Missing Women in Research

[Most recent version here]

Aliénor Bisantis*, Yann Bramoullé, Roberta Ziparo†

December 3, 2025

ABSTRACT.

We provide the first comprehensive analysis of gender differentials in academic trajectories following PhD graduation. Focusing on the universe of Phd graduates in France between 1988 and 2021, we show that raw estimates of gender gaps suggest that women publish less both on the extensive and intensive margin in STEM related fields but are more likely to publish in Biology and Earth Science and Social Sciences. However, controlling for discipline-time-university fixed-effects reveal that women are less likely to publish across all fields and career stage. Our estimates suggest that in the 25 years period from 1988 to 2012, 4,853 female PhD graduates are missing from Academia.

Keywords: Gender Inequality, Career Trajectory, Graduate Education

JEL: I23, J16, J24

*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

Despite improvement in the past 40 years, women are still severely underrepresented in academia, as documented in the *UNESCO Science Report*. In France, the proportion of female University faculty has not reached 30 % yet. This under-representation appears to persist despite the fact that women are now, on average, more educated than men: in 2017, 55 % of French graduates in higher education were women. Is just a question of delay, then? Will the proportion of female faculty track, with a lag, the proportion of women among higher education graduates? Or are there some significant barriers along the way which prevent, somehow, highly educated women to become successful academics? How does the answer to these questions vary across academic disciplines and time? Even though the literature has made some significant progress on these key questions, we still lack comprehensive answers.

In our study, we focus on a critical stage: the transition from a PhD to an academic career. We provide the first comprehensive analysis of gender differentials in academic trajectories following PhD graduation. To do so, we build a novel database by combining data on the universe of PhD graduations in France between 1988 and 2021 from *theses.fr* with data on the universe of academic publications until 2022 from *Scopus*. We then provide a detailed anatomy of the academic gender gaps for an entire country, across all academic disciplines, and over 34 years.[1] Our final sample has information on the academic trajectories of 335,796 PhD graduates. This is, to our knowledge, the first investigation of academic gender gaps at this scale.

We develop our analysis is several stages. We first document how the proportion of women among PhD graduates evolved over time and across disciplines. Encouragingly, female representation increased significantly in all disciplines. Women remain severely underrepresented, however, in *STEM (Sciences, Technology, Engineering, and Mathematics)*. By contrast, women are now slightly overrepresented among PhD graduates in *Humanities and Law*, *Biological and Earth Sciences*, and *Social Sciences*. Do STEM, then, have worse academic gender gaps than disciplines which have reached PhD gender parity?

Our main analysis focuses on PhD graduates' academic trajectories. We have information on the name, gender, discipline, year of defense, university affiliation and academic publications of each candidate, as well as of their PhD supervisor(s).[2]

[1] We had to remove data from one discipline out of twenty two because of a misclassification issue, see Section 2 for details.

[2] For each publication, we have information on year of publication, publication type and outlet, and names and affiliations of authors.

Overall, 48 % of PhD graduates do not publish any academic publication. This motivates separate analyses of the extensive and intensive margins of research, as well as on quality of research. We thus look at both the likelihood to publish at least one publication and the number and impact of publications conditional on publishing. We also distinguish two career stages: early career, until 4 years post-graduation, and mid-career, between 5 and 10 years post-graduation. We run standard reduced-form regressions over the full sample as well as over subsamples of the four main fields of research. In our main specification we include discipline-year-university fixed effects, which control for discipline-specific unobserved determinants of academic productivity that are time-varying and university-specific.

Overall, we document the presence of persistent negative gender gaps in academic trajectories. These gender gaps are generally quantitatively significant across all disciplines. Strikingly, quality is worse in Biological and Earth Sciences, a discipline with a better representation of women among PhD graduates than in STEM. In more detail, we find that women are less likely than men to ever publish across all disciplines, and this is true both at early and mid-career. Results at the intensive margin are consistent with those at the extensive one. Across all fields and career stages, we find that publishing women publish less publications than publishing men. Furthermore, in Biological and Earth Sciences this negative gender gap is aggravated: the average publication impact is lower for publishing women than for publishing men.[3]

Including year-field-university fixed effects leads to a worsening of the estimated gender gaps. This indicates that some male-biased sorting is operating in time and space. This could capture, for instance, differential access to better departments. By contrast, including supervisors' controls has little impact on the estimated gender gaps. This shows that gender gaps are essentially not explained by differential access to better supervisors within departments, once differential access to better departments is controlled for.

We further estimate how these gender gaps evolved over time. While they appear to have been remarkably stable on the extensive margin across disciplines, in Biological and Earth Sciences we find an interesting pattern: the number of publications has increased while average quality has deteriorated over time.

Finally, we use our regressions to estimate the number of missing women in academia, building on studies of missing women in poor countries, see Sen (1990), Coale (1991), Anderson and Ray

[3]In all the other disciplines, there is no discernible difference in average publication impact between publishing men and publishing women.

(2010). We compute the number of women who would become academically active if the barriers affecting women in the post-PhD transition were removed. Based on the field-specific regressions of the mid-career extensive margin, we estimate that 4,853 women are missing in academia in France across all disciplines in the 25 years period from 1988 to 2012. This corresponds to 5,7% of the population of female PhD graduates and to 18% of the population of non-publishing female PhD graduates. The large majority of missing women are in STEM. Note that these numbers are computed by conditioning on the existing population of PhD graduates. They would be larger if we also accounted for gender imbalances at the PhD level, a computation we will perform in future research.

Our analysis contributes to the literature on gender and academia. Earlier studies have documented the importance and persistence of the underrepresentation of women in academia, see Ginther and Kahn (2004), Ceci (2011), Ceci (2014), Meyer et al. (2015), Huang (2020). Many studies focus on STEM, where this underrepresentation is particularly severe. Other studies have documented the fact that men appear to be more productive than women in research, and notably in economics (Ginther and Kahn, 2004; Barbezat, 2006; McDowell et al., 2006; Ductor et al., 2023; Conley et al., 2016), science (Patsali et al., 2024; Stephan and Levin, 1997), and medicine (Rachid et al., 2021).

We also contribute to the literature on career paths in academic jobs. Recent works have identified several barriers to carrier promotion of women in economics linked to recognition of group work, citations, fellowships, writing criteria (Sarsons, 2017; Eberhardt et al., 2023; Card et al., 2019, 2022; Hengel, 2022). It has also been shown that women are substantially underrepresented in senior position in Economics (Bosquet et al., 2019). We contribute to this literature analyzing publication gaps at several carrier stages and across disciplines.

A few studies, combine data on PhD graduations and publications, as we do. Existing studies, however, focus on specific disciplines and shorter time scales, and do not systematically study gender gaps. Gaule and Piacentini (2018) analyze the academic trajectories of PhD graduates in Chemistry in the US who defended between 1999 and 2008. They find a strong impact of the gender match between a PhD graduate and their supervisor. Conley and Önder (2014) study the research productivity of PhD graduates in economics in the US and Canada who defended between 1986 and 2000. They find that the academic rank of the department from which a PhD student graduates provides a surprisingly poor predictor of the student's research success.

Corsini et al. (2022) analyze the determinants of PhD graduates' scientific productivity, based on the population of PhD graduates in France in STEM over the period 2000 - 2014. Patsali et al. (2024) analyze how research independence from the PhD supervisor(s) affects academic outcomes for a similar population of PhD graduates in France in STEM over the period 1995 - 2013.

Our main contributions with respect to these studies are the focus and scale of our analysis. We provide the first analysis of this kind focusing on gender gaps and covering all academic disciplines and over more than 30 years. We look at three main dimensions of academic productivity - probability to ever publish, number and impact of publications - at both early and mid-career. We provide a detailed anatomy of the gender gaps and of how they vary across fields and time. This notably allows us to understand how STEM - and economics - compare with other less studied disciplines.

The remainder of the paper is organized as follows. We describe the data in Section 2. We provide key descriptive statistics in Section 3. We present our empirical strategy and our main findings in Section 4. We compute numbers of missing women in Section 5. We conclude in Section 6.

2 Data

We combine data from two sources: *theses.fr* on PhD graduations and *Scopus* on academic publications. We next describe the data and our sampling restrictions.

