female-by-year coefficients, where the baseline year is 2009 (the year before the reform), showing that these differences are statistically significant. We also estimate differences-in-differences regressions:

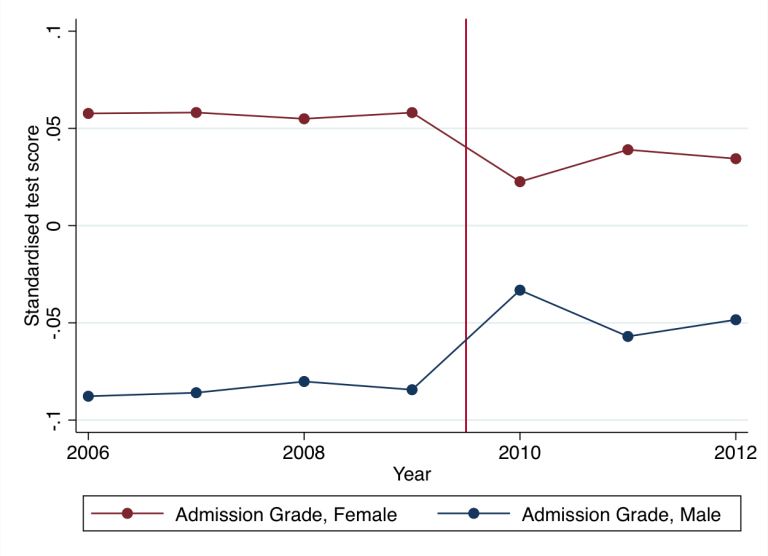
$$Admission\ Grade_{it} = \alpha_t + \beta Female_i + \gamma\left(Female_i \times Post_t\right) + \epsilon_{it}$$

Where we regress the admission grade of student i in year t on year fixed effects α_t, a female indicator, and a post-reform indicator (year = 2010, 2011, 2012) interacted with a female indicator. Table 1 reports point estimates. As suggested by figure 1, the reform had a significant negative effect on females' admission grades. Adding gender-specific time trends and controls (parental and maternal education and occupation dummies, high school, postal code, nationality), the estimates show a very similar picture. Quantitatively, the magnitude of this effect is similar to the effect of taking an exam on a high pollution day (Ebenstein *et al.*, 2016), or to the date of birth effect (January-December) in our sample, as reported in table A2 in the Appendix; or to 15% of the parental college education gradient in admission scores, as reported in table A3 in the Appendix.

It is also interesting to measure the effect in terms of students' rank, which is closely linked to college admissions. Table 2 reports point estimates of the effect on the admission grade rank, where the rank is equal to one for the highest admission grade, and zero for the lowest admission grade. Pre-reform, females were ranked 4% higher, on average, and post-reform this declines to around 2%. Again, adding gender-specific time trends or controls does not substantially change the point estimates.

Regarding the role of W_B and W_C, table A1 in the Appendix show that there are no gender differences, and table A4 in the Appendix reports estimates controlling for W_B and W_C or setting $W_B = W_C = 19$ (i.e., the average weight) for all students, with very similar results.

Figure 1: Effect of the reform on admission grades by gender



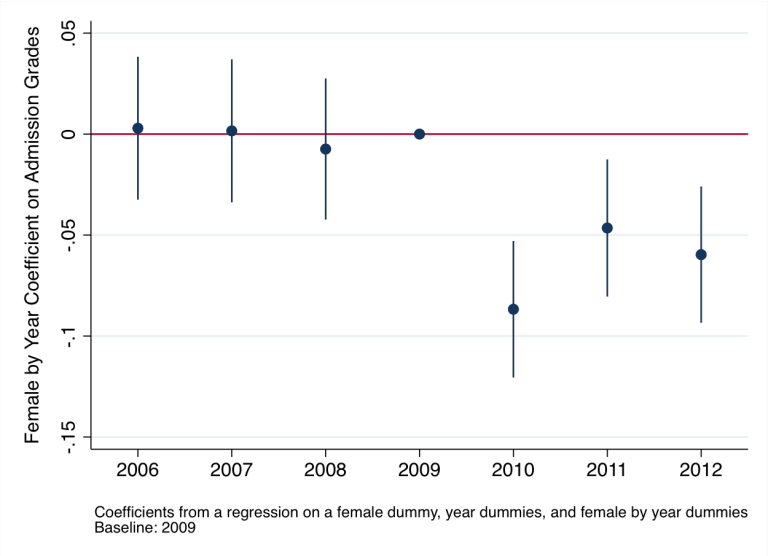


Table 1: Dependent Variable: Admission Grade

	(1)	(2)	(3)	(4)
Female × Post 2009	-0.0636***	-0.0741***	-0.0739***	-0.0662***
	(0.00953)	(0.0189)	(0.00872)	(0.0172)
Female	0.142***		0.110***	
	(0.00645)		(0.00601)	
Female	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Gender-specific trends		✓		✓
Controls			✓	✓
Mean Dep. Var	7.08e-08	7.08e-08	-0.00240	-0.00240
N	183451	183451	182259	182259

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Controls: Mother and father education and occupation, year and month of birth, nationality, high school, postal code.

Table 2: Dependent Variable: Admission Grade Rank

	(1)	(2)	(3)	(4)
Female × Post 2009	-0.0205***	-0.0222***	-0.0229***	-0.0192***
	(0.00275)	(0.00547)	(0.00251)	(0.00497)
Female	0.0430***		0.0328***	
	(0.00187)		(0.00173)	
Female	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Gender-specific trends		✓		✓
Controls			✓	✓
Mean Dep. Var	0.499	0.499	0.499	0.499
N	183451	183451	182259	182259

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Rank is equal to one for the highest score within a cohort and zero for the lowest.
Controls: Mother and father education and occupation, year and month of birth, nationality, high school, postal code.

3.1 Students' response

The reform has a significant effect on gender differences in admission grades, which are a weighted average of high-school and high-stakes grades. There are three possible sources for the estimated effect.

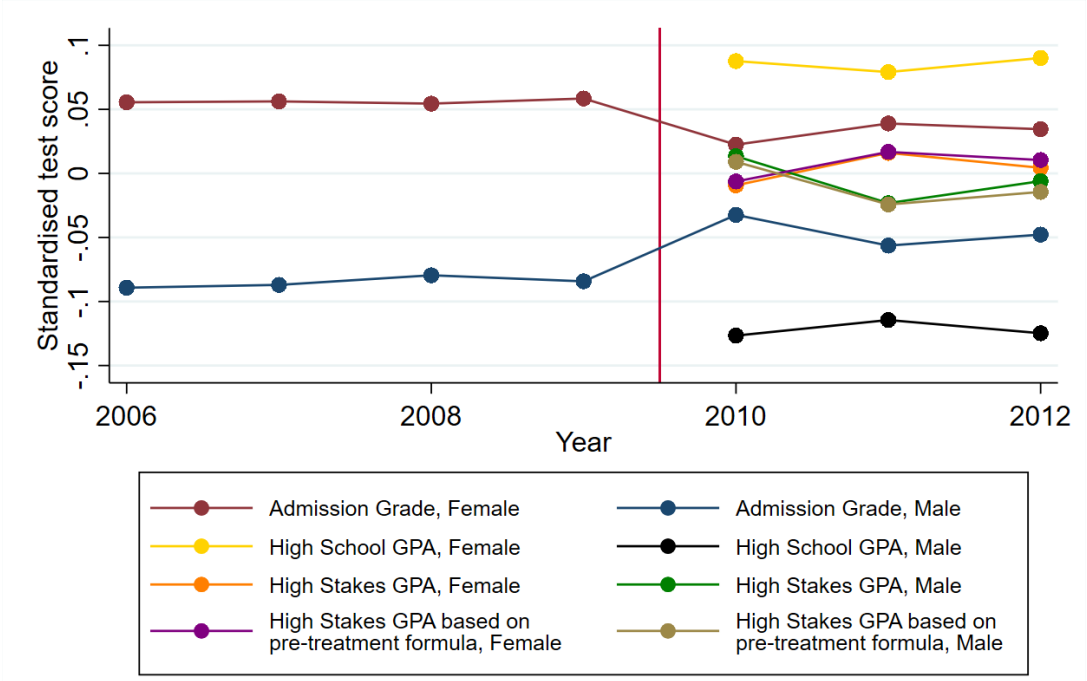
First, an equilibrium effect of the reform on gender differences in high school vs. high-stakes performance. This will happen if students' behavior or effort reacts differently to the increased importance of the high-stakes exam. Second, a re-weighting of baseline gender differences in performance between high school and the high-stakes exam. If female and male students tend to perform relatively differently in high school compared to the high-stakes exam, we would expect the reform to affect admission grades via re-weighting. Third, a re-weighting effect due to differential performance across field and core subjects in the high-stakes exam. As explained in the previous section, the reform also changes the relative weight of core and field subjects in the high-stakes exam. If female and male students tend to perform differently in field subjects compared to core subjects, we would expect this to affect the admission grade as well.

We examine these alternative mechanisms by combining the Selectivitat dataset with administrative data on admission grades, high school grades and high-stakes grades for all post-reform students in Catalan public high schools. We weight the sample of public high schools so that it matches the full sample in terms of average admission grades by year and gender, using entropy balancing (Hainmueller, 2012).[9] Using the weighted sample, we study gender differences in the different components of the admission grades.

