Missing Women in Research

This is a preliminary draft: please do not circulate without consent.

Aliénor Bisantis*, Yann Bramoullé, Roberta Ziparo†

January 2024

ABSTRACT.
Matching the universe of PhD graduations in France across all academic disciplines between 1988 and 2021 and Scopus bibliometric data, we estimate the number of missing women in academia both on the extensive - the probability of publishing - and the intensive margin - the number of missing publications. We explore how the number of missing women varies across time and disciplines. We also show how it correlates with initial gaps in PhD participation and the student-supervisor match characteristics. Preliminary findings show that female PhD students will be less likely to have at least one publication, and if they do, they will publish less in their careers. We observe heterogeneous trends across disciplines.

Keywords: Gender Inequality, Career Trajectory, Graduate Education
JEL: I23, J16, J24

*Presenter, Aix-Marseille University, CNRS, Aix-Marseille School of Economics; address: 5-9 Boulevard Maurice Bourdet, CS 50498 13205 Marseille Cedex 1; tel: +33 7 86 39 96 63; E-mail addresses: alienor.bisantis@univ-amu.fr

†Aix-Marseille University, CNRS, Aix-Marseille School of Economics. E-mail addresses: yann.bramoulle@univ-amu.fr, roberta.ziparo@univ-amu.fr

1 Introduction

The proportion of women researchers varies a lot across disciplines according to the *UNESCO Science Report*; patterns are similar across countries of the world: in sciences, women are poorly represented in engineering, technology, physical, mathematical, and chemical sciences but are more present in biology. Among the humanities, we observe a larger proportion of women, except in philosophy. Women are still under-represented in research in all disciplines. In France, the share of female researchers is below the European average and does not exceed 30%. While women are on average more educated than men[1], women are less likely to enroll in a PhD and even less likely to pursue in academia.

The purpose of this study is to estimate the number of missing women in academia, both on the extensive - the probability of publishing - and the intensive margin - the number of missing publications. To do so, we match data on the universe of PhD graduations in France across all disciplines between 1988 and 2021 and Scopus bibliometric data. We explore how the number of missing women varies across time and disciplines. We also show how it correlates with initial gaps in PhD participation and the student-supervisor match characteristics.

First, we want to document gender differences in PhD graduation across all research disciplines. Second, we want to study whether participation in academia and research productivity differ across gender and disciplines exploring PhD student and supervisor characteristics. To answer these questions, we build a novel database containing information about 336,390 theses defended in France. For each thesis, we have information about the discipline of study, the defense year, and the university affiliation. We use standard methods based on first names to identify the gender of the PhD students and supervisors. Then, we use Scopus bibliometric data to retrieve publications and explore academic participation and productivity among both PhD students and supervisors.

We study the evolution of gender graduation over 20 years. The data set shows a general tendency to the progression over the years of the share of women in PhD as well as the share of women supervisors; the progression remains very weak in *sciences, technology, engineering, and mathematics* (STEM), the share of female PhD students has remained between 20 and 30% since 1988. In *humanities* and *biological and earth science*, the female share was around 40% in the 1990s, rose above 50% around 2000 (except in philosophy), and stagnates around 50% in biological and earth sciences today. Economics and mathematical science are the only disciplines in which the share has stagnated and even decreased since 2010.

In the second part of our study, we look at the extensive and intensive margin of differences in research productivity across genders. Preliminary results indicate gender differences in the probability of publishing at least one paper: on average, female PhD students are less likely to publish than male PhD students, and it seems that the effect becomes more pronounced with career time. We find that the supervisor's research productivity reduces the negative effect of being a female PhD student. We find very heterogeneous results across academic disciplines.

[1] In France, in 2017, 55% of students in higher education were women.

Our study contributes to a better understanding of gender gaps in academia. The literature has shown that in most academic disciplines, men are more productive than women in research: in economics (Ginther and Kahn (2004); Barbezat (2006); McDowell et al. (2006); Sarsons (2017); Ductor (2023); Conley et al. (2016)), in science (Patsali et al. (2021); Stephan and Levin (1997)), in medicine (Rachid, 2021). Our data allows us to focus on less-studied academic disciplines like humanities, and make direct comparisons between the various disciplines. We study researchers at an early career stage: once they finish their PhD and whether or not they continue to publish. Some studies look at early career researcher and the effect of the student-supervisor match: supervisor's experience Campbell (2003), age (Stephan and Levin, 1997), research independence with the PhD student (Patsali et al., 2021), and gender (Tenenbaum et al. (2001); Pezzoni (2016); Gaule and Piacentini (2018)). Our contribution lies first in analyzing the characteristics of PhD students who graduated from French universities and second in comparing those with similar circumstances to identify gender-based differences in academic participation and productivity. We provide the first comprehensive analysis of gender differences in PhD graduation and academic output, based on the universe of PhD graduations in France between 1988 and 2021 and for all academic disciplines.

2 Data

We use data from *Theses.fr* on all theses defended in French universities from 1988 to 2021. For each thesis we have information on the discipline of study, the defense year, the university affiliation, and the names of the PhD student and supervisor(s). *Theses.fr* is a public platform that automatically retrieves data from French university catalogs produced by libraries and documentation centers of higher education and research, making it the most comprehensive and reliable platform for French PhD graduation. Many universities in France have been merged, and we explain the process of standardization of university affiliation codes in Section B in the Appendix.

All data are, at one time or another, manually entered and this can leave room for spelling errors. Moreover, some theses are not submitted by the doctors, are lost or do not pass quality control, and are therefore not reported. We can estimate at 5% the number of theses not submitted or not treated each year[2]. It should also be noted that the processing times of theses in the institutions are long, and therefore, the data for the year 2022 may not be complete at the moment; this is why we decided to remove them. Also, an abnormally low level of observations before 1988 leads us to believe that there are missing theses, so we will finally restrict our data to the years 1988 to 2021.

We use standard methods based on first names to determine the gender of PhD students and supervisors. We use data from *INSEE* which provides for each year the number of newborn boys and girls given specific first names. To ensure accurate gender assignment, we associated a gender with a name only if over 95% of newborns with that name shared the same gender.

[2]this information was given to us by Maïté Roux, the referent of the platform of theses.fr whom I thank very much for the precision of her answers concerning the database

Additionally, to broaden coverage, we supplemented our dataset with government databases from Australia, Canada, Spain, Sweden, Switzerland, the UK, and the US. This methodology successfully identified the gender for 93% of PhD students and 94% of supervisors.

To provide information on research productivity for both PhD students and supervisors, we use a bibliometric dataset from Scopus, which gives us information on publications, journals, year of publication, co-authors, and affiliations until today (2023). This procedure requires the removal of overly common names to avoid mismatches which reduces the final sample to 336,390 theses.

We have organized the theses of our database into twenty-one academic discipline categories (see Appendix Section B for the procedure). We then split the twenty-one fields into four main categories: *Humanity and Law*, *Sciences, Technology, Engineering, and Mathematics* (STEM), *Biological and Earth Sciences* and *Social Sciences*; for each discipline, we detailed the number of observations in Table A1 in the Appendix. Below, we briefly summarize the main features of the dataset.

2.1 Supervisors

We now study how supervisors' characteristics - particularly the number of theses supervised per supervisor - vary depending on the supervisor's gender and across fields and time.