2.1 PhDs

We retrieved data from *theses.fr* on all PhD theses defended in French universities from 1988 to 2021. For each PhD thesis, we have information on discipline of study, defense year, university affiliation, and first and last name of the PhD graduate and the supervisor(s). *theses.fr* is a centralized, public platform that collects data on all PhD graduations in the French academic system. French universities have the legal obligation to report information on PhD graduations. While some spelling errors and reporting delays are unavoidable, this database is considered as being, overall, comprehensive and reliable.[4] The initial sample contains 407,260 observations.

[4]We chose 2021 as the last year to minimize problems of reporting delays. See https://theses.fr/ for more information on data collection and on the underlying institutional arrangement.

We focus on PhD theses with one or two supervisors and with non-missing information on discipline and names, reducing the sample to 397,536 observations.[5]

We distinguish 22 academic disciplines, and exclude theses defended in *Health and Medical Sciences*, because of a misclassification problem specific to this discipline.[6] This further reduces the sample to 340,073 observations. We aggregate the 21 remaining disciplines in 4 broad fields of research: *Humanities and Law*; *Biological and Earth Sciences*; *Science, Technology, Engineering, and Mathematics (STEM)* and *Social Sciences*. In our empirical analysis below, we estimate regressions on the overall sample as well as separate regressions per field. We also pay special attention to *Economics*.

One issue with university affiliations is that France has seen some significant institutional evolution in the past twenty years, including a number of university mergers. We keep track of these evolutions and notably distinguish between universities before and after mergers, see Section C in the Online Appendix for details.

2.2 Gender

We identify the gender of PhD graduates and PhD supervisors using well-established methods based on first names. In a first step, we use data from the French statistical institute, *INSEE*, on the numbers of boys and girls born in France between 1940 and 2020 with a given first name. We associate a gender with a first name when at least 95 % of individuals with this name share the same gender. In a second step, and to broaden coverage to non-French PhD graduates and supervisors, we use similar data from Australia, Canada, Spain, Sweden, the UK, and the US. In the end, we identify the gender of 93% of the PhD graduates and 95% of the supervisors.[7]

2.3 Publications

We next assemble information on academic publications of PhD graduates and PhD supervisors. To do so, we extract bibliometric data from *Scopus* on all publications until 2022 where the first name and last name of one author are identical to the first name and last name of a PhD graduate or a PhD supervisor. The risk of mismatch is still significant, however, when the first

[5]Theses with three supervisors or more represent less than 2% of the original sample.

[6]To become Medical Doctors, French medical students must defend a practical thesis ("Thèse d'exercice") at the end of their studies. These practical theses have very different requirements from PhD theses. However, theses.fr did not distinguish between the two kind of theses until the 2000s, see Section C in the Appendix for details.

[7]See Section C in the Online Appendix for more details on gender association.

and last names are very common, such as "Sarah Lopez" or "Philippe Morin". To address this issue, we remove observations with overly common names, leading to a final sample of 335,796 PhD theses. For each academic publication, we have information on publication type, journal, year of publication, authors and their affiliations.

Table A1 in the Online Appendix describes the number of PhD graduates per discipline in our final sample. It is composed of 335,796 PhD graduates overall: 20% in Humanities and Law, 17 % in Biological and Earth Sciences, 46 % in STEM and 17 % in Social Sciences. Among PhD graduates in Social Sciences, 19 % have graduated in Economics.

2.4 Early and mid-career

In our analysis of academic trajectories, we distinguish two periods: early career and mid-career. We define early career as every year until the fourth year after graduation. E.g., for a researcher graduating in 2000, publications in every year up to 2004 are counted in early career. We define mid-career as the period between five and ten years after graduation, both years included. E.g., for a researcher graduating in 2000, publications in every year between 2005 and 2010 are counted in mid-career. To study academic trajectories post-graduation, we need to trim the last years of our sample - recall that we observe academic publications until 2022. The early career sample is thus composed of all PhD graduates who graduated between 1988 and 2018, corresponding to 281,419 observations. The mid-career sample is composed of all PhD graduates who graduated between 1988 and 2012, corresponding to 220,935 observations.

2.5 Research outcomes

We analyze both the extensive and the intensive margin of academic research. We consider three outcomes: *Any Publi*, *Publications* and *Impact*. *Any Publi* is a binary variable, equal to one if the PhD graduate has at least one academic publication referenced in *Scopus* over the period of interest and to zero otherwise. For PhD graduates for whom *Any Publi*=1, we then compute two measures of research productivity. *Publications* is equal to the fractional publication count, that adds up, over all publications, one over the number of authors. Dividing by the number of authors is a standard way to allocate credit for shared authorship. Our results are robust to using the undivided publication count instead, see Section @ of the Online Appendix. Our third outcome, *Impact*, measures the average impact of the PhD graduate's publications. We build this measure from the Article Influence Score (AIS) of the publication outlet, a classical

measure of impact factor, see Bagues et al. (2017). *Impact* is then equal to the AIS-weighted fractional publication count divided by the fractional publication count.

Formally, consider PhD graduate i with set of publications P_i. Given publication $p \in P_i$, let n_p denote the number of authors of the publication and AIS_p the Article Influence Score of the publication outlet. Then,

$$\begin{aligned} Any\ Publi &= 1 \text{ if } P_i \neq \emptyset \text{ and } 0 \text{ if } P_i = \emptyset \\ Publications &= \sum_{p \in P_i} \frac{1}{n_p} \\ Impact &= \frac{\sum_{p \in P_i} \frac{AIS_p}{n_p}}{\sum_{p \in P_i} \frac{1}{n_p}} \end{aligned}$$

In our regressions below, we further standardize *Publications* and *Impact* to have mean 0 and standard deviation 1 within discipline and career stage.

3 Descriptive statistics

We now provide some key descriptive statistics of the data. We first describe how the proportion of female PhD graduates varied across time and field. We then discuss systematic differences in research productivity across fields.

3.1 Female representation among PhD graduates

Figure 1 depicts how the proportion of female PhD graduates (in solid red) and the total number of PhD graduates (in dotted blue) vary over time in the full sample, Humanities and Law, Biological and Earth Sciences, STEM, Social Sciences and Economics. The red dotted line corresponds to a 50% proportion of female PhD graduates.

The evolution of female representation displays a similar pattern in Humanities and Law, Biological and Earth Sciences, and Social Sciences. Women were underrepresented in the 1990's in these three fields, with a share lying between 30 and 40%. Female representation then improved, and these three fields reached gender parity in the 2000's. Women then became slightly overrepresented among PhD graduates in the three fields in the 2010's. By contrast, women are still severely underrepresented among STEM PhD graduates, with the female share in the best year equal to 32% in 2017. Female representation did increase over time in STEM, although at a lower rate than in the other fields. Overall, Figure 1 reveals important differences