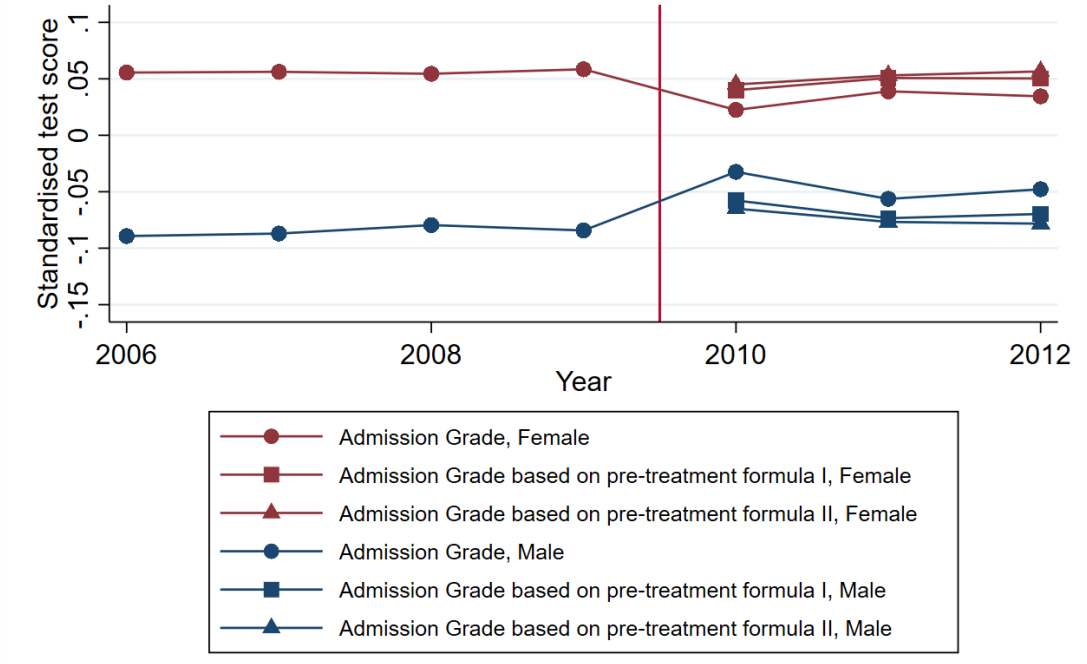
The top panel in figure 2 displays gender differences in standardized high school grades and standardized high-stakes grades. First, it shows a very large gender difference in high

[9]Weights are chosen by the following reweighting scheme that minimizes the entropy distance metric: $\min_{w_i} H(w) = \sum_{i \in Public\ Schools} w_i log(w_i)$; subject to the balance constraint that the first and second moment of the admission grade by year and gender of the re-weighted public school sample is equal to the one in the population.

Figure 2: Re-weighting and the effect of the reform



Sample: public schools, weighted to match admission scores by gender in the population via entropy balancing.
High stakes GPA based on pre-treatment formula gives 50% weight to core subjects.



Sample: public schools, weighted to match admission scores by gender in the population via entropy balancing.
Pre-treatment formula I: pre-treatment weights for high school vs. high stakes.
Pre-treatment formula II: pre-treatment weights for high school vs. high stakes
+ pre-treatment weights for core vs. field subjects within high stakes GPA.

school grades and a very small difference in high-stakes grades. This suggests that the re-weighting between baseline high school and high-stakes grades may have played an important role in the effect of the reform. Second, the figure shows both the post-treatment high-stakes GPA and a high-stakes GPA based on the pre-treatment formula, where core subjects have a 50% weight (as opposed to 60% under the new formula). The figure shows that in both cases, the gender differences in high-stakes performance remain almost identical. This suggests that the change in weights across field and core subjects in the high-stakes exam does not play an important role in the effect of the reform.

The bottom panel in figure 2 displays gender differences in standardized admission grades based on the pre-treatment formula, such that the high-stakes GPA has a weight of 40% for the admission grade, and the high school GPA a weight of 57%. It shows that if high-stakes and high school GPAs were to be re-weighted according to the pre-reform weights, the effect of the reform would have been smaller (and similarly smaller regardless of whether core subjects count for 60% or 50% of the high-stakes exam). In table A5 in the Appendix, we report estimates of the re-weighting effect of the reform. The results suggest that students' reaction to the reform slightly amplified its re-weighting effect, which explains around 75% of the total effect of the reform on admission grades.[10]

[10] We also report results for public schools, without weighting to match admission grades in the population, in table A6 in the Appendix. In this case, the effect also seems largely driven by re-weighting, although the effect of the reform on admission is smaller, which would suggest a slight behavioral reaction of the opposite sign.

4 Students' allocation to college

In this section, we quantify the consequences of the gender differences in admission grades induced by the reform. The consequences will crucially depend on who are the most affected students, and whether they are competing for the same programmes. For instance, in an extreme case where female and male students' preferences were completely segregated, any effects on gender differences in admission grades would not affect the college allocation. We study three outcomes related to the allocation of students to college: enrolment, selectivity of the program of attendance, and career prospects.

Figure 3 displays admission grades over time across the predicted admission grades' distribution. In a first step, we regress admission grades on a vector of pre-determined covariates (namely parental and maternal education and occupation dummies, high school, postal code, month of birth, and nationality), for the pre-treatment sample. Then, we split the sample according to whether students are predicted to be in different quartiles of the admission grade distribution. Figure 3 displays the effect of the reform across these groups. The main takeaway is that the most affected students are those expected to be top performers. For those expected to have lower grades, instead, the differences are small. This is consistent with the findings in Niederle and Vesterlund (2010) that performance gender gaps at high percentiles can partially be explained by the differential manner in which men and women respond to competitive test-taking environments. This also shows that the most affected students are not competing for enrolment into college, but instead for rather selective programmes. Figure 4 displays the number of enrolled students by year and gender, showing no differences due to the reform, as one would expect from figure 3.[11]

[11] Enrolment is increasing during the period of analysis, which includes the great recession, in line with the literature on the counter-cyclicality of education (Arenas and Malgouyres, 2018). Spanish regions most affected by the crisis saw gender differences in educational attainment because of diminished blue-collar labor market opportunities in the construction sector (Aparicio-Fenoll, 2016), but these compliers are unlikely to be at the high school-college enrolment margin.

Figure 3: Effect of the reform on admission grades, along the performance distribution

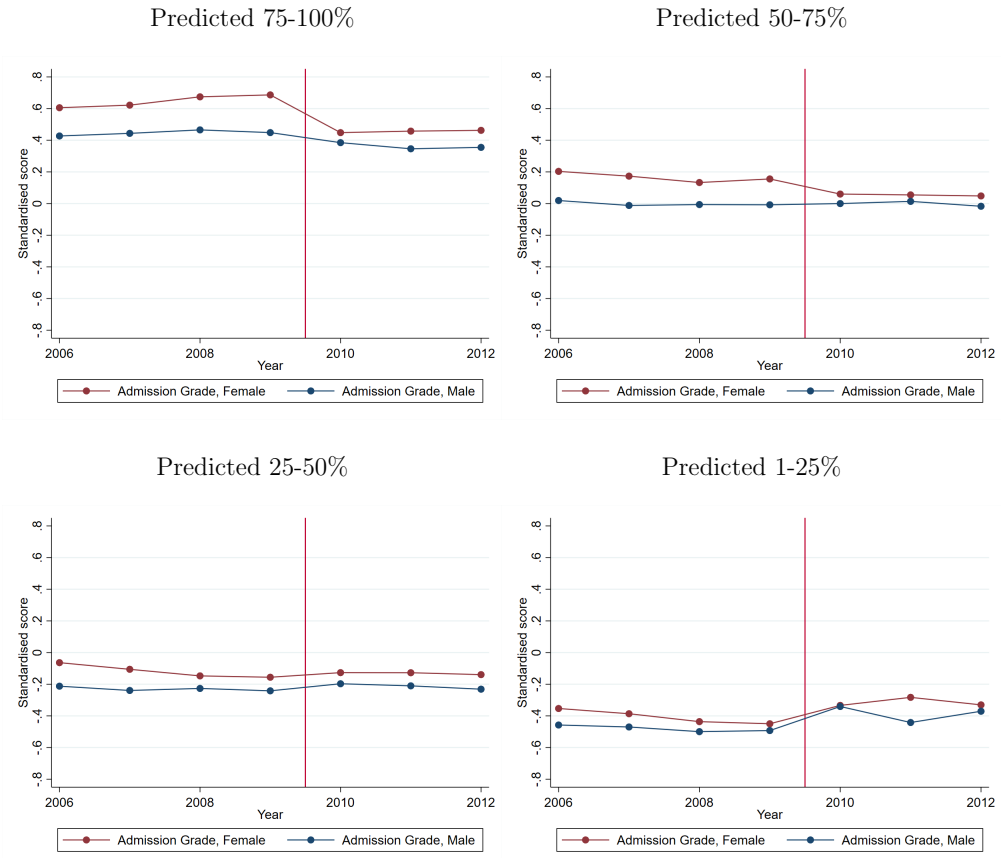
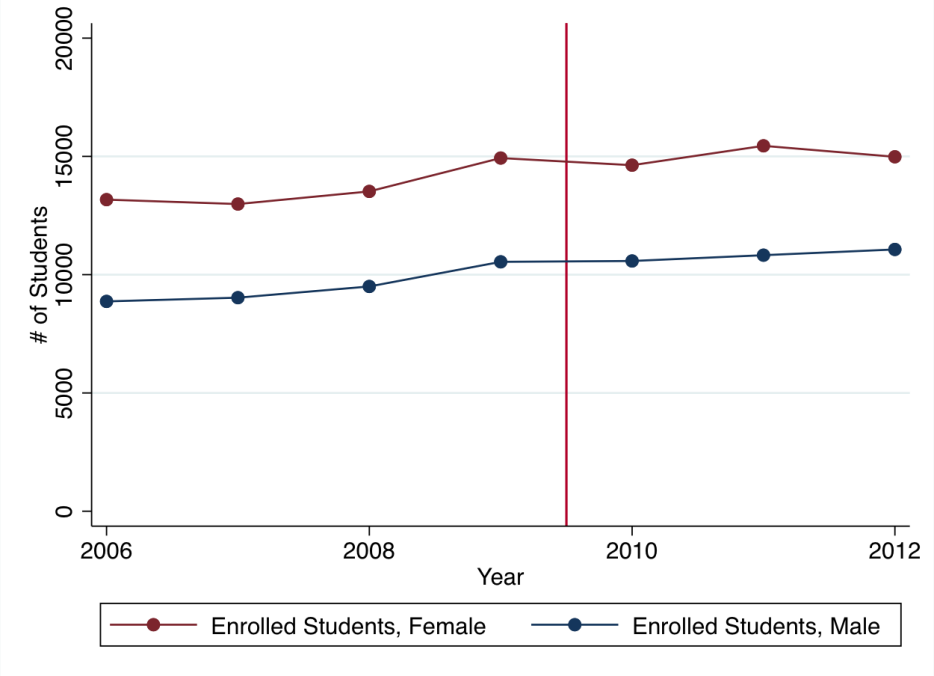