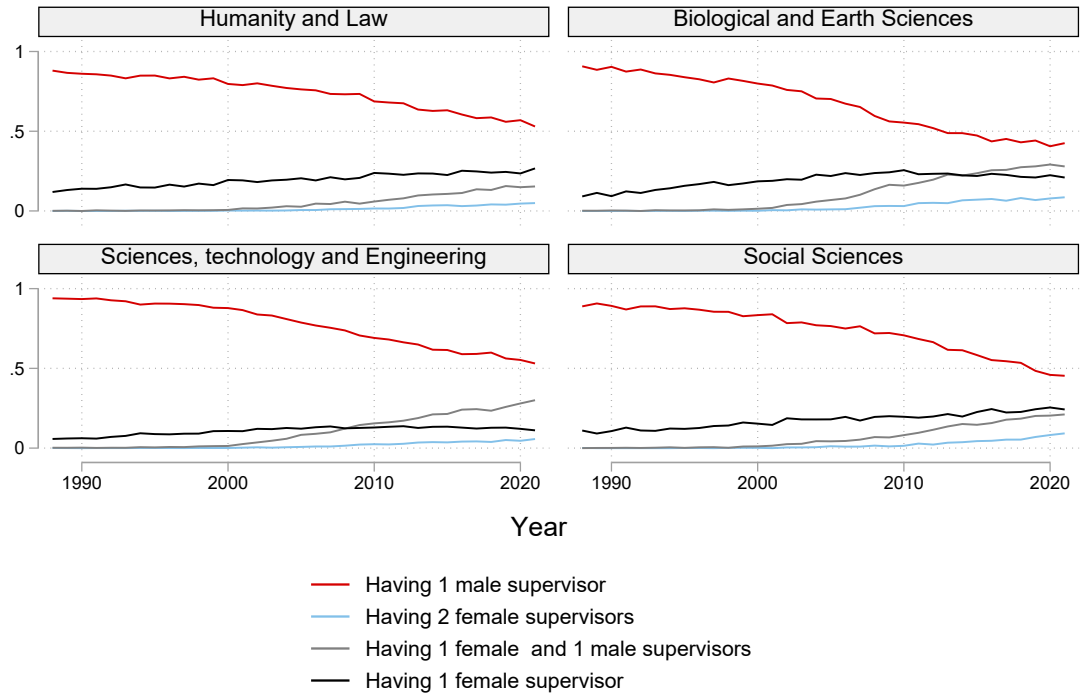


Figure 1: Cumulated Share of Female Supervisors

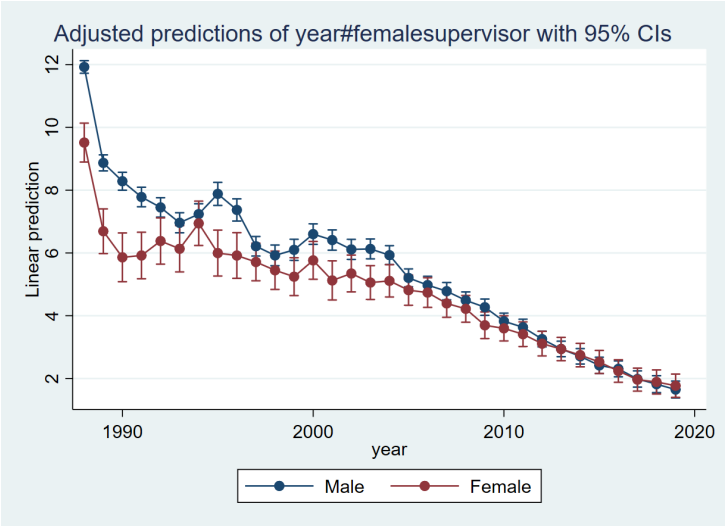
We see in Figure 1 that the cumulative share of female supervisors in our sample is low, in the entire sample, although it is increasing over time. The overall probability that a supervisor

is a woman is 25% (see Table 1). The cumulative number of supervisions of a researcher is an important factor in determining the academic level of the researcher. To be able to be the principal supervisor or the unique supervisor of a student, it is necessary to have a state diploma of HDR (Habilitation to Lead Research)[3]. In France, the proportion of women researchers with HDR degrees is around 30% in 2016[4], which reduces the probability that they will supervise theses.

Table 1: Supervisor's Sample

Disciplines	N	% Female Supervisor	Av. Number of Theses	Female Av. Number of Theses
Humanity and Law	13,638	30.23%	5.54	4.27
Biological and Earth Sciences	21,235	30.87%	3.35	2.73
Sciences, technology, engineering, and math.	41,423	18.78%	4.76	3.91
Social Sciences	13,570	31.56%	5.04	4.71
Entire Sample	89,866	25.32%	4.58	3.59

Figure 2: Predicted probability of the number of Supervision for average female and male supervisor across cohorts



NOTE: The figure shows the predicted number of students per supervisor for an average female and male supervisor. We see the prediction for each cohort, where the year represents the year of first supervision.

To analyze the number of supervisions across gender and to take into account the evolution of the share of female supervisors over time, we perform a linear regression to explain at the individual level the number of theses supervised as a function of a female dummy and a cohort dummy (equal to one for the year of first supervision); Figure 2 shows the coefficients and 95% confidence intervals. We see a gap in the number of supervisions in the older cohorts, but this difference decreases over time. Both curves decrease because the cumulative number of students is lower for the younger cohorts.

[3]Since the decree of May 25, 2016, the research supervisor or co-supervisors, may also be a person holding a doctorate (without HDR) and chosen for her scientific skills, after acceptance by the head of the institution and the research committee.

[4]*Analyse quantitative de la parité entre les femmes et les hommes parmi les enseignants-chercheurs universitaires* from the Ministry of National Education and Higher Education

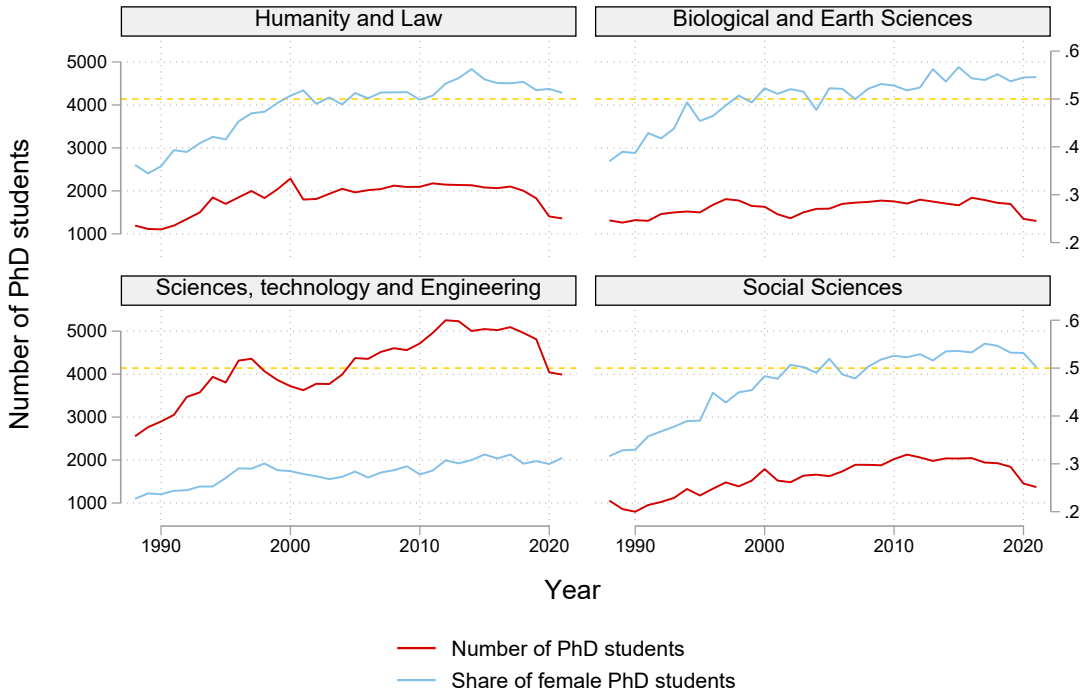


Figure 3: Share of Female PhD Students

2.2 PhD Students

The proportion of female PhD students varied across time and fields in Figure 3, it depicts the proportion of female PhD students and the total number of PhD students by discipline and over time; the yellow line corresponds to a proportion of 50%. The figure shows that the share of female students has been increasing substantially since the 1990s in all fields except for STEM, where the slope of female PhD share is positive but much smaller. Moreover, around 2005, the share of female PhD students exceeded the proportion of males for all discipline categories. In STEM, the female PhD student's share does not exceed 30%.

To make meaningful comparisons on output and avoid that PhD students do not have time to publish their first article, we are interested in the median time between the PhD Defense year and the year of the first publication. In our merged sample, 50% of men publish their first paper in three years while it takes four years for women. Indeed, to look at the evolution of students who have completed their thesis, we need to restrict the temporality of the *Thèses.fr* database before 2019 to leave at least four years between the year of thesis defense and 2023. Our study sample is finally reduced to 286,267 PhD students.

When combining both databases, we find that 42.89% of the PhD students having done a thesis between 1988 and 2019 have made at least one publication. Having one publication does not mean that the student has begun an academic career; indeed, when we look at a large time window of the career[5], we see that 40% of published students have only one publication.

We measure the research productivity in two different ways: first, we use the total number of publications; second, we restrict the publication to articles that take into account the quality

[5]Up to 20 years after PhD defense.

of the journal using the *Article Influence Score* (AIS), and we follow the method used in Bagues et al. (2017). Table 2 describes how productivity variables vary by gender between times t and $t+4$, and between times $t+4$ and $t+10$. Note that the samples are different in the two time periods; before time $t+4$, we consider all PhD students having received a PhD between 1988 and 2019; before time $t+10$, we consider all PhD students having received a PhD between 1988 and 2013. When we look before $t+4$ years, we see that former students publish on average 5.35 articles (5.88 for men and 4.33 for women). We document that former female PhD students publish less than male PhD students on average in the four discipline categories. We also note that the average number of publications is heterogeneous across disciplines.

When looking at the average number of publications between times $t+4$ and $t+10$ conditionally of having at least one publication in this time window - a way to get a better idea of whether the PhD student has effectively continued in the academy - and it seems that the gender gap is increasing with seniority.