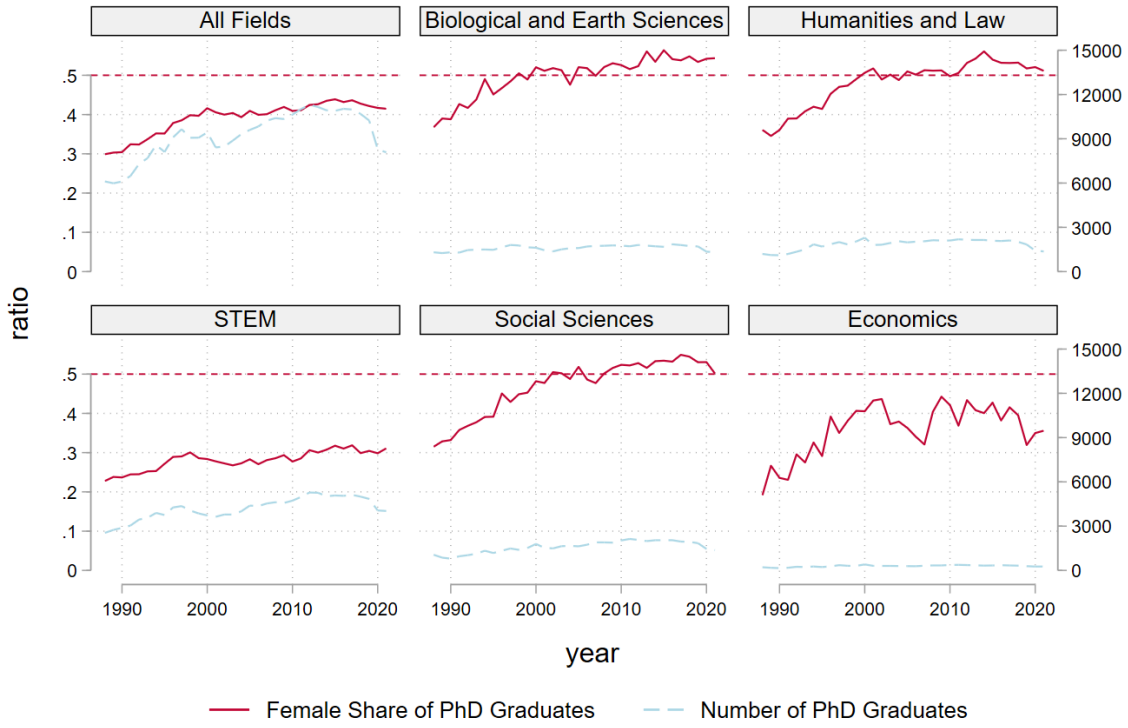


Figure 1: Share of female PhD graduates

in female representation between STEM and the other academic fields.

Figure **??** in the Online Appendix depicts how the share of female PhD graduates and the number of PhD graduates vary over time in each of the 21 academic disciplines. Remarkably, female representation increased over time in all disciplines, with the possible exception of Mathematics. Within Social Sciences, Economics is the only discipline that never reached gender parity. Following good progress in the 1990's, from a low female share of 20% in 1988, female representation among PhD graduates in Economics in France appears to fluctuate below a 45% ceiling.

3.2 Differences in research productivity across fields

Table A7 in the Appendix describes how various measures of research productivity vary across fields and career stages. We report the proportion of publishing PhD graduates for each subsample and the average of the number of undivided publications, of the number of coauthors, and of our two main measures, *Publications* and *Impact*, among publishing PhD graduates.

We see that fields have very different baseline levels of publications, coauthorship and impact, reflecting different academic practices and norms. The likelihood to ever publish, the numbers

of undivided publications and of coauthors at both career stages are all consistent with the following ordering: Humanities and Law < Social Sciences < Biological and Earth Sciences < STEM. Thus, the proportions of PhD graduates with at least one publication four years after graduation are, respectively, 14% < 23% < 46% < 60%. And a publishing PhD graduate at early career has, respectively and on average, 2.4 < 3.1 < 8.3 < 9.1 undivided publications and 1.4 < 2.3 < 12.3 < 17.2 coauthors. Interestingly, differences among fields are much lower, and the previous ordering is not preserved, when looking at fractional publications counts. Allocating credit equally among coauthors, the respective average numbers of *Publications* are 1.78, 1.75, 1.43 and 1.98.

In terms of impact, the two indices at early and mid-career are consistent with a slightly different ordering, Humanities and Law < Social Sciences < STEM < Biological and Earth Sciences. For instance, the average *Impact* at early career for publishing PhD graduates is equal to 0.17 in Humanities and Law, 0.35 in Social Sciences, 0.59 in STEM and 1.63 in Biological and Earth Sciences. While researchers publish less in Biological and Earth Sciences than in STEM, they tend to publish in more influential outlets.

Economics appears to be fairly representative of other Social Sciences in terms of probability to publish and numbers of publications and coauthors conditional on publishing. Publications in Economics have more impact, however, than in other Social Sciences. Average *Impact* at early career for publishing PhD graduates in Economics is equal to 0.68.

The presence of systematic differences in research productivity across fields and disciplines motivates the normalization of *Publications* and *Impact* in our regressions below. More precisely, we compute the mean and standard deviation of *Publications* and *Impact* among all publishing PhD graduates for each of the 21 disciplines and 2 career stages. We then standardize each outcome by subtracting the mean and dividing by the standard deviation. One advantage of this normalization is that it makes the gender gap estimates directly comparable across fields.

4 Empirical analysis

We now analyze differences in academic trajectories between female and male PhD graduates. We look at the three research outcomes, *Any Publi* and, if *Any Publi*=1, *Publications* and *Impact*, at both career stages, overall, for each of the four fields of research, and for Economics. We first compute unconditional estimates, regressing outcome on gender of the PhD graduate

only. These provide differences in average outcomes between female and male PhD graduates. In our main specification, we then include university-discipline-year fixed effects and controls related to supervision.

@One sentence on results@

We first provide raw estimates, only controlling for gender of the PhD graduate. This is equivalent to looking at differences in averages, and yields key descriptive statistics on research outcomes. In our second and preferred specification, we include discipline-year-university fixed effects and we add controls related to supervision: whether the thesis is in cosupervision, and the gender and research productivity of the supervisor(s). Overall, we find strong and negative gender gaps, with some heterogeneity across fields.

4.1 Unconditional gender gaps

Table 1 presents the estimates of gender differences in academic outcomes in the overall sample (first column), in the four fields (second to fifth column), and in Economics (last column).

On the overall sample, we see that women have a lower probability to ever publish and publishing women publish less papers than publishing men. These negative gender gaps are present at early and mid-career and are larger in magnitude at mid-career. These gaps are also quantitatively significant. The early career gender gap at the extensive margin is -7 p.p., from an average probability to ever publish for men equal to 45%. The early career gender gap on publications is -0.19 standard deviation. By contrast, we do not detect much difference in impact. If anything, publications by women have a slightly larger impact.

These average gaps mask some significant heterogeneity across fields. Women are *more* likely to ever publish than men at early career in Biological and Earth Sciences and Social Sciences. This initial advantage is partly dissipated at mid-career. It remains positive, but lower in magnitude, in Biological and Earth Sciences and becomes negative, although statistically insignificant, in Social Sciences. In Humanities and Law and STEM, the gender gap at the extensive margin is negative at both career stages and larger in magnitude at mid-career, especially in STEM.

By contrast, the gender gap is consistently negative at the intensive margin on publications. Publishing women publish less publications than publishing men in the four fields and at the two career stages. In STEM, Biological and Earth Sciences and Social Sciences, these gender gaps are aggravated at mid-career compared to early career. By contrast, the negative gender

Table 1: Unconditional gender gaps

	All Fields	Humanities and Law	Biological and Earth Sc.	STEM	Social Sciences	Economics
			Before t+4			
% Female PhD graduates	39%	49%	50%	28%	48%	37%
Any Publi	**-0.0733*****	**-0.0137*****	**0.0655*****	**-0.0419*****	**0.0147*****	**-0.00955**
	(0.00188)	(0.00287)	(0.00445)	(0.00302)	(0.00380)	(0.00938)
Av. Male Any Publi	0.448	0.146	0.425	0.613	0.226	0.261
Relative Gender Gap	-16%	-9%	15%	-7%	-6%	-4%
# PhD graduates	286814	58008	49859	129402	49545	9303
if Any Publi = 1						
Publications	**-0.190*****	**-0.217*****	**-0.244*****	**-0.236*****	**-0.155*****	**-0.209*****
	(0.00626)	(0.0268)	(0.0134)	(0.00812)	(0.0204)	(0.0460)
Impact	**0.00422**	**-0.0233**	**-0.0255***	**0.0277*****	**-0.0388****	**-0.0346**
	(0.00606)	(0.0227)	(0.0132)	(0.00815)	(0.0188)	(0.0432)
# PhD graduates *if Any Publi = 1*	120248	8054	22838	77807	11549	2399
			Between t+5 and t+10			
% Female PhD Graduates	38%	47%	48%	27%	46%	36%
Any Publi	**-0.0786*****	**-0.0179*****	**0.0141*****	**-0.0961*****	**-0.00493**	**-0.0212****
	(0.00205)	(0.00333)	(0.00488)	(0.00353)	(0.00433)	(0.0105)
Av. Male Any Publi	0.364	0.156	0.366	0.471	0.228	0.256
Relative Gender Gap	-22%	-11%	4%	-20%	-2%	-8%
# PhD graduates	221128	45430	39323	98854	37521	7274
if Any Publi = 1						
Publications	**-0.282*****	**-0.166*****	**-0.392*****	**-0.308*****	**-0.216*****	**-0.288*****
	(0.00864)	(0.0306)	(0.0175)	(0.0123)	(0.0241)	(0.0533)
Impact	**0.0132***	**0.00672**	**-0.0275***	**0.0487*****	**-0.0259**	**-0.0141**
	(0.00787)	(0.0248)	(0.0165)	(0.0114)	(0.0215)	(0.0498)
# PhD graduates *if Any Publi = 1*	73738	6687	14671	43922	8458	1806

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

gap is slightly lower in magnitude at mid-career in Humanities and Law. Economics displays slightly worse gender gaps in publications at both margins than in Social Sciences in general.