Figure 4: College enrolment



We next study the effect of the reform on another margin, namely the selectivity of the academic program attended. The Spanish setting provides a straightforward measure of access to more or less preferred or selective programmes, which is the *threshold grade* of the program of enrolment. The threshold grade is the admission grade of the student with the lowest admission grade who is admitted into a program. It is a measure of how selective is a programme, it is public information and strongly serially correlated. It is also a measure of peer quality and reputation: MacLeod *et al.* (2017) find that in Colombia, programmes' average admission grades across programmes causally matter for labour market outcomes.

To study the effect of the reform on gender differences in the selectivity of the program of enrolment, we rearrange the data and take academic programmes p as the unit of analysis and look at how the reform changes their gender composition depending on pre-reform threshold grades. Studying differences in the allocation according to pre-treatment threshold grades is useful because it keeps the measure of selectivity constant. Threshold Grades themselves are likely to be affected by the reform, and differently depending on the typical gender composition of academic programmes. An extreme case would be a scenario of full gender segregation across programmes: the reform would not change the students' allocation, but it would change average threshold grades by gender. Hence, we estimate the regression:

$$\%Females_{pt} = \alpha_p + \pi_t + \beta \left(\text{Pre-reform Thresh.Grade}_p \times Post_t\right) + \epsilon_{pt}$$

Where the outcome is the % of females in programme p in year t, and where the estimates are weighted by the number of students in each programme. Since the coding of academic programmes is fuzzy, with frequent changes that are difficult to track, we take university $\times$ faculty $\times$ municipality $\times$ field of study as the unit of analysis, for which we obtain a more balanced panel. We obtain 210 units (on average, every unit offers 2.5 programmes per year). Table A7 in the Appendix shows that this is a meaningful grouping since there is a

high serial correlation within this unit of observation in outcomes such as threshold grades, the number of enrolled students or the fraction of female students.

Figure 5 displays the fraction of female students in programs above and below the median pre-reform level of selectivity. It shows that the reform affected the students' allocation, such that the percentage of female students in the most selective programs declines after the reform. The figure suggests that the percentage of female students in the most selective programmes declined by 3pp, compared to the percentage of females in the least selective programmes.

Figure 5: Fraction of Female Students by pre-reform Threshold Grade

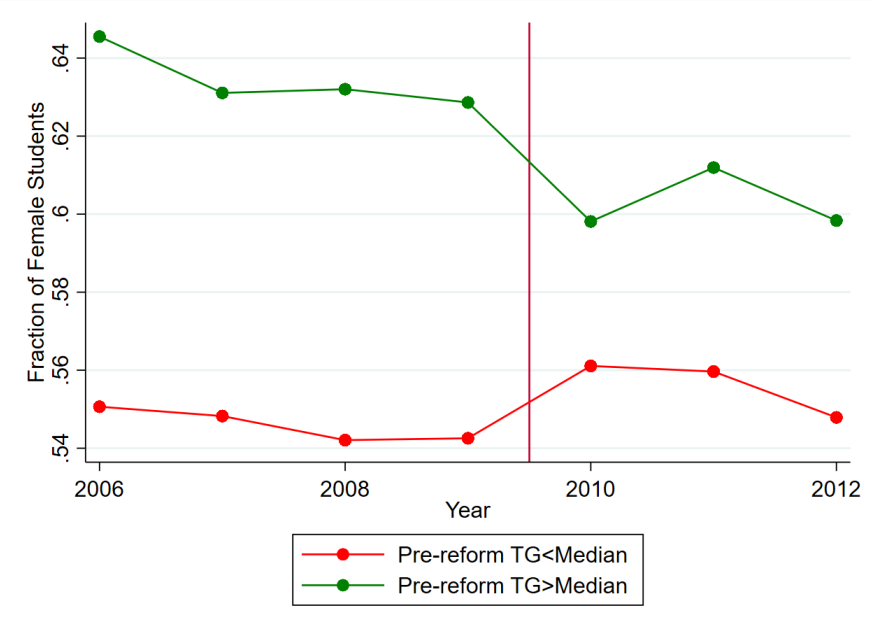


Table 3 reports differences-in-differences estimates with a continuous treatment measure. In a similar vein, the results show that the reform significantly decreased the percentage of female students in the most selective programmes. Compared to a program in the 25th percentile of selectivity, the percentage of female students in a program in the 75th percentile of selectivity declines by around 1.5 pp.

Table 3: Enrolment in selective programs

Dependent variable: fraction of female students

	(1)	(2)
Post × Pre-Reform T.Grade	-0.0146***	-0.0167**
	(0.00418)	(0.00775)
Faculty-Field-Municipality FE	✓	✓
Year FE	✓	✓
Faculty-Field-Municipality trends		✓
Mean Dep. Var	0.588	0.588
N	1018	1018

Standard errors clustered at the panel unit faculty-field-municipality in parentheses.
Estimates weighted by the number of enrolled students.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We also report estimates from individual-level regressions for the effect of the reform on the threshold grade of the program of enrolment in table A8 in the Appendix. In the first two columns, the dependent variable is the average pre-treatment threshold grade of the program of enrolment (again, at the level of university × faculty × municipality × field of study), which keeps the selectivity measure constant. In the third and fourth columns, the dependent variable is the average pre or post-treatment threshold grade of the program of enrolment (again, at the level of university × faculty × municipality × field of study). The results show a negative effect of the reform on the threshold grade of the program of enrolment when keeping its selectivity measure constant (columns 1 and 2), and an even larger effect on actual post-reform threshold grades (columns 3 and 4), which could be due to a decrease in the threshold grades of programs with a large percentage of female students.

Hence, overall, gender differences in admission grades due to the reform translate into significant changes in the colleges' allocation. The magnitude of the effect is again comparable to the date of birth effect on threshold grades in our sample (i.e., the effect of being born in January rather than in December); and to around 15% of the parental college education gradient in threshold grades, as reported by tables A9 and A10 in the Appendix.

In Appendix B, we further study whether this effect on the selectivity of the program of enrolment is associated with changes in career prospects. This is interesting because threshold grades and wages are only positively correlated within field of study, and because female students tend to sort into fields and academic programmes with worse career (wage, employment) prospects. Hence, the effect will depend on whether students very much substitute their most preferred programmes for less selective programmes within the same field. Using a survey of pre-treatment college graduates to compute expected wages and employment by academic program, we estimate that the effect on the college allocation comes along with an increase of 2% in the expected gender wage gap four years after graduation (on top of a 20% wage gap) and with a small but significant effect on expected employment as well.

5 Match quality

The reform has a significant effect on gender differences in admission grades and on the allocation of students to academic programmes, because of gender differences in high school vs. high-stakes performance. However, an open and very policy-relevant question is whether there is a trade-off between gender inequality and the quality of the match between students to college. To address this question, we study how gender differences in high-stakes performance in college admissions relate to college performance skills.

To this aim, we proceed in two steps. First, using machine learning techniques, we identify the types of students who are most likely to benefit from the reform (i.e., predicted winners and losers), based on a large set of detailed pre-determined student characteristics. Then, focusing on pre-treatment cohorts, we compare the college performance of students with the same admission grade and enrolled in the same program, college and (pre-reform) cohort, based on whether they are predicted to be winners or losers from the reform. The aim is to understand whether students who pre-reform were doing better in college (beyond

what one would expect given their admission grades) are those most likely to gain from the reform and whether there are gender differences.

More precisely, in our first step, we estimate a prediction model for the heterogeneous effect of the reform across students, based on individual pre-determined covariates. An important concern about this type of prediction exercise is over-fitting. Over-fitting is a concern because, for instance, OLS coefficient estimates of the heterogeneous effects of the reform maximize the in-sample fit. Instead, Machine Learning methods, such as Lasso (least absolute shrinkage and selection operator), are estimated to maximize their out-of-sample predictive power, although the coefficient estimates cannot be interpreted as indicating any meaningful structure (Mullainathan and Spiess, 2017). Given the large set of covariates at hand and that we are interested in predicting the effect of the reform on admission scores, this is a suitable approach.

Lasso regressions are a form of penalized regression, with a penalty for each non-zero coefficient, that overcome over-fitting via cross-validation: slicing the sample into different parts, a training sample and a testing sample, and delivering estimates that maximize the predictive power of the training samples on the testing samples (Athey and Imbens, 2019).[12] Lasso's $\hat{\beta}$ are the solution to: $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i \boldsymbol{\beta}'\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$, where $\lambda > 0$ is the Lasso penalty parameter that is chosen through cross-validation to maximize the out-of-sample performance of the training sample on the testing sample and p is the number of covariates.