Table 2: Summary Statistics

	Any Publication = 1							
	$t \to t+4$				$t+4 \to t+10$			
	# Publi		**Total AIS**		**# Publi**		**Total AIS**	
Gender	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Humanity and Law								
Male	2.15	2.52	.55	2.13	3.18	4.73	1.02	5.35
Female	1.76	1.87	.37	1.19	2.44	3.36	.60	3.29
Total	1.98	2.26	.48	1.88	2.85	4.16	.83	4.48
Biological and Earth Sciences								
Male	5.64	6.13	11.59	20.10	11.00	13.50	23.35	41.29
Female	4.39	4.66	8.70	14.85	7.34	9.54	15.08	25.50
Total	5.03	5.39	9.95	17.75	9.17	11.82	19.23	34.70
Sciences, technology, engineering, and math.								
Male	6.49	16.41	6.15	28.10	13.16	26.66	11.87	48.60
Female	5.29	16.50	5.35	26.81	9.81	23.98	9.65	45.00
Total	6.18	16.38	5.94	27.66	12.38	26.04	11.36	48.30
Social Sciences								
Male	2.70	2.83	1.43	3.68	4.56	6.05	2.75	7.76
Female	2.30	2.20	1.01	2.75	3.67	4.09	1.78	4.92
Total	2.51	2.59	1.22	3.23	4.13	5.22	2.30	6.59
All Fields								
Male	5.88	14.33	6.34	25.43	11.29	22.98	11.95	43.72
Female	4.33	11.94	5.42	20.88	7.20	16.73	8.96	32.84
Total	5.35	13.53	6.008	23.86	9.94	21.19	10.95	40.87

2.3 PhD Students-Supervisors Matches

We now look at the supervisor's gender composition and student-supervisor matches. We separate the gender composition in 5 categories: having 1 male or 1 female supervisor (single

supervision), having 2 male or 2 female supervisors, and having 1 male and 1 female supervisor (co-supervision). Figure 4 shows the evolution over time of the share of each composition. We see that the share of co-supervision is increasing over the year, while the share of single-supervision is decreasing.

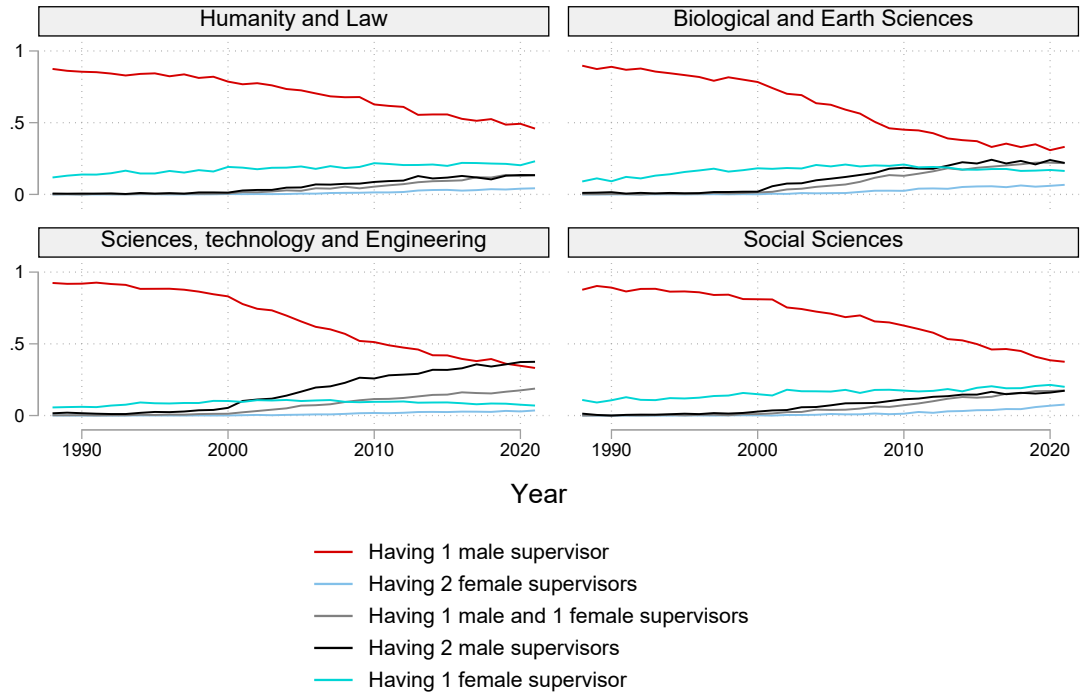


Figure 4: Supervisor's Gender Composition

Table 3 allows us to understand whether the proportion of female doctoral students is greater when the composition of supervisors is more female. It is rare to have female co-supervisors both for female and male students but there seems to be a higher probability when the PhD student is female.

Table 3: Summary Statistics

	Entire Sample		**Humanity and Law**		**STEM**		**Biological and ES**		**Social Sciences**	
VARIABLES	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Student's probabilities										
Being a female student	0.40	0.49	0.49	0.50	0.28	0.45	0.50	0.50	0.48	0.50
Having a 2 supervisors	0.22	0.41	0.12	0.33	0.28	0.45	0.23	0.42	0.18	0.39
Being female with 2 supervisors	0.22	0.42	0.14	0.34	0.30	0.46	0.24	0.42	0.21	0.40
Being male with 2 supervisors	0.22	0.41	0.11	0.31	0.27	0.44	0.22	0.41	0.16	0.36
In simple supervision										
Being female and having 1 female supervisor	0.24	0.43	0.28	0.45	0.16	0.37	0.26	0.44	0.27	0.44
Being male and having 1 female supervisor	0.15	0.35	0.19	0.39	0.11	0.31	0.19	0.39	0.17	0.37
In co-supervision										
Being female and having 2 female supervisors	0.02	0.15	0.02	0.14	0.02	0.14	0.03	0.16	0.03	0.17
Being male and having 2 female supervisors	0.01	0.10	0.01	0.10	0.015	0.10	0.02	0.13	0.01	0.12
Being female and having 2 male supervisors	0.11	0.31	0.06	0.23	0.18	0.38	0.11	0.31	0.09	0.28
Being male and having 2 male supervisors	0.14	0.35	0.06	0.23	0.19	0.39	0.12	0.33	0.09	0.28
Being female and having 1 male and 1 female supervisors	0.08	0.14	0.05	0.22	0.10	0.29	0.09	0.29	0.08	0.27
Being male and having 1 male and 1 female supervisors	0.06	0.24	0.03	0.18	0.07	0.25	0.07	0.26	0.05	0.21

According to the gender composition of the supervisors, we compute the research productivity

of those composition groups at the time of the PhD Defense. When there are two supervisors, we take the maximum productivity among the two supervisors. We plot in Figure 5 the total AIS over years by disciplines and we see that supervisor research productivity is lower when the composition is with one or two female supervisors.

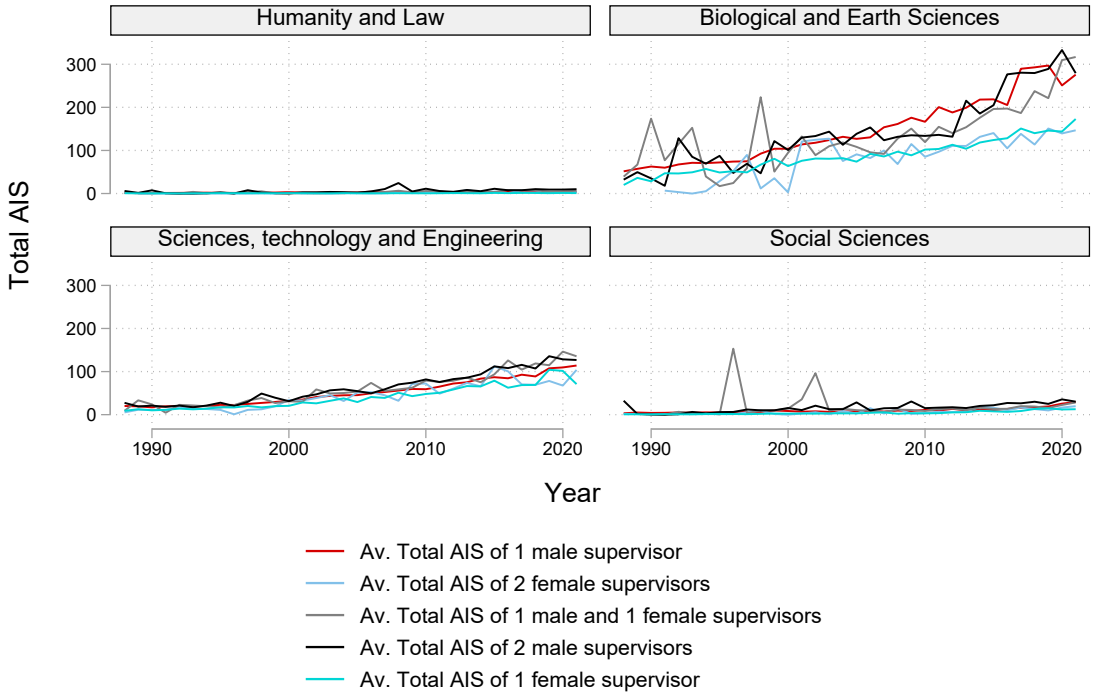


Figure 5: Average Total AIS for each Supervisor's Gender Composition at the Defense Year

3 Empirical Analysis

In our analysis, we look at the extensive and intensive margin (EM and IM) of differences in research productivity between female and male PhD students. First, we investigate the probability of having at least one publication (EM), and then we examine the effect on the research productivity of the PhD student conditional on having at least one publication (IM). We start our analysis by providing descriptive information on the average effect of being a female PhD student on both EM and IM. Secondly, we investigate the effect of the supervisor's gender composition and research productivity on both margins. To deepen the analysis, we introduce an interaction effect to understand the differences in effects between female and male PhD students. Finally, we explore the varied effects across disciplines.