Gender gaps in impact are quantitatively modest, always lower in absolute value than 0.05 standard deviation. Publications by women have a higher impact than publications by men in STEM at both career stages. They have a lower, marginally statistically significant, impact in Biological and Earth Sciences and Social Sciences. Gender gap in impact is statistically not different from zero in Humanities and Law.

Overall, these results demonstrate the presence of strong gender effects in academic trajectories post-graduation and of significant heterogeneity across fields. Of course, unconditional gender differences could be explained by many factors, such as differential access to good universities or good PhD advisors, motivating the inclusion of controls in the regressions.

4.2 Conditional gender gaps

We now estimate conditional gender gaps. We estimate the following model, via OLS, over each field subsample.[8]

$$Y_{iudft} = \beta_{1f} Female_i + \beta_{2f} X_{it} + \mu_{udt} + \epsilon_{iudft} \quad (1)$$

where Y_{iudft} represents research outcome of PhD graduate i, who defended in year t, university u, discipline d and field f. $Female_i$ is a dummy variable equal to 1 if the PhD graduate is female and 0 if he is male. μ_{udt} is a university-discipline-year fixed effect and X_{it} are controls related to supervision.

μ_{udt} is a high-dimensional fixed effect, which captures university-discipline-specific time trends in a non-parametric manner. It controls for time-varying local factors that may affect PhD graduates' academic outcomes such as departments' academic quality and social capital. Identification then comes from comparing academic trajectories of women and men who defended their PhD the same year, in the same discipline and university. X_{it} include a dummy equal to 1 if the thesis is in cosupervision, a dummy equal to 1 if a supervisor is female, and a measure of academic productivity of the supervisors.[9]

We also estimate a version of this model with homogeneous parameters ($\beta_{1f} = \beta_1$, $\beta_{2f} = \beta_2$) over the full sample, as well as a version with discipline-specific gender effect (β_{1d}, β_{2d}) over the Economics subsample.

How are estimated gender gaps affected when controlling for university-discipline-year fixed effects and for supervision characteristics? Estimates are presented in Table 2.

Strikingly, gender gap estimates by field on the likelihood to ever publish are all worse than the unconditional estimates. Gaps that were positive, in favor of women, are now negative and statistically significant, in Biological and Earth Sciences and Social Sciences. Gaps that were negative are now larger in magnitude, in Humanities and Law and STEM. Gender gap estimates on the number of publications for publishing PhD graduates are also very robust. They remain negative and quantitatively and statistically significant for the 4 fields and 2 career stages. They become larger in magnitude in Humanities and Law and Social Sciences and are little affected

[8] Since we include fixed effects at the university-discipline-year level, we remove from the sample observations where only one PhD graduate graduated that year in that university and that discipline. These singleton observations represent about 2% of the overall sample.

[9] For each supervisor s with set of publications at time of defense P_s, we compute the AIS-weighted number of publications at time of defense, equal to $\sum_{p \in P_s} AIS_p$. We then standardize this measure to have standard deviation 1 within discipline. And if the PhD is cosupervised, we take the largest of the two supervisors' productivities.

in Biological and Earth Sciences and STEM. In Economics, gender gaps on publications are worse than the unconditional gaps, and are negative and statistically significant at the extensive and intensive margins and at early and mid-career. Gender gap estimates on impact remain, overall, quantitatively modest and also tend to worsen. In particular, gender gaps on impact in STEM at both career stages, that were positive before, are now close to zero and statistically non-significant.

This shows that a significant part of the unconditional gender gaps captures correlation between gender and the controls. This could notably capture selective sorting by gender across universities, disciplines and years. A worsening of the gender gaps when conditioning is consistent with *positive* female sorting, i.e., women being overrepresented in the university-discipline-years with better research outcomes. Such positive sorting happens, for instance, when women are better represented among PhD graduates of high-quality departments.

Table 2: Conditional gender gaps

	All Fields	Humanities and Law	Biological and Earth Sc.	STEM	Social Sciences	Economics
			Before t+4			
Any Publi	**-0.0316*****	**-0.0273*****	**-0.00918****	**-0.0534*****	**-0.0129*****	**-0.0337*****
	(0.00179)	(0.00297)	(0.00391)	(0.00308)	(0.00409)	(0.00979)
Av. Male Any Publi	0.428	0.145	0.467	0.613	0.227	0.266
Relative Gender Gap	-7%	-19%	-2%	-9%	-6%	-13%
# PhD Graduates	280778	56490	49226	127563	47499	8954
if Any Publi = 1						
Publications	**-0.225*****	**-0.280*****	**-0.220*****	**-0.228*****	**-0.182*****	**-0.239*****
	(0.00698)	(0.0347)	(0.0138)	(0.00875)	(0.0259)	(0.0561)
Impact	**-0.0166****	**0.00169**	**-0.0448*****	**-0.00704**	**-0.0349**	**-0.0150**
	(0.00670)	(0.0293)	(0.0134)	(0.00849)	(0.0237)	(0.0537)
# PhD graduates *if Any Publi = 1*	114827	6721	22569	75845	9692	2045
			Between t+5 and t+10			
% Female PhD Graduates	38%	47%	48%	27%	46%	36%
Any Publi	**-0.0610*****	**-0.0276*****	**-0.0509*****	**-0.0990*****	**-0.0279*****	**-0.0451*****
	(0.00212)	(0.00352)	(0.00457)	(0.00371)	(0.00476)	(0.0112)
Av. Male Any Publi	0.354	0.155	0.399	0.469	0.230	0.265
Relative Gender Gap	-17%	-18%	-14%	-21%	-12%	-18%
# PhD graduates	216142	44149	38786	97336	35871	6988
if Any Publi = 1						
Publications	**-0.318*****	**-0.222*****	**-0.378*****	**-0.314*****	**-0.265*****	**-0.343*****
	(0.00996)	(0.0390)	(0.0182)	(0.0135)	(0.0321)	(0.0669)
Impact	**-0.0138**	**-0.00587**	**-0.0531*****	**0.00758**	**-0.0365**	**-0.0385**
	(0.00888)	(0.0266)	(0.0171)	(0.0123)	(0.0273)	(0.0617)
# PhD graduates *if Any Publi = 1*	69071	5564	14437	42186	6884	1519

Notes: Discipline X Year X University FE + controls (CoSup std_max_total_AIS_supervisor female supervisor). Standard errors are reported in parentheses. Significance levels are defined as follows: $p < 0.1$ *, $p < 0.05$ **, $p < 0.01$ ***.

The analysis of female representation in Section 3 shows a clear distinction between STEM, where women are severely underrepresented, and other fields, where women are well represented. This naturally raises the question of whether gender gaps are worse in STEM. At the intensive margin, since *Publications* and *Impact* are standardized, we can directly compare estimates across fields. By contrast, estimates are not directly comparable across fields for *Any Publi* given the different baseline levels. We then compute the relative gender gap, equal to the estimated gender gap divided by the average likelihood to ever publish among men. We see that at early career, gender gaps on publications at the extensive and intensive margin are both worse in Humanities and Law than in STEM. At the extensive margin, gender gaps in Biological and Earth Sciences and Social Sciences are comparable in magnitude to gaps in STEM. At mid-career, gender gap at the extensive margin in STEM is worst among the 4 fields, but close in magnitude to the one in Humanities and Law. At the mid-career intensive margin, the gender gap in STEM is lower in magnitude than in Biological and Earth Sciences. Overall, the estimations reveal that the worse representation of women in STEM is not associated with worse gender gaps in academic trajectories.