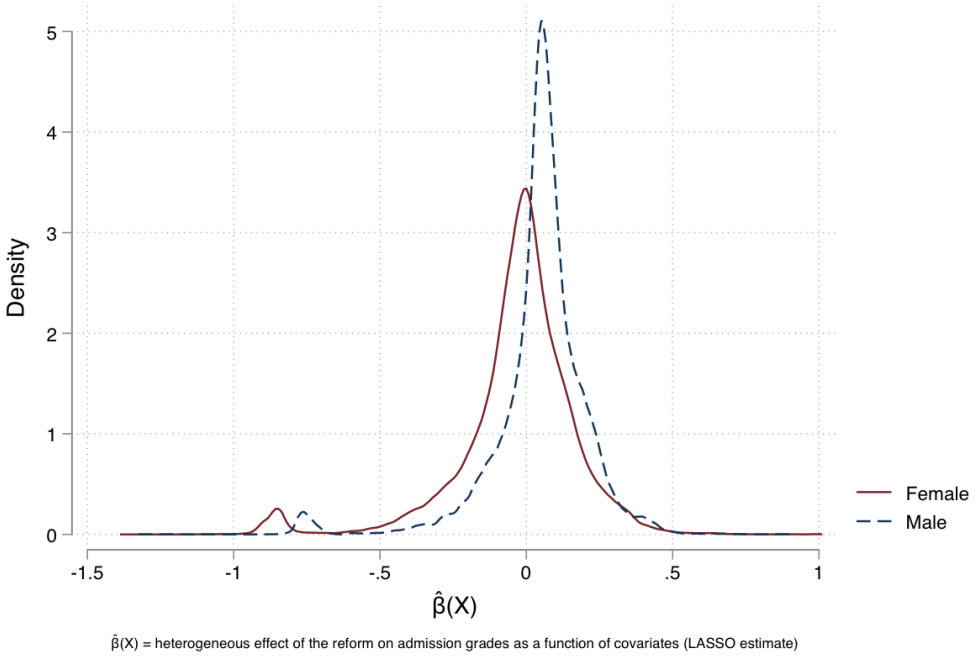
In this case, we fit the Lasso models separately for the pre and post-reform periods, to obtain $\hat{\beta}(X)$, the predicted gain of the reform as a function of covariates X, where X is a vector of parental and maternal education and occupation dummies, postal code, high school, and month of birth dummies, all of them interacted with a gender indicator.

[12] In this case, the pre-treatment sample is sliced into ten different parts, as suggested by Kuhn and Johnson (2013) and Kohavi (1995).

$$\hat{\beta}(X) = \widehat{Admission\ Grade}(X)^{Post} - \widehat{Admission\ Grade}(X)^{Pre}$$

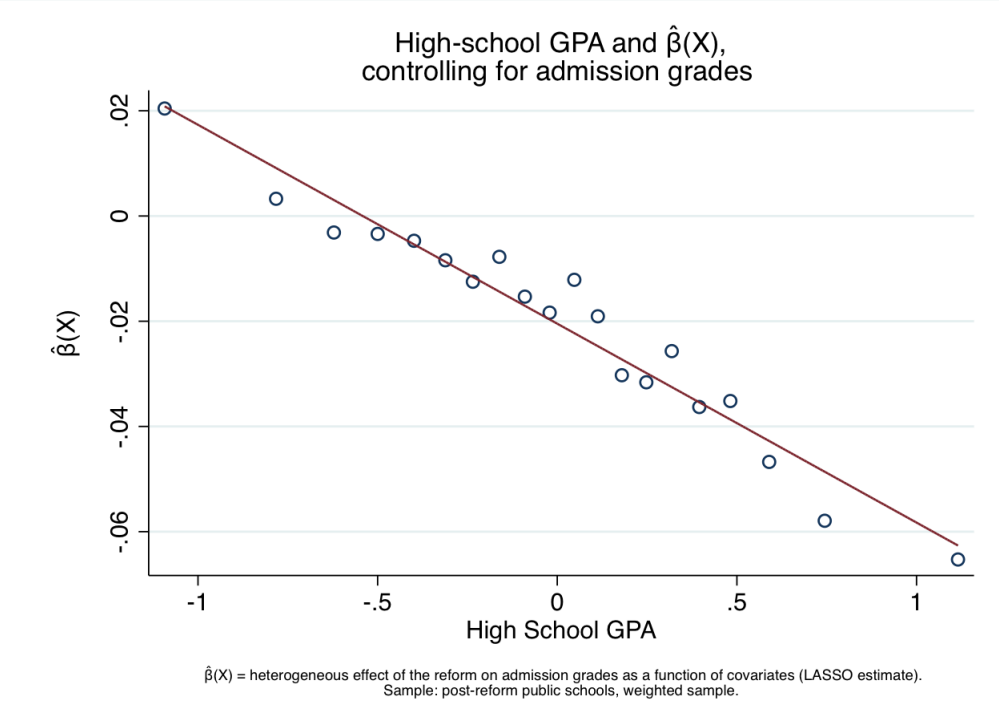
Figure 6 plots the distribution of predicted effects of the reform $\hat{\beta}(X)$ by gender, where on average $\hat{\beta}(X, Female)$ = -0.03 and $\hat{\beta}(X, Male) = 0.045$ (note that this is not symmetric because there are 60% of female students).

Figure 6: Distribution of expected gains from the reform



We further validate this measure by looking at its correlation with high school performance for the post-treatment cohorts (sample of public schools). We would expect that students predicted to gain from the reform are those doing relatively worse in high school. Figure 7 shows that within students with similar admission grades, those predicted to benefit from the reform are indeed those with worse high school grades (relative to their high-stakes performance).

Figure 7



Once we have obtained an individual-level measure of the predicted effects of the reform ($\hat{\beta}(X)$), the second step is to relate it to college performance skills. The data on college performance by pre-treatment students enrolled comes from UB (Universitat de Barcelona), UAB (Universitat Autònoma de Barcelona) and UPF (Universitat Pompeu Fabra), which enrol around 61% of students in Catalan Public Universities.[13] We merge these data with the college applications data (i.e., the *selectivitat* dataset).[14] European undergraduate degrees are structured into subjects. Subjects have a number of credits (usually around 6 per subject, where a credit represents a certain amount of coursework time, which is standardized across all EU countries), and completion of an undergraduate degree typically requires passing 180 credits.

[13] UB: 29.5%, UAB: 22.5%, UPF: 9%.

[14] We match the main college applications dataset with the college performance datasets, which are provided by universities, based on detailed demographics, matching 72.3% of students.

For UB, we observe, for all students in the 2006 to 2009 enrolling cohorts, for every year they are enrolled, the number of subjects (credits) they enrol, the number of credits they pass, and the average GPA in the passed subjects. For UAB, for all students in the 2006 to 2009 enrolling cohorts, the number of credits they enrol and pass, for the academic years 2008 to 2012.[15] For UPF, for all students in the 2006 to 2009 enrolling cohorts, the yearly number of credits they enrol and pass. Hence, we use as the main measure of college performance the fraction of credits that a student passes out of the credits she enrols during her time in college. We also present results with students' college GPA (average GPA in the completed subjects, unconditional on graduation, available for UB) in the Appendix.

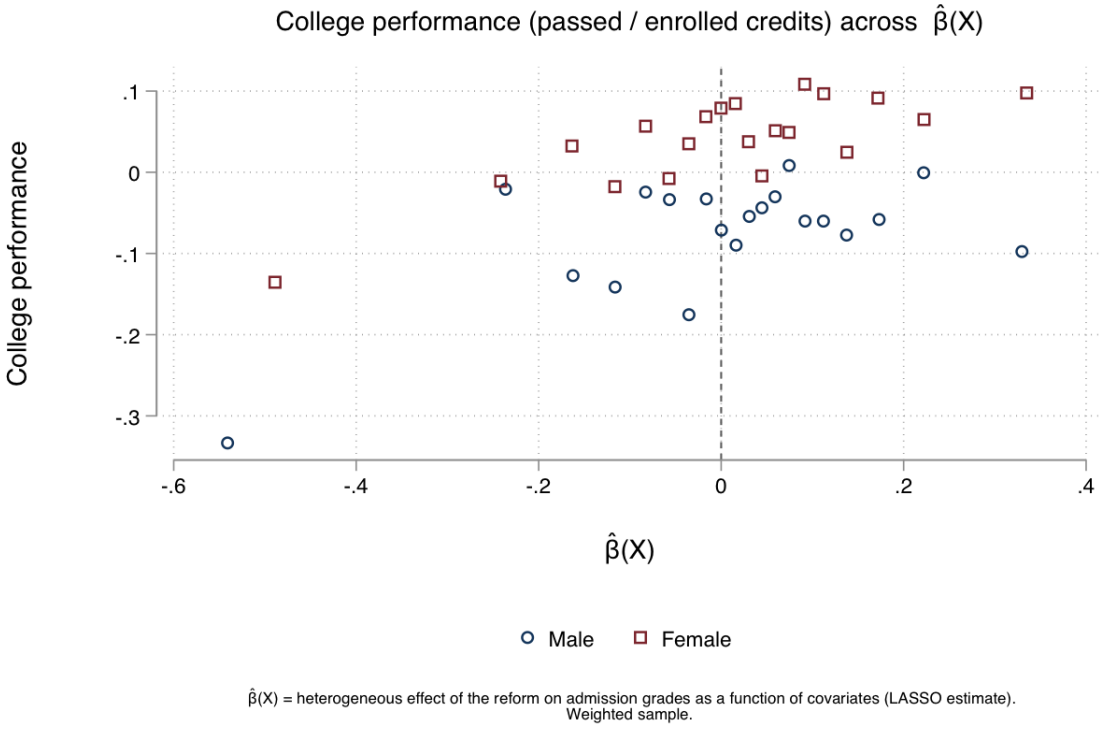
We measure college performance with the residuals of a regression of the raw measure of college performance (fraction of credits passed out of credits enrolled and GPA, both standardized by cohort by academic programme) on admission grades: $\widetilde{CP}_i = CP_i - \widehat{CP}_i$. We weight the observations so that the college performance sample matches the population in admission grades by gender and cohort, using entropy balancing (Hainmueller, 2012), but report unweighted results in the Appendix as well.