3.1 Gender Gap

We estimate the gender gap of having at least one publication (EM), and whether having one publication, the number of publications and taking into account the AIS quality index (IM),

using the ordinary least squares (OLS) method:

$$Y_{it+\tau} = \beta_0 + \beta_1 F_i + \mu_i + \epsilon_{it+\tau}$$

where $Y_{it+\tau}$ is the research productivity outcome variable for each PhD student i at time $t+4$ and $t+10$ ($\tau = 4, 10$), with t the year of PhD defense. F_i is a dummy variable indicating the gender of the PhD student. We finally add three fixed effects controlling for universities, years, and disciplines (μ_i).

Table 4 shows the baseline for our estimations. We see that the probability of having at least one publication between 4 years after the year of defense t ($t+4$) and 10 years after the year of defense ($t+10$) is 37% of male, and decreases of 7.8% for female PhD students. We observe a heterogeneity across disciplines: the probability is higher at 1.5% for females compared to male PhD students in biological and earth science. The probability of publishing is lower for humanity and law and social science - it is easier to have a publication in biological and earth science and STEM. Otherwise, there is a negative gender gap for female PhD students in the other disciplines. For the IM, conditionally of having at least one publication male PhD students have on average 11 publications and females have 4 publications less. In terms of the number of publications, we observe a negative gender gap for women in all disciplines. For the average quality per publication using AIS, we have a positive coefficient: the result shows that women publish less in quantity, but better in quality in terms of AIS.

Table 4: Summary statistics - Gender Gap Intensive and Extensive Margin

	All Fields	Humanity and Law	Biological and Earth Sc.	STEM	Social Science	Economics
# PhD students	220,935	45,259	39,421	98,885	37,370	7,239
% Female PhD Students	38%	47%	48%	27%	46%	36%
Between t+4 and t+10						
Gender Gap Any_Publi	-7.8%***	-1.9%***	1.5%***	-9.6%***	-0.5%	-2.5%**
Av. Male Any_Publi	37%***	16%***	38%***	48%***	23%***	26%***
Between t+4 and t+10 and Any_Publi = 1						
Gender Gap # Publi	-4.08***	-0.73***	-3.67***	-3.35***	-0.89***	-1.26***
Av. Male # Publi	11.29***	3.19***	11.01***	13.16***	4.56***	5.20***
Gender Gap Av. Quality	0.18***	-0.01	-0.06	0.08***	-0.06***	-0.06
Av. Male Av. Quality	0.82***	0.19***	2.09***	0.66***	0.45***	0.80***

Notes: Standard errors are reported in parentheses. Significance levels are defined as follows: $p < 0.1$ *, $p < 0.05$ **, $p < 0.01$ ***.

Table 5 shows the estimates for all disciplines controlling for disciplines-fixed effect and year-fixed effect. We see that the probability of publishing is negative and robust for the controlled fixed effect. It seems that there is a stronger variation across disciplines and less if we add year-fixed effects. The trend across years seems to not vary a lot, in the next steps we will estimate the coefficient for each year.

Table 5: Gender Gap Intensive and Extensive Margin

All Disciplines	(1)	(2)	(3)	(4)
Any Publication between t+4 and t+10				
Female PhD	**-0.079*****	**-0.040*****	**-0.060*****	**-0.0503*****
	(0.002)	(0.002)	(0.002)	(0.00203)
Constant	0.371***	0.356***	0.364***	0.360***
	(0.001)	(0.001)	(0.001)	(0.00124)
Observations	220935	220935	220935	220935
# Publication between t+4 and t+10				
Female PhD	**-4.085*****	**-2.809*****	**-3.601*****	**-2.995*****
	(0.164)	(0.168)	(0.165)	(0.167)
Constant	11.29***	10.87***	11.13***	10.93***
	(0.094)	(0.094)	(0.094)	(0.094)
Av. Quality between t+4 and t+10				
Female PhD	**0.185*****	**0.016**	**0.168*****	**-0.0086**
	(0.012)	(0.011)	(0.012)	(0.011)
Constant	0.822***	0.877***	0.827***	0.886***
	(0.007)	(0.006)	(0.007)	(0.006)
Observations	75196	75196	75195	75195
Controls				
Year FE				Yes
University FE			Yes	Yes
Disciplines FE		Yes		Yes
Disciplines X Year FE				Yes

3.2 Effect of the Supervisor's Gender Composition and Research Productivity

We add some supervision characteristics, whether the thesis is co-supervised, supervisor's research productivity using the Article Influence Score (AIS). We also control for the supervisor's gender composition: if it's a female co-supervision (FF), female supervision (F), or mixed co-supervision (FM). We add a female interaction for each supervision's characteristic. We add control for the same fixed effects as the precedent model. We estimate the following econometric model through OLS:

$$Y_{it+\tau} = \beta_0 + \beta_1 F_i + \beta_2 X_{it} + \beta_3 F_i \times X_{it} + \beta_4 \delta_i + \beta_5 F_i \times \delta_i + \mu_i + \gamma_{t+\tau} + \epsilon_{it+\tau}$$

with $\tau = 4, 10$

Table 6 shows the results of the EM: we see that the negative effect of being a female PhD student is robust. It seems that there is a positive effect of the supervisor's research productivity, which tends to affect men more positively than women if we look at column (5). We find a negative effect of having one female supervisor compared to having one male supervisor without a gender-differentiated effect.

Table 6: All Fields - Dependent Variable: Any Publication

	(1)	(2)	(3)	(4)	(5)	(6)
Any Publication between t+4 and t+10						
Female PhD	-0.0818***	-0.0508***	-0.0834***	-0.0521***	-0.0768***	-0.0782***
	(0.00206)	(0.00204)	(0.00208)	(0.00204)	(0.00229)	(0.00250)
Having 2 supervisors	0.0852***	0.0165***	0.0883***	0.0158***	0.0939***	0.0971***
	(0.00294)	(0.00314)	(0.00352)	(0.00365)	(0.00376)	(0.00436)
Total AIS of supervisor(s)	0.000367***	0.000152***	0.000369***	0.000156***	0.000392***	0.000392***
	(0.00000824)	(0.00000845)	(0.00000824)	(0.00000846)	(0.0000114)	(0.0000114)
Having 1 female supervisor			0.0236***	0.0234***		0.0241***
			(0.00307)	(0.00300)		(0.00427)
Having 2 female supervisors			0.0111	0.0373***		-0.00942
			(0.0131)	(0.0126)		(0.0203)
Having 1 female and 1 male supervisors			-0.000624	0.0105*		-0.000511
			(0.00615)	(0.00592)		(0.00816)
Female X Having 2 supervisors					-0.0227***	-0.0258***
					(0.00603)	(0.00739)
Female X Total AIS of supervisor(s)					-0.0000512***	-0.0000497***
					(0.0000165)	(0.0000165)
Female X Having 1 female supervisor						-0.00198
						(0.00615)
Female X Having 2 female supervisors						0.0441*
						(0.0268)
Female X Having 1 female and 1 male supervisors						0.00561
						(0.0125)
University FE		Yes		Yes		
Disciplines X Year FE		Yes		Yes		
Observations	218837	218837	218837	218837	218837	218837

Standard errors in parentheses

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

Once we look at the IM in Table 7 conditionally of having published between $t + 4$ and $t + 10$, we have a positive effect of being with a female supervisor compared to having one male

supervisor and a positive effect of being with both one male and one female supervisor compared to be with two male supervisors. We don't find a clear effect of being in co-supervision.