To sum up, accounting for university-discipline-year fixed effects and supervision controls, we find that women are less likely to ever publish than men and that publishing women publish less than publishing men in the 4 fields of research and 2 career stages. Gender gap estimates are generally larger in magnitude than in the raw data, consistent with positive female sorting in space, discipline and time. Gender gaps in STEM, while bad, are not particularly worse than in the other fields of research.

4.3 Evolution of gender gaps over time

Since female representation has increased over time in all fields, a natural question is whether the gender gaps have also evolved.

We next analyze how these gender gaps evolved over time, focusing on mid-career.

We partition the 1988 - 2012 time window into six periods: five 4-year periods and one 3-year period. We rerun separate regressions by field for each of these six periods, including university-discipline-year fixed effects and supervision controls. Figure 2 depicts how estimated gender gaps vary over time for the four fields and three indicators.

The results reveal a limited evolution of these estimates over time, with two exceptions. In Biological and Earth Sciences we find an interesting pattern: the number of publications has

increased while average quality has deteriorated over time. STEM, that is consistently worse in the probability of publishing and as bad in the number of publications, appears to improve over the last period of time in terms of average quality.

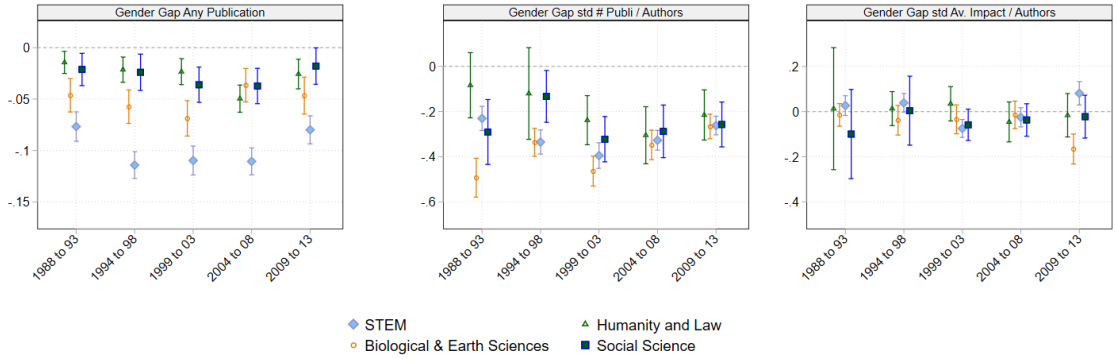


Figure 2: Between t+5 and t+10

5 Missing women

In this Section, we compute the number of "missing women in academia". More precisely, we focus on the mid-career sample, formed of PhD graduates in French universities in all disciplines except Health and Medical Sciences from 1988 to 2012. Our aim is to compute the number of women from this sample who "should" have been academically productive but were not.

Any computation of missing women relies on counterfactual scenarios, as discussed in the literature on missing women in poor countries (e.g. Sen (1990), Coale (1991), Anderson and Ray (2010)). Here, we build these counterfactual scenarios from the models estimated in the previous section. More precisely, we compute the number of missing women in academia from the field-specific models of the mid-career extensive margin that accounts for university-discipline-year fixed effects.

Consider a female PhD graduate i who defended in year t, university u and field f. Let g_i denote the female dummy. From the estimated model, we obtain two probabilities. First, the actual estimated probability that i publishes at least one publication between 5 and 10 years post-graduation, $\hat{p}(g_i = 1)$. And second, the predicted probability of publishing if i had been a male PhD graduate defending in the same year, university and field, $\hat{p}(g_i = 0)$.

The estimated number of missing women n_{MW} in a field is then simply the sum of the differences in probabilities over all PhD graduates from the field:

$$n_{MW} = \sum_i \hat{p}(g_i = 0) - \hat{p}(g_i = 1).$$

Given our linear specification, this is simply equal to the estimated gender gap at the extensive margin, $\hat{\beta}_f$, times the number of female PhD graduates in the field.

Table 3 provides estimates of the number of missing women in academia and of the number of missing publications by publishing women.

Table 3: Missing Women and Missing Papers - Computations

Fields	n_W	$n_{W\|P=1}$	$-\hat{\beta}$ (Any Publi)	$-\hat{\beta}$ (# Publi)	n_{MW}	Conf Inter. 95% (n_{MW})	n_{MP}	Conf Inter. 95% (n_{MP})
Humanities and Law	21,475	2,958	0.0276	0.222	593	[445; 741]	657	[431; 883]
Biological and Earth Sc.	19,043	7,243	0.0509	0.378	969	[799; 1140]	2,738	[2479; 2996]
STEM	27,154	10,173	0.0990	0.314	2,688	[2491; 2886]	3,194	[2925; 3463]
Social Sc.	17,354	3,866	0.0279	0.265	484	[322; 646]	1,024	[781; 1268]
Economics	2,641	620	0.0451	0.343	119	[61; 177]	213	[131; 294]
All Fields	85,026	24,240	0.0610	0.318	4,853	[4468; 5238]	7,825	[7433; 8217]

$\hat{\beta}$: female estimates for Y between t+5 and t+10 (Any publication or Number of publication) from eq. 1. See appendix B. n_W : Number of women who defended their PhD (before 2013). $n_{W|P=1}$: Number of women who have at least one publication between t+5 and t+10

Overall, we estimate that in the 25 years period from 1988 to 2012, 4,853 female PhD graduates out of the 24,682 who did not publish at mid-career should have been academically active. This represents approximately 18 % of the population of non-publishing female PhD graduates and 6 % of the full population of female PhD graduates. The bulk of these numbers is coming from STEM, both because of a direct size effect - STEM is a larger field and has more female PhD graduates than the other fields, despite the fact that women are less well-represented in STEM - and of the fact that the estimated negative gender gap is larger in STEM.

Note that this methodology has a number of limitations. First, these computations are performed conditionally on the actual population of PhD graduates. This means that the number of missing women in a field with a lower proportion of female PhD graduates is mechanically lower than in a field with a higher proportion of female PhD graduates. Accounting for gender imbalances in the PhD population is an important direction for future research - and would increase the estimated number of missing women in academia given their persistent underrepresentation in STEM. Second, these computations rely on the assumption that there is no predetermined limit on the number of PhD graduates in a field and year who can end up being academically active. An alternative scenario would be to assume that this number is fixed - which would decrease the estimated number of missing women.

6 Conclusion and next steps

In this paper, we provide the first systematic analysis of gender gaps in academic trajectories post-PhD graduation for an entire country, across 30 years and all academic disciplines. Female representation in the PhD population has improved over time in all fields. While women are still severely underrepresented among PhD graduates in STEM, women are now slightly overrepresented in all fields outside of STEM. Despite these encouraging increases in representativity, however, we document the presence of persistent negative gender gaps in academic trajectories.

We find that women are less likely than men to ever publish in STEM, Humanities and Law, and Social Sciences. Interestingly, women are more likely than men to ever publish in the first stage of their career in Biological and Earth Sciences. This advantage is dissipated by mid-career, characterized by gender equality in Biological and Earth Sciences. Results on the intensive margin are even more clear-cut. Across all fields and career stages, publishing women publish less publications than publishing men. In STEM, this is partly counterbalanced by the fact that women have higher average publication impact than men. In STEM, women publish less publications but of higher impact. In Biological and Earth Sciences and Social Sciences, however, this negative gender gap is aggravated. Women publish less publications and of lower impact.

Our preferred econometric specification includes university-field and year-field fixed effects. Gender gaps worsen a bit when including the fixed effects, indicating that some male-biased sorting is operating in time and space. By contrast, including controls for supervision characteristics - the number, gender and productivity of PhD supervisors - has very little impact, which shows that these gender gaps are not explained by differential access to better supervisors. We use our regressions to compute the number of missing women in academia, and estimate that 15 % of the population of non-publishing female PhD graduates, corresponding to 4 % of the population of female PhD graduates, should have been academically active.