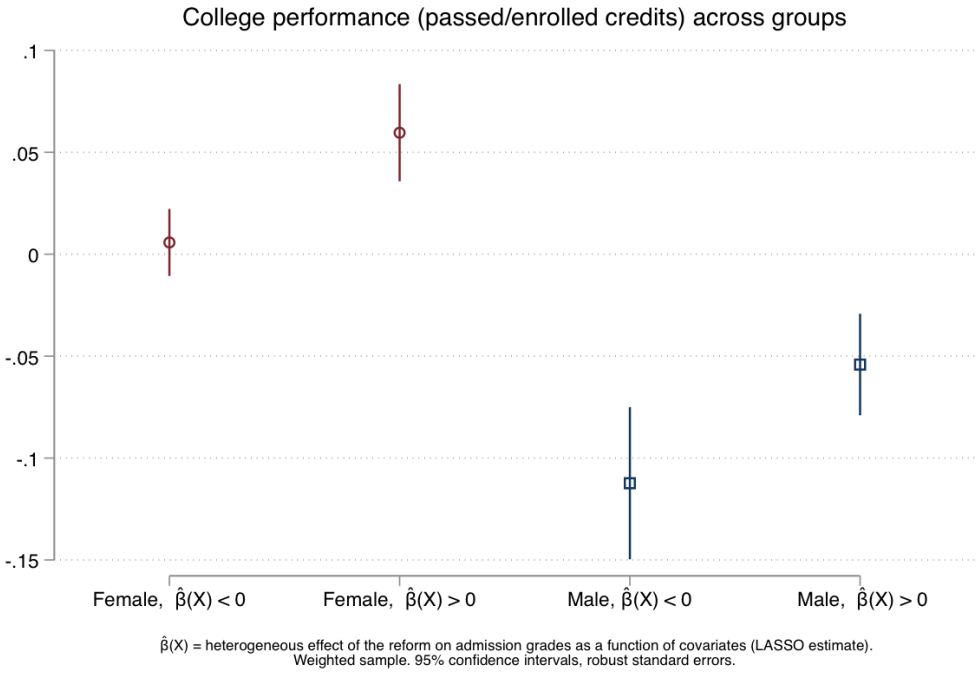
The top panel in figure 8 displays college performance of pre-reform students $\widetilde{CP}_i$ over $\hat{\beta}(X)$. It shows three interesting patterns. First, a positive unconditional and within-gender correlation between college performance and expected gains from the reform, which suggests that high-stakes performance skills correlate positively with college performance skills.

Second, that female students perform better in college than male students with the same expected gains from the reform. This can be seen more precisely in the bottom panel of the figure, which splits college performance by gender and by expected winners and losers from the reform. The college performance of females with $\hat{\beta}(X) < 0$ is larger than the performance of males with $\hat{\beta}(X) < 0$, and the same for expected winners ($\hat{\beta}(X) > 0$).

[15] This means that the students from the 2006 and 2007 cohorts are slightly positively selected because we observe them conditional on enrolment in their second or third year. However, dropping those cohorts does not change the results.

Figure 8: College performance and expected gains from the reform





Third, that females predicted to lose from the reform perform better in college than males predicted to benefit from it. Again, this is shown more precisely in the bottom panel. The college performance of females expected to lose (i.e., those with $\hat{\beta}(X) < 0$) is larger than the performance of males expected to win from the reform (i.e. those with $\hat{\beta}(X) > 0$).

The results show that although within gender, high-stakes performance skills positively correlate with college performance skills, the gender difference in high-stakes performance is negatively related to college performance skills. This means that the gender differences in admission scores induced by the reform may go against policy-makers objective functions aiming at selecting students based on their college performance potential.

Figure A1 in the Appendix displays the same figure for GPA rather than the fraction of credits passed, and figure A2 shows unweighted results, with a very similar pattern.

We also report results disaggregated by field of study in figure A3 the Appendix.[16] The figure indicates that the results are largely driven by social science students. Investigating the mechanisms driving these heterogeneous effects is left for future research.

[16]The field composition in our sample vs. the population is the following: Arts-Humanities (14% in sample vs. 10% population), Science (17% vs. 9%), Social Sciences (51% vs. 43%), Health Science (12% vs. 15%), and Engineering (5% vs. 23%).

6 Conclusions

Our results show that a college choice mechanism giving more weight to high-stakes exams for admissions has important cross-gender effects. In general, female students tend to outperform male students in high school, but gender differences in high-stakes performance are much smaller. Using administrative data on the population of applicants to Catalan universities, we find a significant negative effect on female college admission scores of a reform that increased the weight of the comprehensive high-stakes exam at the end of high school for college admissions. A very substantial part of this effect is due to a re-weighting of the baseline high school vs. high-stakes performance differences, but the overall effect is slightly larger, suggesting that the effect of the reform is amplified by behavioral responses.

We further document that these effects have important consequences for the allocation of students to college. Most gender differences in admission scores induced by the reform happen at the top of the ability distribution, and as a result, the reform does not affect college enrolment. Nevertheless, the percentage of female students in the most selective degrees decreases significantly, and this comes along with a decline in their career prospects, widening expected gender gaps in the labour market.

Finally, we study whether the reform entails a trade-off between gender inequality and match quality. We find that within gender, good college performers tend to benefit from the reform. However, the results show that female students expected to lose from the reform are better college performers than male students expected to gain from the reform. Hence, the results show that gender differences in high-stake exam performance are not positively related to determinants of college performance (if anything, these are negatively related). This is an important result for policy-makers designing college admission policies aiming at maximizing college performance potential in admissions while also taking into account gender differences in performance in different settings.

References

Aparicio-Fenoll, A. (2016). Returns to education and educational outcomes: The case of the Spanish housing boom. *Journal of Human Capital*, **10** (2), 235–265.

Arenas, A., Calsamiglia, C. and Loviglio, A. (2021). What is at stake without high-stakes exams? students' evaluation and admission to college at the time of covid-19. *Economics of Education Review*, **83**, 102143.

— and Malgouyres, C. (2018). Countercyclical school attainment and intergenerational mobility. *Labour Economics*, **53**, 97–111.

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

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.

Balafoutas, L. and Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, **335** (6068), 579–582.

Cai, X., Lu, Y., Pan, J. and Zhong, S. (2018). Gender Gap under Pressure: Evidence from China's National College Entrance Examination. *The Review of Economics and Statistics*.

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

Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, **69** (1), 9–15.

Gneezy, U., Niederle, M. and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, **118** (3), 1049–1074.

— and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, **94** (2), 377–381.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, pp. 25–46.

Iriberri, N. and Rey-Biel, P. (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal*, **129** (620), 1863–1893.

Jurajda, Š. and Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to Czech universities. *American Economic Review*, **101** (3), 514–18.

Kaufmann, K. M., Messner, M. and Solis, A. (2021). *Elite Higher Education, the Marriage Market and the Intergenerational Transmission of Human Capital*. CRC TR 224 Discussion Paper Series CRC TR 224, University of Bonn and University of Mannheim, Germany.

Kirkebøen, L., Leuven, E. and Mogstad, M. (2021). *College as a marriage market*. Tech. rep., National Bureau of Economic Research.

Kirkebøen, L. J., Leuven, E. and Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, **131** (3), 1057–1111.

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.

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

MacLeod, W. B., Riehl, E., Saavedra, J. E. and Urquiola, M. (2017). The big sort: College reputation and labor market outcomes. *American Economic Journal: Applied Economics*, **9** (3), 223–61.

Montolio, D. and Taberner, P. A. (2018). Gender differences under test pressure and their impact on academic performance: a quasi-experimental design. *IEB Working Paper 2018/21*.

Morin, L.-P. (2015). Do men and women respond differently to competition? Evidence from a major education reform. *Journal of Labor Economics*, **33** (2), 443–491.

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

Niederle, M. (2015). Gender. In *Handbook of Experimental Economics*, Princeton University Press.

—, Segal, C. and Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, **59** (1), 1–16.

— and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, **122** (3), 1067–1101.

— and — (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, **24** (2), 129–44.

— and — (2011). Gender and competition. *Annu. Rev. Econ.*, **3** (1), 601–630.

Ors, E., Palomino, F. and Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, **31** (3), 443–499.

Saygin, P. O. (2018). Gender bias in standardized tests: evidence from a centralized college admissions system. *Empirical Economics*, pp. 1–29.

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** (10), 2916–2948.

UNESCO (2017). Global education monitoring report.

Gender differences in high-stakes performance and college admission policies

Appendix

Table A1: Heterogeneity in field subjects' weights

	(1) 1(Taking all exams)	(2) Average weight
Female	0.00631*	0.0663
	(0.00343)	(0.0514)
Year FE	✓	✓
Mean Dep. Var	0.585	13.69
N	84677	84677

Sample: post-reform. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Avg. weight=$\frac{W_B+W_C}{2}$, with $or = 0$ if the exam is not taken.

Table A2: Date of birth effect

Dependent variable: admission grade

	(1)	(2)
Born in January	0.0727***	0.0724***
	(0.0115)	(0.0115)
Year FE		✓
Mean Dep. Var	-0.00120	-0.00120
N	30255	30255

Sample: born in January or December.