Table 7: All Fields - Dependent Variable: Number of Publications

	(1)	(2)	(3)	(4)	(5)	(6)
# Publications between t+4 and t+10 if Any Publication = 1						
Female PhD	-4.338***	-3.034***	-4.287***	-3.023***	-3.710***	-3.665***
	(0.163)	(0.167)	(0.164)	(0.167)	(0.189)	(0.207)
Having 2 supervisors	1.321***	0.219	1.583***	0.355	1.378***	1.719***
	(0.205)	(0.225)	(0.244)	(0.261)	(0.251)	(0.290)
Total AIS of supervisor(s)	0.0199***	0.0176***	0.0199***	0.0176***	0.0238***	0.0238***
	(0.000552)	(0.000585)	(0.000552)	(0.000586)	(0.000705)	(0.000705)
Having 1 female supervisor			-0.360	-0.0553		-0.372
			(0.236)	(0.238)		(0.310)
Having 2 female supervisors			-2.062**	-1.081		-1.529
			(0.911)	(0.901)		(1.346)
Having 1 female and 1 male supervisors			-0.818*	-0.349		-1.398***
			(0.423)	(0.418)		(0.533)
Female X Having 2 supervisors					-0.235	-0.563
					(0.435)	(0.535)
Female X Total AIS of supervisor(s)					-0.0101***	-0.0102***
					(0.00113)	(0.00113)
Female X Having 1 female supervisor						-0.0628
						(0.479)
Female X Having 2 female supervisors						-0.739
						(1.841)
Female X Having 1 female and 1 male supervisors						1.611*
						(0.882)
University FE		Yes		Yes		
Disciplines X Year FE		Yes		Yes		
Observations	74953	74952	74953	74952	74953	74953

Standard errors in parentheses

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

3.3 Heterogeneous Effect Across Disciplines

Table 8: By Disciplines - Dependent Variable: Any Publication

	Humanity and Law		Biological and Earth Sc.		STEM		Social Sc.	
Any Publi t+4 to t+10	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female PhD	-0.035***	-0.035***	-0.002	-0.007	-0.098***	-0.098***	-0.025***	-0.027***
	(0.003)	(0.003)	(0.005)	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)
Having 2 supervisors	0.070***	0.078***	-0.087***	-0.103***	0.024***	0.025***	0.072***	0.061***
	(0.007)	(0.008)	(0.008)	(0.009)	(0.005)	(0.005)	(0.007)	(0.009)
Total AIS Supervisor(s)	-0.000000442	-0.000000643	0.000116***	0.000127***	0.000157***	0.000158***	0.000217***	0.000225***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Having 1 female supervisor		0.003		0.064***		0.004		0.023***
		(0.004)		(0.007)		(0.006)		(0.006)
Having 2 female supervisors		0.010		0.118***		0.011		0.046*
		(0.025)		(0.026)		(0.022)		(0.028)
Having 1 female and 1 male supervisors		-0.023*		0.070***		-0.004		0.039***
		(0.013)		(0.014)		(0.009)		(0.014)
Controls								
University	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	44689	44689	39174	39174	97840	97840	37120	37120

Standard errors in parentheses

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

We compute the same model for the different disciplines' categories and we find heterogeneous effects across disciplines. Table 8 shows the estimates for having any publication between t+4 and t+10 by disciplines. We see that in STEM the probability of publishing is smaller for women, we find no gender effect in the probability of having published for women in biological and earth science. In humanity and law, we have a higher female share but lower female participation.

Table 9: Intensive Margin by Disciplines

	Humanity and Law		Biological and Earth Sc.		STEM		Social Sc.	
t+4 to t+10	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	# Publi	Av. Quality	# Publi	Av. Quality	# Publi	Av. Quality	# Publi	Av. Quality
Female PhD	-0.754***	0.00000213	-3.755***	-0.117***	-3.834***	0.0283**	-1.055***	-0.0498***
	(0.103)	(0.017)	(0.191)	(0.038)	(0.288)	(0.012)	(0.115)	(0.019)
Having 2 supervisors	0.428**	0.045	0.204	-0.402***	0.013	-0.053***	0.499**	0.029
	(0.204)	(0.034)	(0.417)	(0.083)	(0.385)	(0.016)	(0.213)	(0.035)
Total AIS of supervisor(s)	0.02***	0.002***	0.004***	0.001***	0.045***	0.003***	0.005***	0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Having 1 supervisor	-0.160	-0.017	-0.059	0.119**	-0.046	0.083***	-0.068	-0.053**
	(0.139)	(0.023)	(0.250)	(0.049)	(0.440)	(0.018)	(0.162)	(0.027)
Having 2 female supervisors	-0.771	-0.151	-0.916	0.058	-1.512	0.296***	1.070*	-0.089
	(0.582)	(0.096)	(1.011)	(0.200)	(1.597)	(0.066)	(0.599)	(0.099)
Having 1 female and 1 male supervisors	-0.451	-0.093*	0.059	0.053	-0.521	0.102***	-0.184	-0.064
	(0.333)	(0.055)	(0.586)	(0.116)	(0.640)	(0.026)	(0.325)	(0.053)
Controls								
University	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6704	6704	15094	15094	44586	44586	8542	8542

Standard errors in parentheses

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

Table 9 shows the estimates for the intensive margin by discipline. It seems that the gender gap deepens with career time. We see that it's better to have two supervisors in social science and humanity and law for academic participation, but there is an opposite and negative effect for STEM and biological and earth Science - where having one female supervisor increases the

probability of publishing compared to having one male supervisor. In biological and earth science women publish less and less quality, but in STEM women publish with a better quality.

The tables in Section A in the appendix show the estimates of the same model with the gender-differentiated effects across all disciplines. We see that in humanity and law and biological and earth science, there is a penalty on the number of publications if a female PhD student has two supervisors. For all disciplines, the research productivity of the supervisor(s) affects less positively female PhD students than males in the number of publications.

4 Conclusion and Next Steps

Our next phase involves improving our analysis by adding new variables. First, we will retrieve the number and name of co-authors for each publication. It will allow us to improve the quality measure of publication and control for the gender distribution of the supervisor's network; we can also analyze the effect of the supervisor's network on her PhD student. Second, we would like to have a better idea of the work environment by adding some controls on the university: female share of supervisors available in the university at the time of the thesis, the research quality of the university using research quality of supervisors, and the number of PhD students.

The aim of this study is to understand some factors contributing to both the under-representation of women in academia and the gender research productivity gap. As we continue to grapple with this question, our findings, coupled with existing literature, underscore that gender disparities take root early in researchers' careers. To address this challenge, proactive policies are crucial, offering a pathway to mitigate inequalities and foster an environment that encourages women to pursue and sustain careers in academia after their PhDs.

References

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva, "Does the Gender Composition of Scientific Committees Matter?," *American Economic Review*, April 2017, *107* (4), 1207–38.

Barbezat, Debra A., "Gender Differences in Research Patterns Among PhD Economists," *The Journal of Economic Education*, 2006, *37* (3), 359–375.

Campbell, Robert A., "Preparing the Next Generation of Scientists: The Social Process of Managing Students," *Social Studies of Science*, 2003, *33* (6), 897–927.

Conley, John P., Ali Sina Önder, and Benno Torgler, "Are all economics graduate cohorts created equal? Gender, job openings, and research productivity," *Scientometrics*, August 2016, *108* (2), 937–958.

et al Rachid E., Noureddine T. Tamim H., "Gender disparity in research productivity across departments in the faculty of medicine: a bibliometric analysis.," *Scientometrics*, 01 2021, *126*, 4715–4731.

et Anja Prummer Ductor Lorenzo, Sanjeev Goyal, "Gender and Collaboration," Technical Report 2023.

Gaule, Patrick and Mario Piacentini, "An advisor like me? Advisor gender and post-graduate careers in science," *Research Policy*, 2018, *47* (4), 805–813.

Ginther, Donna and Shannon Kahn, "Women in Economics: Moving Up or Falling Off the Academic Career Ladder?," *Journal of Economic Perspectives*, 02 2004, *18*, 193–214.

M., Mairesse J. Stephan P. Lane J. Pezzoni, "Gender and the publication output of graduate students: A case study," *PLos One*, 2016, *11.*

McDowell, John M., Larry D. Singell, and Mark Stater, "Two to Tango? Gender Differences in the Decisions to Publish and Coauthor," *Economic Inquiry*, January 2006, *44* (1), 153–168.

Patsali, Sofia, Michele Pezzoni, and Fabiana Visentin, "The Impact of Research Independence on PhD Students' Careers: Large-Scale Evidence from France," in "Investments in Early Career Scientists: Data and Research Gaps" NBER Online, France November 2021.

Sarsons, Heather, "Recognition for Group Work: Gender Differences in Academia," *American Economic Review*, May 2017, *107* (5), 141–145.

Stephan, Paula and Sharon Levin, "The Critical Importance of Careers in Collaborative Scientific Research," *Revue d'Économie Industrielle*, 01 1997, *79*, 45–61.

Tenenbaum, Harriet R., Faye J. Crosby, and Melissa D. Gliner, "Mentoring Relationships in Graduate School," *Journal of Vocational Behavior*, 2001, *59* (3), 326–341.