There are a number of further things we want to do before finalizing the paper. First, we are currently obtaining and cleaning the data on coauthors. This will allow us to assess the robustness of our analysis at the intensive margin to measures that account for the number of coauthors. We will also investigate the impact of supervisors' networks on the academic trajectories of PhD graduates and gender differences in coauthorships (see, e.g., **?**). Second, we would like to study the impact of time-varying university controls, such as the female share

among supervisors during the thesis period, the overall research quality of the university, and the overall number of PhD students. This will allow us to better understand how the work environment during PhD may affect academic trajectories post-graduation. Third, we recently obtained comprehensive data on "qualification" at the junior and senior level. Qualification is a federal-level exam which is required to apply to university positions in France. We intend to combine qualification data with thesis and publication data. This will allow us to cross-validate our current analysis, and to obtain further insights into the determinants of academic trajectories.

References

Anderson, Siwan and Debraj Ray, "Missing Women: Age and Disease," *The Review of Economic Studies*, 2010, *77* (4), 1262–1300.

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.

Benveniste, Stéphane, "Like Father, Like Child: Intergenerational Mobility in the French Grandes Écoles throughout the 20 th Century," 2023.

Bosquet, Clément, Pierre-Philippe Combes, and Cecilia García-Peñalosa, "Gender and promotions: Evidence from academic economists in France," *The Scandinavian Journal of Economics*, 2019, *121* (3), 1020–1053.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri, "Are Referees and Editors in Economics Gender Neutral?*," *The Quarterly Journal of Economics*, 11 2019, *135* (1), 269–327.

_ , _ , _ , and _ , "Gender differences in peer recognition by economists," *Econometrica*, 2022, *90* (5), 1937–1971.

Coale, Ansley J., "Excess Female Mortality and the Balance of the Sexes in the Population: An Estimate of the Number of "Missing Females"," *Population and Development Review*, 1991, *17* (3), 517–523.

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.

_ and _ , "The Research Productivity of New PhDs in Economics: The Surprisingly High Non-success of the Successful," *Journal of Economic Perspectives*, September 2014, *28* (3), 205–16.

Corsini, Alberto, Michele Pezzoni, and Fabiana Visentin, "What makes a productive Ph.D. student?," *Research Policy*, 2022, *51* (10), 104561.

Ductor, Lorenzo, Sanjeev Goyal, and Anja Prummer, "Gender and collaboration," *Review of Economics and Statistics*, 2023, *105* (6), 1366–1378.

Eberhardt, Markus, Giovanni Facchini, and Valeria Rueda, "Gender differences in reference letters: Evidence from the economics job market," *The Economic Journal*, 2023, *133* (655), 2676–2708.

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.

Hengel, Erin, "Publishing while female: Are women held to higher standards? Evidence from peer review," *The Economic Journal*, 2022, *132* (648), 2951–2991.

J., Williams W. M. Ceci S., "Understanding current causes of women's underrepresentation in science.," *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 2011.

J., Gates A. J. Sinatra R. Barabási A.-L. Huang, ""Historical Comparison of Gender Inequality in Scientific Careers across Countries and Disciplines."," *Proceedings of the National Academy of Sciences*, 2020.

J., Ginther D. K. Kahn S. Williams-W. M. Ceci S., "Women in Academic Science: A Changing Landscape," *Psychological science in the public interest : a journal of the American Psychological Society*, 2014.

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.

Meyer, Meredith, Andrei Cimpian, and Sarah-Jane Leslie, "Women are underrepresented in fields where success is believed to require brilliance," *Frontiers in Psychology*, 2015, *6.*

Patsali, Sofia, Michele Pezzoni, and Fabiana Visentin, "Research independence: drivers and impact on PhD students' careers," *Studies in Higher Education*, 02 2024, pp. 1–24.

Rachid, Elza, Tania Noureddine, Hani Tamim, Maha Makki, Sally Naalbandian, and Christiane Al-Haddad, "Gender disparity in research productivity across departments in the faculty of medicine: a bibliometric analysis," *Scientometrics*, 2021, *126*, 4715–4731.

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

Sen, Amartya, "More Than 100 Million Women Are Missing," *The New York Review of Books*, 12 1990.

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

A Figures and Tables

Table A1: List of the 21 disciplines

Disciplines	Observations
Humanities and Law	**67,084**
History and Archaeology	12,486
Journalism, Librarianship and Curatorial Studies	1,863
Language and Culture	20,308
Law, Justice and Law Enforcement	18,592
Philosophy and Religion	6,292
The Arts	7,543
Biological and Earth Sciences	**57,109**
Biological Sciences	42,786
Earth Sciences	14,323
Sciences, technology and Engineering	**153,103**
Agricultural, Veterinary and Environmental Sciences	2,419
Chemical Sciences	22,369
Engineering and technology	50,864
Information, computing and Communication Sciences	23,752
Mathematical Sciences	11,157
Physical Sciences	42,542
Social Sciences	**58,500**
Architecture, Urban Environment and Building	1,485
Behavioural and Cognitive Sciences	14,509
Commerce, Management, Tourism and Services	8,165
Economics	10,983
Education	3,738
Policy and Political Science	4,633
Studies in Human Society	14,987
Entire Sample	**335,796**

Table A2: P-values for Pairwise Differences in Female PhD Coefficient of Any publication between t+5 and t+10

	(1) All	(2) Humanity	(3) Bio	(4) STEM	(5) Social	(6) Econ
(1) All	–	0.000***	0.045**	0.000***	0.000***	0.162
(2) Humanity and Law	0.000***	–	0.000***	0.000***	0.960	0.140
(3) Biological Sciences	0.045**	0.000***	–	0.000***	0.000***	0.628
(4) STEM	0.000***	0.000***	0.000***	–	0.000***	0.000***
(5) Social Sciences	0.000***	0.960	0.000***	0.000***	–	0.161
(6) Economics	0.162	0.140	0.628	0.000***	0.161	–

Notes: P-values for tests of equality of female_phd coefficient between disciplines. H0: $\beta_{\text{row}} - \beta_{\text{column}} = 0$. Lower values indicate stronger evidence of differences. *** p<0.01, ** p<0.05, * p<0.10.

Table A3: P-values for Pairwise Differences in Female PhD Coefficient of std # Publi/authors between t+5 and t+10

	(1) All	(2) Humanity	(3) Bio	(4) STEM	(5) Social	(6) Econ
(1) All	–	0.017**	0.004***	0.805	0.114	0.712
(2) Humanity and Law	0.017**	–	0.000***	0.025**	0.392	0.117
(3) Biological Sciences	0.004***	0.000***	–	0.005***	0.002***	0.620
(4) STEM	0.805	0.025**	0.005***	–	0.159	0.670
(5) Social Sciences	0.114	0.392	0.002***	0.159	–	0.292
(6) Economics	0.712	0.117	0.620	0.670	0.292	–

Notes: P-values for tests of equality of female_phd coefficient between disciplines. H0: $\beta_{\text{row}} - \beta_{\text{column}} = 0$. Lower values indicate stronger evidence of differences. *** p<0.01, ** p<0.05, * p<0.10.

Table A4: P-values for Pairwise Differences in Female PhD Coefficient of std Av impact between t+5 and t+10

	(1) All	(2) Humanity	(3) Bio	(4) STEM	(5) Social	(6) Econ
(1) All	–	0.778	0.041**	0.158	0.429	0.692
(2) Humanity and Law	0.778	–	0.136	0.647	0.422	0.628
(3) Biological Sciences	0.041**	0.136	–	0.004***	0.607	0.819
(4) STEM	0.158	0.647	0.004***	–	0.141	0.464
(5) Social Sciences	0.429	0.422	0.607	0.141	–	0.977
(6) Economics	0.692	0.628	0.819	0.464	0.977	–

Notes: P-values for tests of equality of female_phd coefficient between disciplines. H0: $\beta_{\text{row}} - \beta_{\text{column}} = 0$. Lower values indicate stronger evidence of differences. *** p<0.01, ** p<0.05, * p<0.10.