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Parental education gradient in admission grades

Dependent variable: admission grade

	(1)	(2)	(3)	(4)
Both college educated	0.461***			
	(0.00557)			
At least one college educated		0.388***		
		(0.00464)		
Mother college educated			0.407***	
			(0.00491)	
Father college educated				0.384***
				(0.00494)
Year FE	✓	✓	✓	✓
Mean Dep. Var	7.08e-08	7.08e-08	7.08e-08	7.08e-08
Mean Indep. Var	0.233	0.455	0.347	0.341
N	183451	183451	183451	183451

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4: Robustness Admission Grades

	(1)	(2)	(3)	(4)
Female × Post 2009	-0.0506***	-0.0608***	-0.0661***	-0.0533***
	(0.00952)	(0.00872)	(0.00761)	(0.00712)
Female	0.142***	0.109***	0.142***	0.108***
	(0.00645)	(0.00600)	(0.00645)	(0.00596)
Female	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Post × Subject weights			✓	✓
Controls		✓		✓
Mean Dep. Var	-1.57e-08	-0.00244	7.08e-08	-0.00240
Subject weights	19	19	Baseline (enrolment)	Baseline (enrolment)
N	183451	182259	183451	182259

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Controls: Mother and father education and occupation, year and month of birth, nationality, high school, postal code.

Table A5: Re-weighting effects of the reform

	Admission Grade		$\Delta^{\text{Admission Grade}}_{\text{Admission Grade, based on pre-treatment formula}}$	
	(1)	(2)	(3)	(4)
Female × Post 2009	-0.0641***	-0.0621***	-0.0475***	-0.0452***
	(0.0165)	(0.0152)	(0.00256)	(0.00254)
Female	0.141***	0.114***		
	(0.0106)	(0.00985)		
Year FE	✓	✓	✓	✓
Controls		✓		✓
Mean Dep. Var	-0.000151	-0.0000259	-0.000338	-0.000303
N	70228	70067	70228	70067

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Sample: public schools, weighted to match admission scores by gender/year in the population via entropy balancing. Pre-treatment formula: pre-treatment weights for high school vs. high stakes and for core vs. field subjects within the high stakes GPA.
Controls: Mother and father education and occupation, year and month of birth, nationality, high school, postal code.

Table A6: Re-weighting effects of the reform, public schools, unweighted.

	Admission Grade		$\Delta^{\text{Admission Grade}}_{\text{Admission Grade, based on pre-treatment formula}}$	
	(1)	(2)	(3)	(4)
Female × Post 2009	-0.0354**	-0.0389***	-0.0482***	-0.0458***
	(0.0149)	(0.0140)	(0.00264)	(0.00261)
Female	0.103***	0.0839***		
	(0.00978)	(0.00916)		
Year FE	✓	✓	✓	✓
Controls		✓		✓
Mean Dep. Var	-0.139	-0.138	-0.000646	-0.000609
N	70228	70067	70228	70067

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Sample: public schools. Pre-treatment formula: pre-treatment weights for high school vs. high stakes and for core vs. field subjects within the high stakes GPA. Controls: Mother and father education and occupation, year and month of birth, nationality, high school, postal code.

Table A7: Within faculty-field-municipality autocorrelation in outcomes

	(1) Thresh. Grade	(2) #Enrolled Students	(3) %Female Enrolled
Lagged T.Grade	0.949***		
	(0.0167)		
Lagged #Enrolled Students		0.987***	
		(0.00750)	
Lagged %Female Enrolled			0.951***
			(0.00840)
Mean Dep. Var	-0.873	341.6	0.584
N	874	874	874

Standard errors clustered by the panel unit faculty-field-municipality.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Estimates weighted by the number of enrolled students.

Table A8: Threshold grades, program of enrolment

	Thresh G. (pre-treat).		Thresh G. (actual).	
	(1)	(2)	(3)	(4)
Female	0.168***		0.168***	
	(0.00498)		(0.00498)	
Female × Post 2009	-0.0262***	-0.0537***	-0.0720***	-0.107***
	(0.00769)	(0.0155)	(0.00795)	(0.0158)
Year FE	✓	✓	✓	✓
Gender-specific trends		✓		✓
Mean Dep. Var	-0.833	-0.833	-0.872	-0.872
N	166372	166372	170082	170082

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Thresh G. (pre-treat): based on pre-reform avg. values by faculty-field-municipality.

Thresh G. (actual): based on pre and post-reform averages by faculty-field-municipality.

Table A9: Date of birth effect

Dependent variable: threshold grade		
	(1)	(2)
Born in January	0.0585***	0.0591***
	(0.0112)	(0.0112)
Year FE		✓
Mean Dep. Var	-0.870	-0.870
N	28063	28063

Sample: born in January or December.

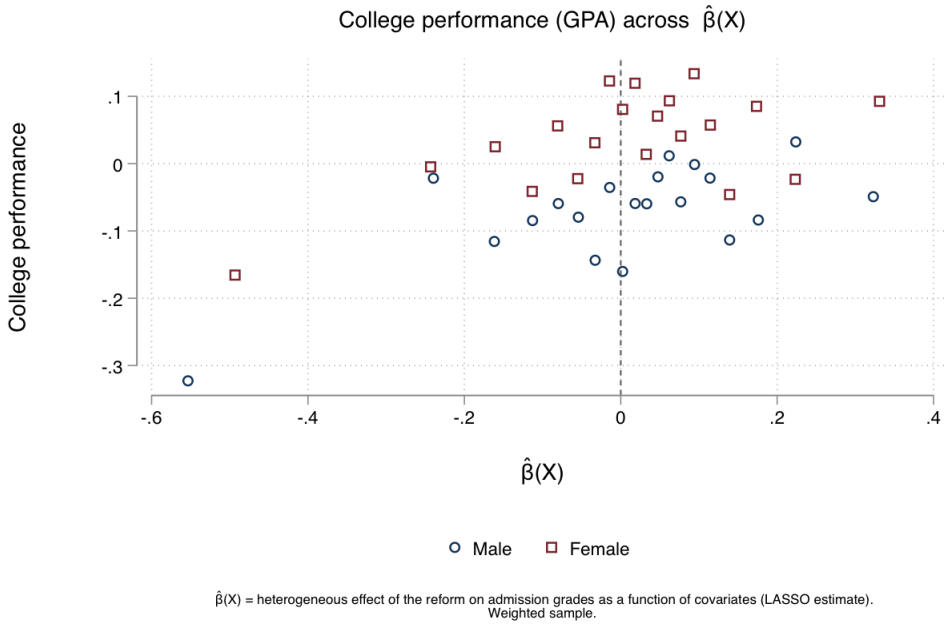
Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A10: Parental education gradient in threshold grades

Dependent variable: threshold grade				
	(1)	(2)	(3)	(4)
Both college educated	0.390***			
	(0.00584)			
At least one college educated		0.318***		
		(0.00457)		
Mother college educated			0.330***	
			(0.00496)	
Father college educated				0.328***
				(0.00499)
Year FE	✓	✓	✓	✓
Mean Dep. Var	-0.872	-0.872	-0.872	-0.872
Mean Indep. Var	0.230	0.452	0.344	0.337
N	170082	170082	170082	170082

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure A1: College performance and expected gains from the reform



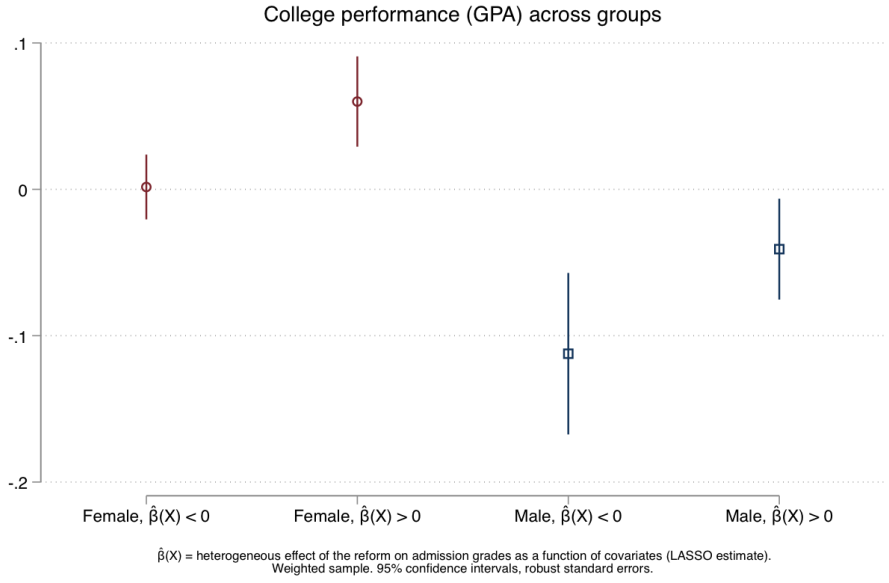


Figure A2: College performance and expected gains from the reform (unweighted)