A Figures and Tables

Table A1: Description of Disciplines

Disciplines	Observations
Humanity and Law	**67,104**
History and Archaeology	12,487
Journalism, Librarianship and Curatorial Studies	1,870
Language and Culture	20,319
Law, Justice and Law Enforcement	18,590
Philosophy and Religion	6,295
The Arts	7,543
Biological and Earth Sciences	**57,330**
Biological Sciences	42,969
Earth Sciences	14,361
Sciences, technology and Engineering	**153,461**
Agricultural, Veterinary and Environmental Sciences	2,428
Chemical Sciences	22,440
Engineering and technology	51,022
Information, computing and Communication Sciences	23,756
Mathematical Sciences	11,160
Physical Sciences	42,655
Social Sciences	**58,495**
Architecture, Urban Environment and Building	1,478
Behavioural and Cognitive Sciences	14,512
Commerce, Management, Tourism and Services	8,168
Economics	10,980
Education	3,740
Policy and Political Science	4,631
Studies in Human Society	14,986
Entire Sample	**336,390**

Table A2: Biological and Earth Science - Dependent Variable: Number of Publication

	(1)	(2)	(3)	(4)	(5)	(6)
# Publications between t+4 and t+10 if Any Publication = 1						
Female PhD	-3.669***	-3.761***	-3.674***	-3.755***	-3.446***	-3.502***
	(0.190)	(0.190)	(0.190)	(0.191)	(0.226)	(0.254)
Having 2 supervisors	0.578**	0.164	0.600	0.204	1.073***	0.779
	(0.294)	(0.322)	(0.395)	(0.417)	(0.415)	(0.536)
Total AIS of supervisor(s)	0.00455***	0.00405***	0.00458***	0.00403***	0.00491***	0.00492***
	(0.000399)	(0.000413)	(0.000400)	(0.000415)	(0.000526)	(0.000527)
Having 1 female supervisor			0.220	-0.0596		0.1000
			(0.244)	(0.250)		(0.355)
Having 2 female supervisors			-0.570	-0.916		0.167
			(1.008)	(1.011)		(1.628)
Having 1 female and 1 male supervisors			0.197	0.0597		0.850
			(0.585)	(0.586)		(0.829)
Female X Having 2 supervisors					-0.985*	-0.383
					(0.588)	(0.794)
Female X Total AIS of supervisor(s)					-0.000827	-0.000807
					(0.000806)	(0.000809)
Female X Having 1 female supervisor						0.209
						(0.490)
Female X Having 2 female supervisors						-1.068
						(2.083)
Female X Having 1 female and 1 male supervisors						-1.182
						(1.173)
University FE		Yes		Yes		
Disciplines X Year FE		Yes		Yes		
Observations	21625	21618	21625	21618	21625	21625

Standard errors in parentheses

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

Table A3: STEM - Dependent Variable: Number of Publication

	(1)	(2)	(3)	(4)	(5)	(6)
# Publications between t+4 and t+10 if Any Publication = 1						
Female PhD	-3.824***	-3.850***	-3.821***	-3.834***	-3.077***	-2.989***
	(0.288)	(0.287)	(0.288)	(0.288)	(0.345)	(0.369)
Having 2 supervisors	-0.128	-0.148	0.101	0.0134	-0.239	0.194
	(0.307)	(0.344)	(0.351)	(0.385)	(0.355)	(0.401)
Total AIS of supervisor(s)	0.0475***	0.0449***	0.0475***	0.0449***	0.0523***	0.0523***
	(0.00121)	(0.00126)	(0.00121)	(0.00126)	(0.00144)	(0.00144)
Having 1 female supervisor			0.385	-0.0462		0.580
			(0.437)	(0.440)		(0.520)
Having 2 female supervisors			-1.745	-1.512		-1.565
			(1.602)	(1.597)		(2.124)
Having 1 female and 1 male supervisors			-0.565	-0.521		-1.529**
			(0.642)	(0.640)		(0.763)
Female X Having 2 supervisors					0.362	-0.502
					(0.705)	(0.830)
Female X Total AIS of supervisor(s)					-0.0162***	-0.0163***
					(0.00265)	(0.00265)
Female X Having 1 female supervisor						-0.751
						(0.961)
Female X Having 2 female supervisors						-0.439
						(3.249)
Female X Having 1 female and 1 male supervisors						3.328**
						(1.419)
University FE		Yes		Yes		
Disciplines X Year FE		Yes		Yes		
Observations	44591	44586	44591	44586	44591	44591

Standard errors in parentheses

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

Table A4: Humanity and Law - Dependent Variable: Number of Publication

	(1)	(2)	(3)	(4)	(5)	(6)
# Publications between t+4 and t+10 if Any Publication = 1						
Female PhD	-0.752***	-0.767***	-0.743***	-0.754***	-0.667***	-0.639***
	(0.103)	(0.103)	(0.103)	(0.103)	(0.111)	(0.124)
Having 2 supervisors	0.672***	0.275*	0.815***	0.428**	1.083***	1.299***
	(0.160)	(0.167)	(0.199)	(0.204)	(0.222)	(0.266)
Total AIS of supervisor(s)	0.0217***	0.0175***	0.0215***	0.0171***	0.0190***	0.0189***
	(0.00389)	(0.00396)	(0.00389)	(0.00397)	(0.00489)	(0.00490)
Having 1 female supervisor			-0.0560	-0.160		-0.00501
			(0.138)	(0.139)		(0.193)
Having 2 female supervisors			-0.669	-0.771		-1.355
			(0.586)	(0.582)		(0.993)
Having 1 female and 1 male supervisors			-0.356	-0.451		-0.559
			(0.334)	(0.333)		(0.473)
Female X Having 2 supervisors					-0.860***	-1.102***
					(0.320)	(0.401)
Female X Total AIS of supervisor(s)					0.00705	0.00703
					(0.00805)	(0.00806)
Female X Having 1 female supervisor						-0.122
						(0.275)
Female X Having 2 female supervisors						1.362
						(1.236)
Female X Having 1 female and 1 male supervisors						0.547
						(0.670)
University FE		Yes		Yes		
Disciplines X Year FE		Yes		Yes		
Observations	6713	6704	6713	6704	6713	6713

Standard errors in parentheses

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

Table A5: Social Science - Dependent Variable: Number of Publication

	(1)	(2)	(3)	(4)	(5)	(6)
# Publications between t+4 and t+10 if Any Publication = 1						
Female PhD	-0.917***	-1.055***	-0.932***	-1.055***	-0.799***	-0.776***
	(0.114)	(0.115)	(0.115)	(0.115)	(0.125)	(0.140)
Having 2 supervisors	0.959***	0.526***	0.945***	0.499**	1.208***	0.994***
	(0.163)	(0.171)	(0.208)	(0.213)	(0.232)	(0.279)
Total AIS of supervisor(s)	0.00587***	0.00524***	0.00598***	0.00526***	0.00830***	0.00847***
	(0.00129)	(0.00132)	(0.00129)	(0.00133)	(0.00185)	(0.00186)
Having 1 female supervisor			0.131	-0.0677		0.215
			(0.159)	(0.162)		(0.236)
Having 2 female supervisors			0.987	1.070*		3.247***
			(0.602)	(0.599)		(1.061)
Having 1 female and 1 male supervisors			-0.0900	-0.184		0.349
			(0.326)	(0.325)		(0.485)
Female X Having 2 supervisors					-0.487	-0.113
					(0.326)	(0.418)
Female X Total AIS of supervisor(s)					-0.00465*	-0.00482*
					(0.00258)	(0.00258)
Female X Having 1 female supervisor						-0.179
						(0.320)
Female X Having 2 female supervisors						-3.264**
						(1.294)
Female X Having 1 female and 1 male supervisors						-0.734
						(0.659)
University FE		Yes		Yes		
Disciplines X Year FE		Yes		Yes		
Observations	8547	8542	8547	8542	8547	8547

Standard errors in parentheses

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

B Data Thèses.fr - Detailed Procedure

We use data from *Theses.fr* on all theses defended in French universities from 1988 to 2021. *Theses.fr* is a public platform that automatically retrieves data from French university catalogs produced by libraries and documentation centers of higher education and research, making it the most comprehensive and reliable platform for French PhD graduation.

All data are, at one time or another, manually entered and this can leave room for spelling errors. Moreover, some theses are not submitted by the doctors, are lost or do not pass quality control, and are therefore not reported. We can estimate at 5% the number of theses not submitted or not treated each year. Furthermore, the processing times of theses in the institutions are long, and therefore, the data for the year 2022 may not be complete at the moment; this is why we decided to remove them. Also, an abnormally low level of observations before 1988 leads us to believe that there are missing theses, so we will finally restrict our data to the years 1988 to 2021.

Starting with a total of 407,260 theses defended on *Theses.fr* between 1988 and 2021, we implement a series of exclusions. Theses with more than two supervisors (2% of the total) are eliminated, resulting in a refined dataset of 399,118 theses. Additional exclusions are applied to

ensure completeness, including dropping theses with incomplete names of PhD students and supervisors, and missing discipline. After these steps, the dataset is further refined to 397,726 theses.

For each thesis, we have information on the discipline of study, the defense year, the university affiliation, and the names of the PhD student and supervisor(s). In the subsequent sections, we will delve into the details of the cleaning procedures applied to discipline and university affiliation. Then, we will explain why we exclude health and medical sciences from our sample. Following that, we will explain the methodology employed to associate gender with the first names.