Table A5: P-values for Female PhD Coefficient Across Time Periods

	(1) Any Publication	(2) Std # Publi/authors	(3) Std Av. Impact/authors
Humanities and Law			
Period 1 vs 2 ($\leq$ 1993 vs 1994-1998)	0.477	0.809	0.999
Period 2 vs 3 (1994-1998 vs 1999-2003)	0.859	0.400	0.741
Period 3 vs 4 (1999-2003 vs 2004-2008)	0.018**	0.511	0.258
Period 4 vs 5 (2004-2008 vs 2009-2013)	0.044**	0.381	0.711
Biological and Earth Sciences			
Period 1 vs 2 ($\leq$1993 vs 1994-1998)	0.428	0.016**	0.643
Period 2 vs 3 (1994-1998 vs 1999-2003)	0.432	0.021**	0.948
Period 3 vs 4 (1999-2003 vs 2004-2008)	0.026**	0.041**	0.711
Period 4 vs 5 (2004-2008 vs 2009-2013)	0.495	0.115	0.006***
STEM			
Period 1 vs 2 ($\leq$1993 vs 1994-1998)	0.001***	0.024**	0.733
Period 2 vs 3 (1994-1998 vs 1999-2003)	0.705	0.207	0.001***
Period 3 vs 4 (1999-2003 vs 2004-2008)	0.944	0.122	0.163
Period 4 vs 5 (2004-2008 vs 2009-2013)	0.008***	0.073*	0.009***
Social Sciences			
Period 1 vs 2 ($\leq$1993 vs 1994-1998)	0.848	0.160	0.497
Period 2 vs 3 (1994-1998 vs 1999-2003)	0.419	0.042**	0.542
Period 3 vs 4 (1999-2003 vs 2004-2008)	0.934	0.712	0.725
Period 4 vs 5 (2004-2008 vs 2009-2013)	0.198	0.744	0.840

Notes: P-values for tests of equality of female_phd coefficient between consecutive time periods in all fields. H0: $\beta_{\text{period i}} = \beta_{\text{period i+1}}$. Lower values indicate stronger evidence of differences. *** p<0.01, ** p<0.05, * p<0.10.

Table A6: Descriptive statistics - Supervisors

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Humanities and Law				**Biological and Earth Sc.**				**STEM**				**Social Sc.**			
	All	Male	Female	p-value	All	Male	Female	p-value	All	Male	Female	p-value	All	Male	Female	p-value
Female	0.307				0.312				0.191				0.318			
All Publication	10.678	11.882	8.054	0.000	99.370	111.949	70.478	0.000	94.864	98.131	72.315	0.000	19.765	21.422	16.219	0.000
Total Impact	4.532	5.605	2.215	0.022	179.184	201.281	129.724	0.000	85.427	87.825	67.512	0.000	12.918	14.970	8.682	0.000
Average Impact per publication	0.170	0.181	0.141	0.000	1.616	1.606	1.655	0.028	0.724	0.723	0.737	0.296	0.436	0.464	0.375	0.000
Number of supervision	5.740	6.363	4.356	0.000	3.409	3.737	2.740	0.000	4.810	5.064	3.931	0.000	5.170	5.836	3.826	0.000
Observations	12,801				20,583				39,584				12,939			

Table A7: Descriptive statistics - PhD Graduates

	All Fields	Humanities & Law	Biological & Earth Sciences	STEM	Social Sciences	Economics
Before 4 years						
Female	0.40	0.49	0.50	0.28	0.48	0.37
Any Publi	0.42	0.14	0.46	0.60	0.23	0.26
Observations	286,814	58,008	49,859	129,402	49,545	9,303
if Any Publi = 1						
Undivided publis	7.94	2.38	8.31	9.12	3.13	3.26
Publications	1.84	1.78	1.43	1.98	1.75	1.77
Impact	0.74	0.17	1.63	0.59	0.35	0.68
Coauthors	13.80	1.42	12.32	17.23	2.28	2.12
Observations	120,248	8,054	22,838	77,807	11,549	2,399
Between 5 and 10 years						
Female	0.38	0.47	0.48	0.27	0.46	0.36
Any Publi	0.33	0.15	0.37	0.44	0.23	0.25
Observations	221,128	45,430	39,323	98,854	37,521	7,274
if Any Publi = 1						
Undivided publis	10.00	2.87	9.34	12.43	4.15	4.77
Publications	2.26	2.05	1.50	2.58	2.09	2.22
Impact	0.79	0.17	1.81	0.62	0.39	0.72
Coauthors	9.40	1.16	9.21	14.23	2.37	1.81
Observations	73,738	6,687	14,671	43,922	8,458	1,806

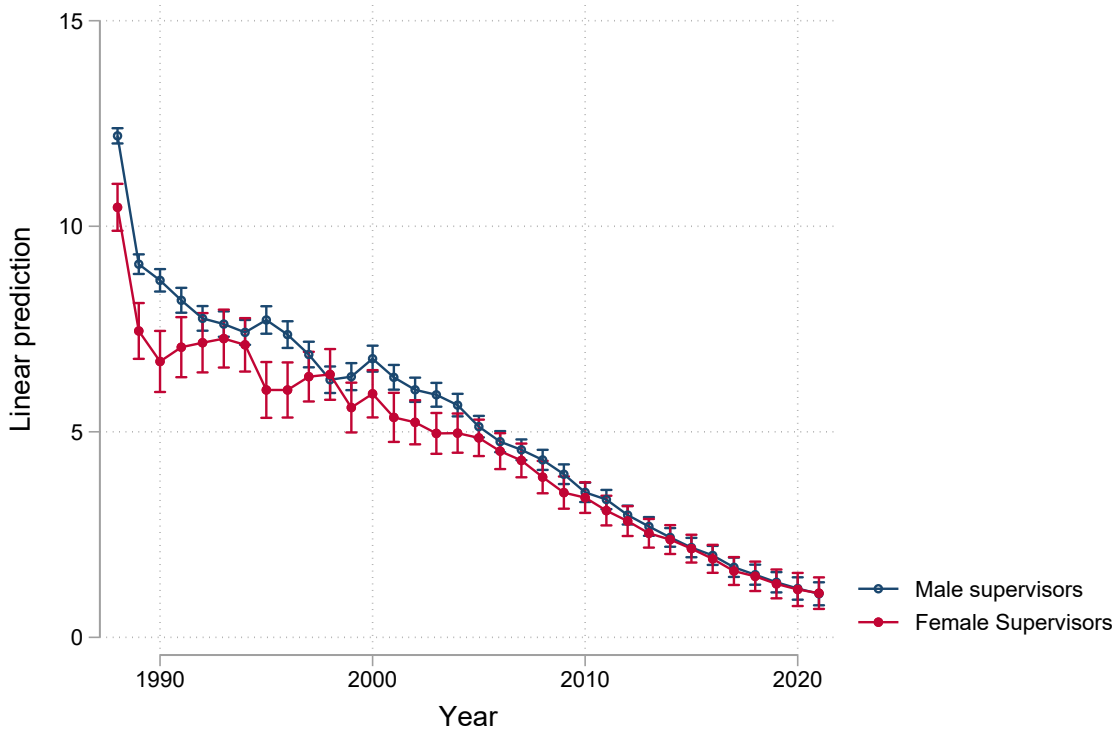
Table A8: Economics - Descriptive statistics of PhD Graduates

	1	2	3	4
	All	Male	Female	p-value
Female	0.37			
Co-supervision	0.129	0.113	0.146	0.027
Any publication, before 4 years:	0.256	0.268	0.256	0.208
All publication, before 4 years:	3.243	3.529	2.767	0.000
Total Impact, before 4 years:	2.339	2.698	1.846	0.000
Average Impact per publication, before 4 years:	0.708	0.737	0.683	0.367
Any publication, between 5 and 10 years:	0.246	0.264	0.237	0.011
All publication, between 5 and 10 years:	4.761	5.193	3.949	0.000
Total Impact, between 5 and 10 years:	4.369	5.002	3.156	0.000
Average Impact per publication, between 5 and 10 years:	0.782	0.801	0.746	0.400
Publication during PhD	0.357	0.351	0.371	0.205

Table A9: Economics - Descriptive statistics of Supervisors

	1	2	3	4
	All	Male	Female	p-value
Female	0.228			
All Publication	24.597	26.373	19.464	0.000
Total Impact	25.785	29.353	16.094	0.000
Average Impact per publication	0.819	0.856	0.732	0.065
Number of supervision	5.227	5.722	3.543	0.000
Observations	2,374			

Figure A1: Predicted probability of the number of Supervision for average female and male supervisor across cohorts



(a) 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.