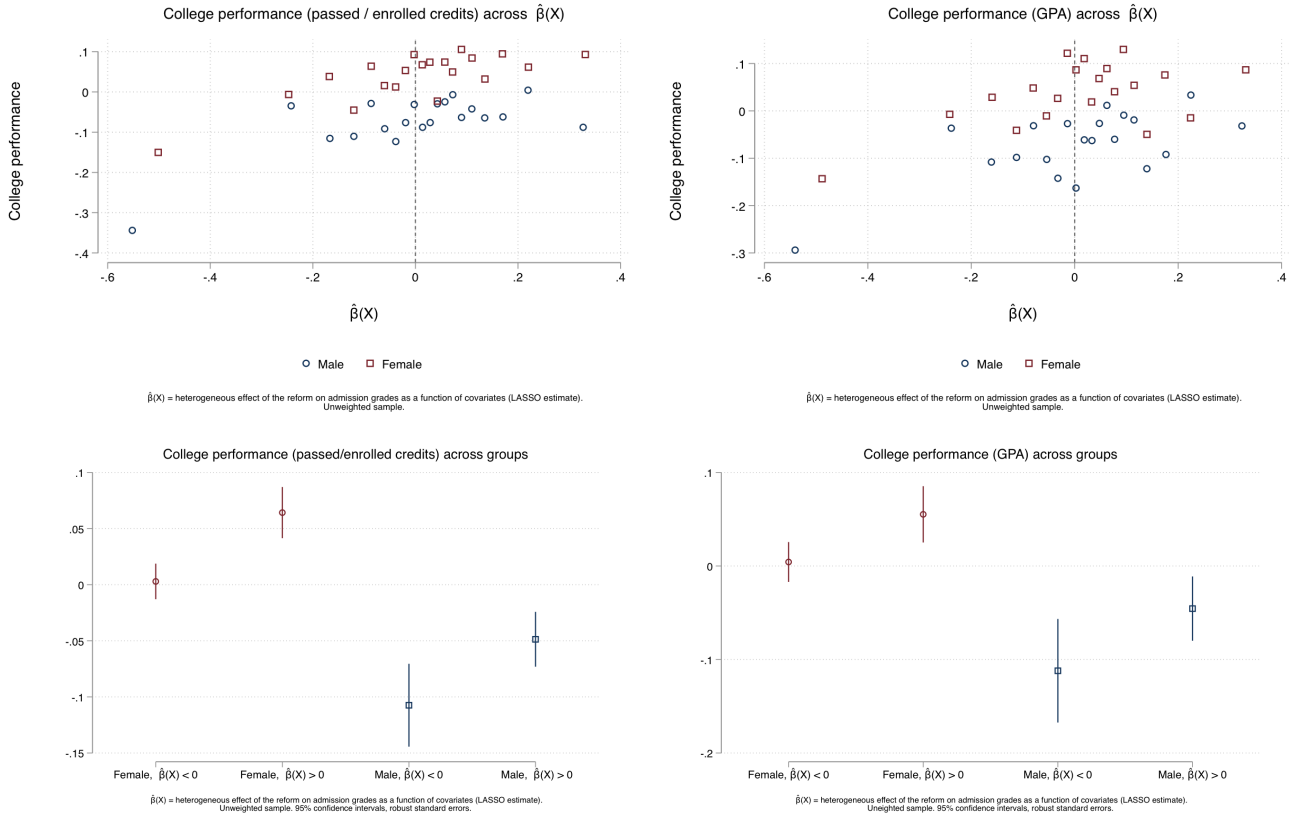
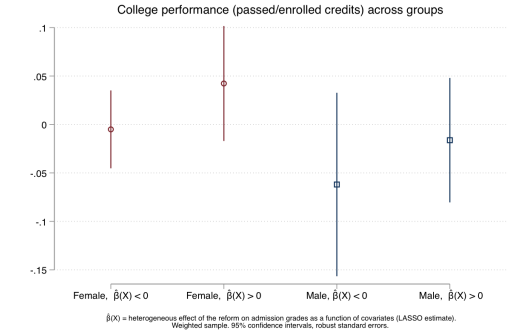
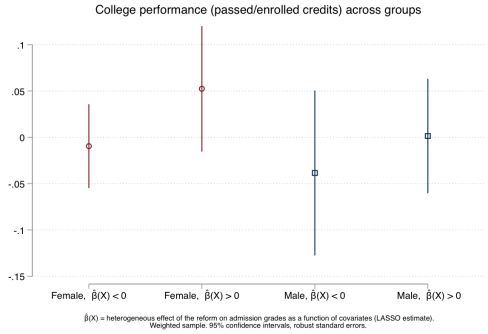


Figure A3: College performance and expected gains from the reform across fields

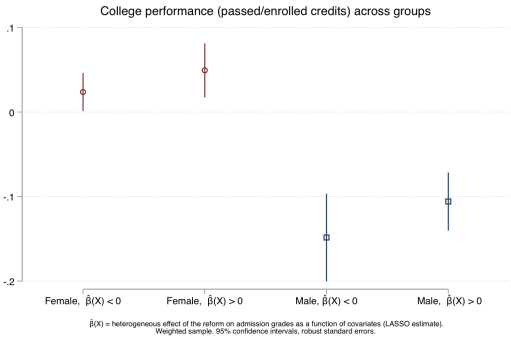
(a) Arts and humanities



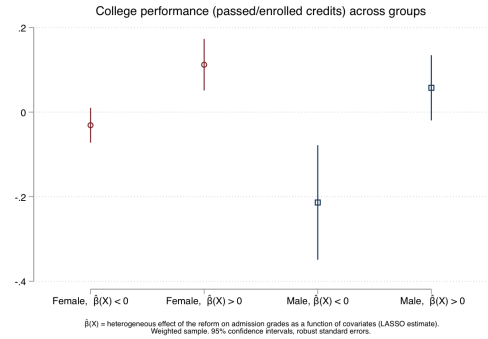
(b) Science



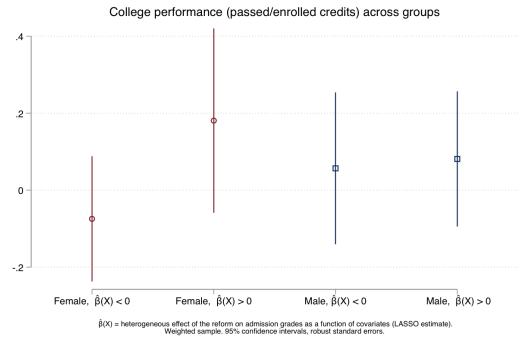
(c) Social Sciences



(d) Health Sciences



(e) Technical/engineering degrees



Appendix B: Career Prospects

In this Appendix, we study how the change in students' allocation changes students' career prospects. To this aim, we use survey data on a sample of pre-treatment college graduates from Catalan universities, with information on labour market outcomes four years after graduation. For every student, we do not observe the exact academic programme, but area (i.e., field of study) indicators, and the university. There are enough area indicators (more than 50) which combined with the university of enrolment make it a meaningful measure, despite some measurement error. Figure A4 displays the social science classification to illustrate the level of detail of the field of study that we observe.[17]

Figure A4: Degree classification example: social science

Agència per a la Qualitat del Sistema Universitari de Catalunya

Catàleg de titulacions

CODI	ENSENYAMENT	SUBÀMBIT DETALLAT (1r NIVELL)		SUBÀMBIT AMPLIAT (2n NIVELL)		ÀMBIT	
2010101	Economia	20101	Economia	201	Economia, Empresa i Turisme	2	Ciències socials i jurídiques
2010102	Comptabilitat i finances						
2010201	Administració i direcció d'empreses	20102	Administració d'Empreses				
2010202	Màrqueting i investigació de mercats						
2010203	Ciències empresarials						
2010204	Estudis internacionals d'economia i empresa						
2010301	Turisme	20103	Turisme				
2020101	Dret	20201	Dret	202	Dret, laboral i polítiques		
2020201	Criminologia	20202	Laboral				
2020202	Relacions laborals						
2020203	Ciències del treball						
2020204	Prevenció i seguretat integral						
2020301	Gestió i administració pública	20203	Polítiques				
2020302	Ciències polítiques i de l'administració						
2020401	Sociologia	20204	Sociologia, Geografia				
2020402	Antropologia social i cultural						
2020403	Geografia						
2030101	Comunicació audiovisual	20301	Comunicació	203	Comunicació i Documentació		
2030102	Periodisme						
2030103	Publicitat i relacions públiques						
2030201	Informació i documentació	20302	Documentació				
2040101	Educació infantil	20401	Mestres	204	Educació		
2040102	Educació primària						
2040103	Mestre. Especialitat d'Educació Especial						
2040104	Mestre. Especialitat d'Educació Física						
2040105	Mestre. Especialitat d'Educació Musical						
2040106	Mestre. Especialitat de Llengua Estrangera						
2040201	Pedagogia	20402	Pedagogia i Psicopedagogia				
2040202	Psicopedagogia						
2040203	Formació de professorat						
2050101	Treball social	20501	Treball i educació social	205	Intervenció Social		
2050102	Educació social						
2050201	Psicologia social i organitzacional	20502	Psicologia				
2100101	Titulacions Mixtes	21001	Titulacions Mixtes	210	Titulacions Mixtes		

[17] We do not observe academic programmes (columns 1 and 2), but sub-sub-area indicators (columns 3 and 4).

In this representative survey, although girls outperform boys in educational attainment, females earn 23% less than males on average (9.3% less when accounting for field of study), as reported by table A11.[18]

Table A11: Gender wage gap of college graduates

Dependent variable: ln(wage)

	(1)	(2)	(3)
Female	-0.232***	-0.0923***	-0.0939***
	(0.00915)	(0.00930)	(0.00930)
Cohort FE	✓	✓	✓
Field of study FE		✓	✓
University-by-field of study FE			✓
Mean Dep. Var	9.662	9.662	9.662
N	11729	11729	11724

Field of study: sub-sub-area.

Sample of 2006-2009 cohorts, 4 years after graduation.

All regressions control for year of survey FE

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

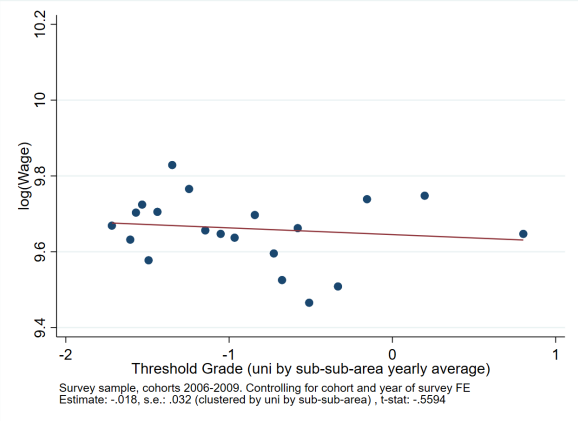
The effect of the reform on gender differences in career prospects is not straightforward, for two reasons. First, because of the relationship between program selectivity and career prospects. The left panel of figure A5 shows the unconditional correlation that there is almost no unconditional correlation between college selectivity and earnings. However, the right panel shows that a positive correlation exists within field of study. This is because selectivity is determined by capacity constraints and demand, and some high-paying technical degrees have good career prospects but low capacity constraints and low demand; while some degrees in the humanities have worse career prospects but high capacity constraints and demand. However, within field of study, where demand and capacity constraints are more homogeneous, the correlation is positive, as one would expect. Hence, the effect of the reform on career prospects will depend on the extent to which students' alternative to their most preferred option tends to be in the same field of study.