B.1 Gender association

In this study, the gender association of both students and supervisors is determined using first names. We use *INSEE* database, which compiles first names given in France from 1900 to 2020, along with the corresponding gender distribution considering the total number of attributions between 1940 and 2020. We establish a reliable gender ratio for each first name that is associated with both genders, we reject those representing less than 5% of the distribution, and we can't distinguish for the others. However, due to potential issues with foreign students' names, it may be necessary to explore databases from other countries. *INSEE* database allows gender association for 314,553 out of 340,175 PhD first names. Collecting data from other governmental databases from New Zealand, Spain, the United Kingdom, and the United States, we completed the gender association of 9,248 PhD students. We identify the gender of 2,000 more PhD students by employing a methodology that relies on combinations of the last two letters in their first names and the corresponding probability linked to a particular gender[6]. We remain with 3% of non-gendered first names - that can be associated both to a male or female - and that we can't identify. Following the same method, we can identify 95% of supervisors.

B.2 Disciplines

The management of discipline categories in *Thèses.fr* is not precise since part of the data has been collected manually. The database had around 22,000 unique values for the discipline variable to be categorized into twenty-two subcategories of discipline and coarsely in four larger categories following The Australian and New Zealand Standard Research Classification (ANZSRC). We proceeded by researching keywords to associate them manually by starting with first precise filtering with words specific to the various categories, as illustrated in the following example.

Example
"CHIMIE ORGANIQUE" for "Chemical Sciences"
"INFORMATIQUE" for ""Information, computing and Communication Sciences"
"SCIENCES BIOLOGIQUES" for "Biological Sciences" ...

Then, we filter with more and more general keywords by checking manually that there is no mis-association, some general keywords are given in the following example.

[6]We follow the same methodology used in Benveniste, 2023

Example
"MAGNETISME" for "Physical Sciences"
"LANGUES" for "Language and Culture"
"VEGETAL" for "Biological Sciences" ...

We continue by filtering from the most precise to the largest, so the order of compilation is very important to avoid errors. For example, the keyword "INFORMATION" can be associated with both "Journalism, Librarianship and Curatorial Studies" and "Information, computing and Communication Sciences". We can associate it with one of the categories once there is no more discipline with this keyword in the other category. For unknown or ambiguous disciplines, we must refer to the thesis title and repeat the same procedure by keywords. The association of categories can involve errors, especially for multidisciplinary theses that we have to associate with only one of the disciplines. This is why we have decided to create 4 main categories which make it possible to encompass similar subjects for which an association error is more frequent.

Drop *Health and Medical Sciences* discipline. In this section, we will explain why the data theses on Health and Medical sciences were not reliable before the 2000s. We realized that there were irregularities in the theses done in medicine around the year 1994. We determined the origin of these errors: *Thèses.fr* automatically selects the defended doctoral theses, and then the algorithm rejects everything that is not classified as a thesis. In the context of health studies in France, the exercise's theses (*thèses d'exercice*) are defended to obtain a State Diploma of Doctor, which allows practicing medicine; they are not doctoral theses, defended to obtain the national diploma of doctor (*diplôme national de doctorat*). Unfortunately, in the importation of the data in *Thèses.fr*, a large number of *thèses d'exercice* are indicated as "theses" which cannot be identified and thus distort the results. Figure B1 presents the number of theses defended since 1988 in health and medical sciences; it shows that institutions began to resolve this error by making a clear distinction between doctoral theses and theses of exercise around the 2000s. Indeed, we had confirmation that the data were reliable from the early 2000s.

B.3 University

For several years, French universities have been undergoing a process of merging institutions, ostensibly to boost their international attractiveness[7]. We had to standardize the code of the universities for the analysis. We follow the documentation of the university's coding provided by *Thèses.fr*[8] and we track changes in names. There are 26 universities created between 2007 and 2020 which allows the merging of 76 universities. For example, in 2013 Aix-Marseille University was created, merging Aix-Marseille 1, Aix-Marseille 2, and Aix-Marseille 3. However, sometimes there are splits that no longer allow us to distinguish between previous codes. It's then easier to use a single code for universities that have split up, even if you lose precision. For example, the University Paris-Saclay is a merger of 11 institutions in 2015 which finally separated at the end

[7] https://www.enseignementsup-recherche.gouv.fr/fr/premier-bilan-des-fusions-d-universites-realisees-entre-2009-et-2017-47515

[8] https://documentation.abes.fr/guide/html/regles/CodesUnivEtab.htm

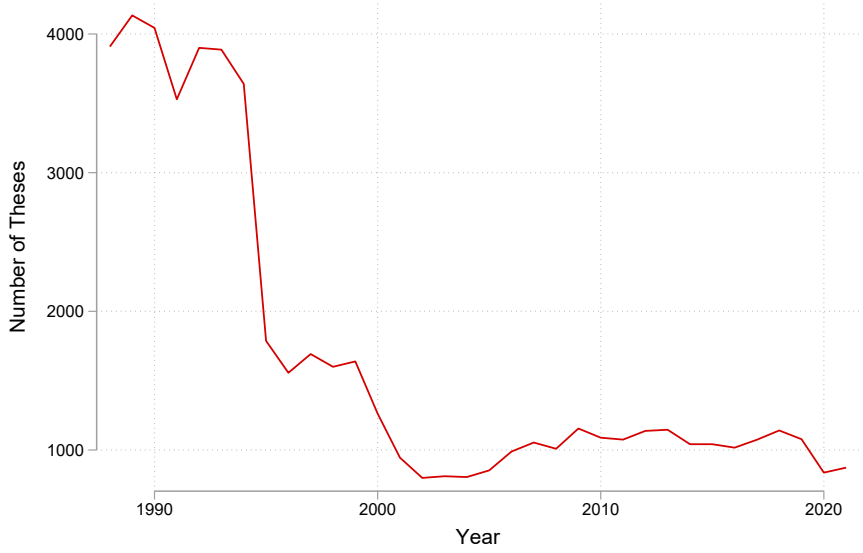


Figure B1: Number of Thesis by Year of Defense in Health and Medical Sciences

of 2019 into two different institutions. We detail the list of all institutions and their change in coding in the following tables: Table B6, B7, and B8 are the universities, Table B9 the National Institute of Polytechnics, and Table B10 the Higher Education Establishment. All the codes of the universities are associated with their name and the evolution of their code over the years. We focus on the period 1988 to 2021, any changes and code that appears before or after are taken into account. If the description is empty, it means that there is no change during the period.

Code	University	Description
AGUY+ANTI+YANE*	Antilles-Guyane	ANTI and YANE since 2015
AIX1	Aix-Marseille 1	See AIXM since 2012
AIX2	Aix-Marseille 2	See AIXM since 2012
AIX3	Aix-Marseille 3	See AIXM since 2012
AIXM	Aix-Marseille	Creation 2012
AMIE	Amiens	
ANGE	Angers	
ANTI	Antilles	Creation 2015
ARTO	Artois	
AVIG	Avignon	
AZUR (=COAZ)**	Univ. Côte d'Azur (ComUE)	Creation 2016, changing code in 2020
BELF	Belfort Montbéliard	See UBFC since 2017
BESA	Besançon	See UBFC since 2017
BOR1 + BOR4***	Bordeaux 1 + 4	See BORD since 2014
BOR2	Bordeaux 2	See BORD since 2014
BOR3	Bordeaux 3	See BORD since 2014
BORD	Bordeaux	Creation 2014
BRES	Brest - Bretagne occidentale	
CAEN	Caen	See NORM since 2017
CERG (=CYUN)	Cergy-Pontoise	Changing code CYUN in 2020
CHAM	Chambéry	See GREN since 2010
CLF1	Clermont-Ferrand 1	See CLFA since 2021
CLF2	Clermont-Ferrand 2	See CLFA since 2021
CLFA (=UCFA)	Univ. Clermont Auvergne	Changing code UCFA in 2020
COMP	Compiègne	
CORT	Corte	
DIJO	Dijon	See UBFC since 2017
DUNK	Littoral Dunkerque	
EVRY	Evry Val d'Essonne	See SACL since 2015
GRAL	Univ. Grenoble Alpes	
GRE1	Grenoble 1	See GREN since 2010
GRE2	Grenoble 2	See GREN since 2010
GRE3	Grenoble 3	See GREN since 2010
GREN(=GREA = GRAL)	Grenoble	Changing code in 2015, 2020
LARE	La Réunion	
LARO	La Rochelle	
LEHA	Le Havre	See NORM since 2017
LEMA	Le Mans	

Table B6: Universities

All the code of the universities associated with their name and the evolution of their code over the years. We focus on the period 1988 to 2021, any changes and code that appears before or after are taken into account. If the description is empty, it means that there is no change during the period. * Guyane and Antilles were part of the same university at the beginning and then split, so we have to do only one university with all(because we don't know who was in which university); ** The sign equal, when the code name changed but represents the same university; *** BOR4 since 1995 for Law, Social Sciences and politics, Economics and Management theses), so we have to merge the two universities

Code	University	Description
LIL1	Lille 1	See LILU since 20185
LIL2	Lille 2	See LILU since 2018
LIL3	Lille 3	See LILU since 2018
LILU	Univ.polfLille	Creation 2018
LIMO	Limoges	
LORI	Lorient-Bretagne sud	
LORR	Univ. de Lorraine	Creation 2012
LYO1	Lyon 1	See LYSE since 2015
LYO2	Lyon 2	See LYSE since 2015
LYO3	Lyon 3	See LYSE since 2015
LYSE	Lyon (COMUE)	Creation 2015
MARN	Marne la Vallée	See PEST since 2008
METZ	Metz	See LORR since 2012
MON1	Montpellier 1	See MONT since 2015
MON2	Montpellier 2	See MONT since 2015
MON3	Montpellier 3	
MONT	Montpellier	Creation 2015
MULH	Mulhouse	
NAN1	Nancy 1	See LORR since 2012
NAN2	Nancy 2	See LORR since 2012
NANT	Nantes	
NCAL	Nouvelle Calédonie	
NICE	Nice	See AZUR since 2016
NIME	Nîmes	
NORM	Normandie (COMUE)	Creation 2017
PA01	Paris 1	
PA02	Paris 2	
PA03	Paris 3	See USPC de 2015 à 2019
PA04	Paris 4	See SORU since 2018
PA05	Paris 5	See USPC de 2015 à 2019 See UNIP since 2019
PA06	Paris 6	See SORU since 2018
PA07	Paris 7	See USPC de 2015 à 2019 See UNIP since 2019
PA08	Paris 8	
PA09	Paris 9	See PSLE since 2016
PA10	Paris 10	
PA11	Paris 11	See SACL since 2015
PA12	Paris 12	See PEST de 2008 à 2020
PA13	Paris 13	See USPC de 2015 à 2019
PACI+NCAL+POLF*	Pacifique	NCAL and POLF since 1999
PAUU	Pau	
PERP	Perpignan	
PEST(=PESC)**	Paris Est (COMUE)	
POIT	Poitiers	
POLF	Polynésie française	

Table B7: Universities

All the code of the universities associated with their name and the evolution of their code over the years. We focus on the period 1988 to 2021, any changes and code that appears before or after are taken into account. If the description is empty, it means that there is no change during the period. * Nouvelle Calédonie and Polynésie française were part of the same university at the beginning and then split, so we have to use only one code with both as we can't distinguish them. ** PEST changed its name in 2015 to PESC

Code	University	Description
REIM	Reims	
REN1	Rennes 1	
REN2	Rennes 2	
ROUE	Rouen	
SACL+UPAS+IPPA+IAVF*	Univ. Paris-Saclay (ComUE)	Creation in 2015
SORU	Sorbonne Univ.	
STET	Saint-Etienne	See LYSE since 2015
STR1	Strasbourg 1	See STRA since 2009
STR2	Strasbourg 2	See STRA since 2009
STR3	Strasbourg 3	See STRA since 2009
STRA	Strasbourg	Creation 2009
TOU1	Toulouse 1	
TOU2	Toulouse 2	
TOU3	Toulouse 3-Ec. nationale vétérinaire	
TOUL	Toulon	
TOUR	Tours	
TROY	Troyes	
UBFC	Bourgogne Franche-Comté	Creation 2017
UCFA	Univ. Clermont-Auvergne	
UEFL	Univ. Gustave Eiffel	
UNIP	Univ. de Paris	Creation 2019
UPHF	Univ. Polytech. Hauts-de-France - Valenciennes	
USPC+PA03+PA13 +INAL+UNIP**	Sorbonne Paris Cité	Creation in 2019
VALE	Valenciennes	See UPHF since 2019
VERS	Versailles St Quentin en Yvelines	See SACL since 2015
YANE	Guyane	Creation 2015

Table B8: Universities

All the code of the universities associated with their name and the evolution of their code over the years. We focus on the period 1988 to 2021, any changes and code that appears before or after are taken into account. If the description is empty, it means that there is no change during the period. * IAVF is a new branch in 2016 and SACL was divided into UPAS and IPPA in 2019, as we can't distinguish, we use the same code for the three. ** There is a merge and then a split of universities, so we use one code for PA03, PA13, INAL, and UNIP only after 2019.

Code	Institute	Description
INPG	Institut national polytechnique - Grenoble	See GREN since 2009
INPL	Institut national polytechnique - Lorraine	
INPT	Institut national polytechnique - Toulouse	
IPPA	Institut Polytechnique de Paris	

Table B9: National Institute of Polytechnics

All the code of the universities associated with their name and the evolution of their code over the years. We focus on the period 1988 to 2021, any changes and code that appears before or after are taken into account. If the description is empty, it means that there is no change during the period.

Code	Establishment	Description
AGPT+EIAA +ENGR+INAP*	AgroParisTech	See SACL since 2015
CLIL	Centrale Lille Institut	
CNAM	Conservatoire national des arts et métiers	
CSUP	CentraleSupélec	See SACL since 2015
DENS	Ec. normale supérieure - Cachan	See SACL since 2015
ECAP	Ec. centrale des arts et manufactures de Paris	See SACL since 2015
ECDL	Ec. centrale de Lyon	See LYSE since 2015
ECDM	Ec. centrale de Marseille	
ECDN	Ec. centrale de Nantes	See CLIL since 2020
ECLI	Ec. centrale de Lille	See CLIL since 2020
EHEC	Ec. des hautes études commerciales	See SACL since 2015
EHES	Ec. des hautes études en sciences sociales	
EIAA	Ec. nationale supérieure des industries alimentaires - Massy	See AGPT since 2007-
EMAC	Ec. nationale des Mines d'Albi-Carmaux	
EMAL	IMT Mines Alès	
EMNA	Ec. des Mines de Nantes	See IMTA since 2017
EMSE	Ec. nationale supérieure des Mines - Saint-Etienne	
ENAM	Ec. nationale supérieure d'arts et métiers	See HESA since 2020
ENCM	Ec. nationale supérieure de chimie de Montpellier	
ENCP	Ec. nationale des chartes	
ENCR	Ec. nationale supérieure de chimie de Rennes	
ENGR	Ec. nationale du génie rural, des eaux et forêts	See AGPT since 2007
ENIB	Ec. nationale d'ingénieurs de Brest	
ENIS	Ec. nationale d'ingénieurs de Saint-Etienne	See LYSE since 2015
ENMP	Ec. nationale supérieure des Mines - Paris	See PSLE since 2016
ENPC	Ec. nationale des ponts et chaussées	See PEST since 2008
ENSL	Ec. normale supérieure (sciences) - Lyon	See LYSE since 2015
ENSR	Ec. normale supérieure de Rennes	
ENST	Ec. nationale supérieure des télécommunications	See SACL since 2015
ENSU	Ec. normale supérieure- Paris (rue d'Ulm)	See PSLE since 2016
ENTA	Ec. nationale supérieure de techniques avancées Bretagne	
ENTP	Ec. nationale des travaux publics	See LYSE since 2015
EPHE	Ec. pratique des hautes études	See PSLE since 2016
EPXX	Ec. polytechnique	See SACL since 2015
ESAE	ISAE	
ESEC	Ec. supérieure des sciences économiques et commerciales	
ESMA	Ec. nationale supérieure de mécanique et d'aérotechnique	
ESTA	Ec. nationale supérieure de techniques avancées	See SACL since 2015
GLOB	Institut de physique du Globe	See USPC since 2015
HESA	HESAM	
IAVF	Institut agronomique, vétérinaire et forestier de France - Paris	
IEPP	Institut d'études politiques - Paris	
IMTA	Ec. nationale supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire	
INAL	Institut national des langues et civilisations orientales (INALCO)	See USPC since 2015
INAP	Institut national d'agronomie - Paris Grignon	See AGPT since 2007
IOTA	Institut d'optique théorique et appliquée - Palaiseau	SACL UPAS
ISAB	Institut national des sciences appliquées Val de Loire - Bourges	
ISAL	Institut national des sciences appliquées - Lyon	See LYSE since 2015
ISAM	Institut national des sciences appliquées - Rouen	See NORM since 2017
ISAR	Institut national des sciences appliquées - Rennes	
ISAT	Institut national des sciences appliquées - Toulouse	
MNHN	Museum d'histoire naturelle	
MTLD	Ec. nationale supérieure Mines-Télécom Lille Douai	
NSAI	Ec. nationale de la Statistique et de l'Analyse de l'Information - Rennes	
NSAM	SupAgro - Montpellier	
NSAR	Agrocampus Ouest - Rennes	
OBSP	Observatoire de Paris	See PSLE since 2016
ONIR	Ec. nationale vétérinaire - Nantes	
ORLE		
PSLE(=UPSL)	Paris Sciences et Lettres (ComUE)	Creation 2016
TELB	Ec. nationale supérieure des TelecompolfBretagne - Brest	See IMTA since 2017
TELE	Institut national des télécommunications	See SACL since 2015

Table B10: Higher Education Establishment

All the code of the universities associated with their name and the evolution of their code over the years. We focus on the period 1988 to 2021, any changes and code that appears before or after are taken into account. If the description is empty, it means that there is no change during the period. * EIAA+ENGR+INAP merged to become AGPT in 2007 we use one code for the three. ** Change code in 2020