[18] We use the 2014 and 2017 waves of the survey, conducted by the Catalan Agency for the Quality of Universities (AQU), which include students from the 2006-2009 enrolling cohorts.

Second, the effect on gender differences in career prospects is not straightforward because females tend to sort into academic programmes with worse labour market prospects. Figure A6 displays wages against the gender composition of academic programmes. First, it shows that within academic programmes, females earn lower wages. Second, it also shows that programmes with a higher percentage of female students tend to pay less (for both males and females). This is important because the right panel of figure A6 shows that due to the reform, females enrol less in programmes with a higher pre-reform percentage of female students.

Figure A5: Threshold grades and wages

Unconditionally



Within field

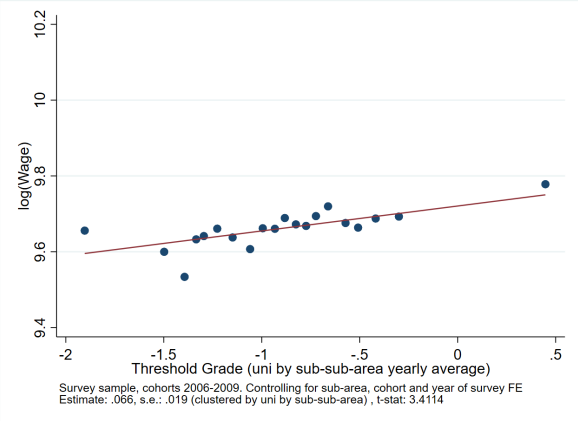
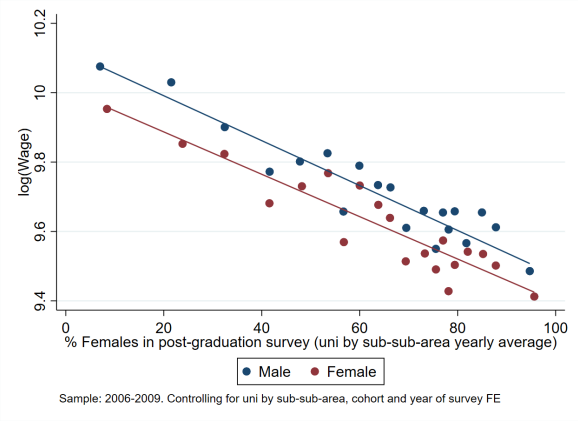
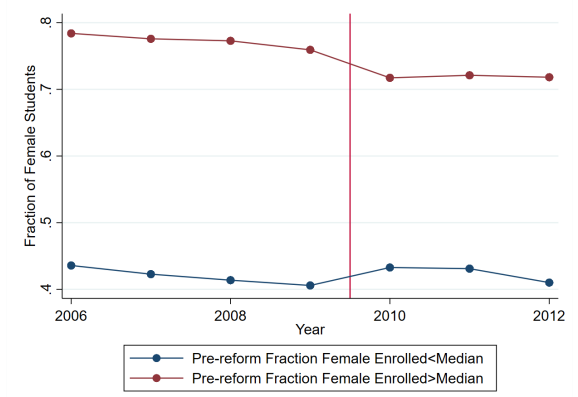


Figure A6

% of female students and wages



% female enrolment by pre-treatment female enrolment



To estimate the effect of the reform on career prospects by gender, we first estimate expected labour market outcomes by academic programme using the survey data, which includes cohorts enrolling into college between 2006 and 2009 (i.e., pre-reform cohorts):

$$Outcome_{it} = \delta Female_i + \alpha(Area \times Uni)_i + \beta Trend_t + \gamma Trend_t \times (Area \times Uni)_i + \epsilon_{it}$$

Where labour market outcomes of student i in enrolling cohort t are measured for the 2006-2009 enrolling cohorts (and survey FE have been partialled out), and where $Area \times Uni$ are dummies for study subarea (or sub-sub-area) by university.

In a 2nd step, we combine the predicted labour market outcomes from the previous regression with the college enrolment data from the Selectivitat dataset, and we estimate:

$$\widehat{Outcome}_{it} = \delta_t + \gamma Female_i + \beta Female_i \times Post_t + \epsilon_{it}$$

Figure A7 displays female by year coefficients, where the baseline year is 2009, the last pre-reform year. It shows that the gender gap in career prospects becomes larger after the reform. Table A12 reports point estimates, indicating an increase of around 2.5pp in the gap, on top of a 23pp pre-reform gap. Table A13 reports point estimates on the expected employment rate. Given the high employment rate among Catalan university graduates (around 87% according to the survey), the magnitude of the effect is smaller, but still significant. To benchmark the magnitude of these effects, it is interesting to compare them with the findings in Ebenstein *et al.* (2016) that pollution in matriculation exam days leads to lower test scores, resulting in a decline in post-secondary education and earnings. It turns out that the effect on female test scores and career prospects is similar in magnitude to the effect of one standard deviation in pollution exposure on the day of the exam. Tables A14 and A15 report placebo tests showing that the change in the post-treatment period is large and significant compared to any changes within the pre-treatment period.

Figure A7: Career prospects: predicted log(wages)

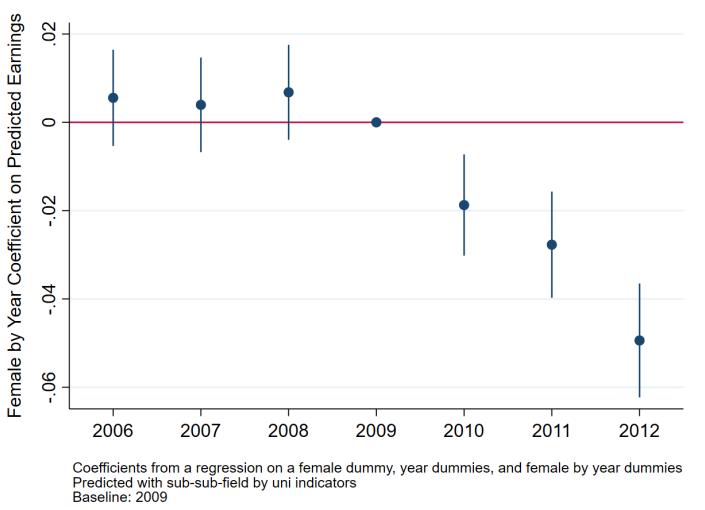


Table A12: Dependent Variable: Predicted log(wage)

	(1)	(2)	(3)	(4)
Female	-0.229***	-0.224***	-0.244***	-0.229***
	(0.00167)	(0.00172)	(0.00195)	(0.00203)
Female × Post 2009	-0.0212***	-0.0140***	-0.0361***	-0.0274***
	(0.00276)	(0.00282)	(0.00341)	(0.00354)
Year FE	✓	✓	✓	✓
Main Predictor	Sub-field × Uni	Sub-field × Uni × Female	Sub-sub-field × Uni	Sub-sub-field × Uni × Female
Mean Dep. Var	9.684	9.688	9.661	9.665
N	170082	170082	170082	170082

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A13: Dependent Variable: Predicted Employment Rate

	(1)	(2)	(3)	(4)
Female	-0.0109***	-0.00983***	-0.0148***	-0.0134***
	(0.000515)	(0.000600)	(0.000695)	(0.000787)
Female × Post 2009	-0.00340***	-0.00110	-0.00868***	-0.00551***
	(0.00104)	(0.00110)	(0.00169)	(0.00174)
Year FE	✓	✓	✓	✓
Main Predictor	Sub-field × Uni	Sub-field × Uni × Female	Sub-sub-field × Uni	Sub-sub-field × Uni × Female
Mean Dep. Var	0.873	0.872	0.874	0.872
N	170082	170082	170082	170082

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A14: Dependent Variable: Predicted log(wage)

	(1)	(2)	(3)	(4)
Female	-0.228***	-0.224***	-0.243***	-0.229***
	(0.00241)	(0.00249)	(0.00276)	(0.00288)
Female × Post 2009	-0.0203***	-0.0142***	-0.0353***	-0.0276***
	(0.00319)	(0.00326)	(0.00392)	(0.00407)
Female × Post 2007	-0.00181	0.000519	-0.00152	0.000392
	(0.00334)	(0.00344)	(0.00390)	(0.00406)
Year FE	✓	✓	✓	✓
Main Predictor	Sub-field × Uni	Sub-field × Uni × Female	Sub-sub-field × Uni	Sub-sub-field × Uni × Female
Mean Dep. Var	9.684	9.688	9.661	9.665
N	170082	170082	170082	170082

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A15: Dependent Variable: Predicted Employment Rate

	(1)	(2)	(3)	(4)
Female	-0.0117***	-0.0113***	-0.0144***	-0.0137***
	(0.000853)	(0.000986)	(0.000997)	(0.00116)
Female × Post 2009	-0.00408***	-0.00242**	-0.00835***	-0.00583***
	(0.00109)	(0.00117)	(0.00182)	(0.00189)
Female × Post 2007	0.00143	0.00277**	-0.000702	0.000686
	(0.00105)	(0.00122)	(0.00139)	(0.00158)
Year FE	✓	✓	✓	✓
Main Predictor	Sub-field × Uni	Sub-field × Uni × Female	Sub-sub-field × Uni	Sub-sub-field × Uni × Female
Mean Dep. Var	0.873	0.872	0.874	0.872
N	170082	170082	170082	170082

Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$