Table A10: By Fields - Before 4 years after the Defense

	Humanities and Law			Biological and Earth Sc.			STEM			Social Sc.		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact
Female PhD	-0.0285***	-0.590***	-0.0224	0.0533***	-2.795***	-0.100***	-0.0546***	-2.361***	0.0207**	-0.0163***	-0.567***	-0.0674***
	(0.00339)	(0.0946)	(0.0164)	(0.00550)	(0.192)	(0.0317)	(0.00373)	(0.227)	(0.00969)	(0.00457)	(0.0965)	(0.0199)
Total Impact Supervisors	0.00368*	0.174***	0.0245***	0.0379***	1.184***	0.247***	0.00803***	6.642***	0.399***	0.0233***	0.253***	0.125***
	(0.00205)	(0.0476)	(0.00827)	(0.00329)	(0.103)	(0.0170)	(0.00193)	(0.107)	(0.00456)	(0.00276)	(0.0377)	(0.00776)
Having 2 supervisors	0.0975***	0.488***	0.0142	-0.162***	-0.827**	-0.384***	0.0104**	-0.540**	-0.0613***	0.0556***	0.659***	-0.0213
	(0.00840)	(0.158)	(0.0274)	(0.00998)	(0.406)	(0.0670)	(0.00451)	(0.249)	(0.0106)	(0.00956)	(0.163)	(0.0336)
Having 1 female supervisor	-0.00483	-0.155	-0.0193	0.108***	-0.725**	-0.0293	0.00744	0.297	0.0850***	0.0254***	-0.0518	-0.0465
	(0.00570)	(0.146)	(0.0254)	(0.00910)	(0.293)	(0.0483)	(0.00596)	(0.339)	(0.0144)	(0.00806)	(0.155)	(0.0319)
Having 2 female supervisors	-0.0364	-0.326	-0.000306	0.165***	-1.920**	0.00383	-0.0210	-0.167	0.136***	0.0390	-0.0672	-0.0259
	(0.0262)	(0.488)	(0.0848)	(0.0270)	(0.963)	(0.159)	(0.0180)	(0.985)	(0.0420)	(0.0299)	(0.472)	(0.0972)
Having 1 female	0.00344	-0.105	-0.00744	0.121***	0.633	0.0684	-0.00151	-0.627	0.0444***	0.0141	-0.359	-0.0133
and 1 male supervisors	(0.0137)	(0.244)	(0.0424)	(0.0146)	(0.562)	(0.0927)	(0.00729)	(0.396)	(0.0169)	(0.0154)	(0.253)	(0.0520)
Female X Total Impact Supervisors	0.000319	-0.0797	0.0157	0.000438	-0.441***	-0.0350	0.00867***	-1.150***	-0.0478***	-0.00143	-0.0805	0.0544***
	(0.00279)	(0.0675)	(0.0117)	(0.00469)	(0.145)	(0.0240)	(0.00337)	(0.185)	(0.00790)	(0.00408)	(0.0615)	(0.0126)
Female X Having 2 supervisors	-0.0264**	-0.291	0.0454	0.0166	1.023*	0.0725	0.0255***	-0.260	0.0118	0.00834	-0.141	0.0366
	(0.0116)	(0.229)	(0.0397)	(0.0138)	(0.546)	(0.0901)	(0.00787)	(0.445)	(0.0189)	(0.0133)	(0.229)	(0.0472)
Female X Having 1 female supervisor	-0.00276	0.189	0.0622*	-0.0491***	0.657*	0.0456	0.0150	0.182	-0.0460*	-0.00127	-0.0278	0.0634
	(0.00761)	(0.202)	(0.0350)	(0.0121)	(0.387)	(0.0639)	(0.0100)	(0.588)	(0.0251)	(0.0106)	(0.206)	(0.0423)
Female X Having 2 female supervisors	0.0396	0.666	-0.0601	-0.0619*	1.290	-0.165	0.0257	0.162	-0.0692	0.0386	-0.0954	-0.00779
	(0.0320)	(0.605)	(0.105)	(0.0350)	(1.237)	(0.204)	(0.0269)	(1.482)	(0.0631)	(0.0364)	(0.568)	(0.117)
Female X Having 1 female	-0.0235	0.331	-0.0370	-0.0552***	-0.573	-0.108	-0.0209*	0.137	-0.0202	0.0120	-0.0342	0.000142
and 1 male supervisors	(0.0183)	(0.344)	(0.0597)	(0.0201)	(0.758)	(0.125)	(0.0126)	(0.697)	(0.0297)	(0.0204)	(0.333)	(0.0685)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
University FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	56545	8002	8002	48664	22834	22834	126597	77842	77842	48634	11556	11556

Standard errors in parentheses. All regressions include also the variables *female, Total Impact of the supervisors*, and the interaction between the two.
Number of publications and *average impact* are computed conditionally of *any publication equal to 1*
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A11: By Fields - Between 5 and 10 years after the Defense

	Humanities and Law			Biological and Earth Sc.			STEM			Social Sc.		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact
Female PhD	-0.0290***	-0.663***	0.00343	0.00576	-3.939***	-0.108**	-0.101***	-3.829***	0.0229	-0.0235***	-1.000***	-0.0557**
	(0.00385)	(0.124)	(0.0203)	(0.00584)	(0.244)	(0.0465)	(0.00431)	(0.354)	(0.0146)	(0.00506)	(0.140)	(0.0224)
Total Impact Supervisors	-0.000706	0.593***	0.0373*	0.0537***	1.099***	0.288***	0.0269***	6.416***	0.397***	0.0232***	0.419***	0.202***
	(0.00247)	(0.120)	(0.0196)	(0.00455)	(0.161)	(0.0307)	(0.00328)	(0.227)	(0.00938)	(0.00368)	(0.0713)	(0.0114)
Having 2 supervisors	0.0962***	0.687**	0.0781*	-0.101***	0.0251	-0.410***	0.0226***	0.473	-0.0379**	0.0762***	0.439	-0.00125
	(0.0117)	(0.270)	(0.0443)	(0.0125)	(0.577)	(0.110)	(0.00598)	(0.432)	(0.0178)	(0.0126)	(0.286)	(0.0458)
Having 1 female supervisor	0.00785	-0.184	-0.0144	0.104***	-0.429	0.116	0.00278	0.381	0.0678***	0.0231**	0.000178	-0.0665*
	(0.00668)	(0.195)	(0.0320)	(0.01000)	(0.377)	(0.0718)	(0.00703)	(0.523)	(0.0216)	(0.00939)	(0.241)	(0.0386)
Having 2 female supervisors	-0.0231	-0.363	-0.187	0.126***	0.217	0.109	-0.00107	-1.538	0.292***	0.0257	3.391***	-0.0557
	(0.0443)	(1.009)	(0.165)	(0.0420)	(1.696)	(0.323)	(0.0294)	(2.102)	(0.0867)	(0.0496)	(1.067)	(0.171)
Having 1 female	-0.0190	-0.527	-0.147*	0.0850***	0.849	0.217	-0.00511	-1.004	0.0702**	0.0388*	0.358	-0.0610
and 1 male supervisors	(0.0203)	(0.465)	(0.0763)	(0.0201)	(0.865)	(0.165)	(0.0106)	(0.758)	(0.0312)	(0.0226)	(0.487)	(0.0778)
Female X Total Impact Supervisors	-0.00173	-0.188	0.0370	-0.0158**	-0.304	0.0503	0.00347	-1.267***	-0.0660***	-0.00575	-0.192*	-0.0346*
	(0.00340)	(0.233)	(0.0381)	(0.00653)	(0.241)	(0.0459)	(0.00576)	(0.411)	(0.0170)	(0.00534)	(0.112)	(0.0179)
Female X Having 2 supervisors	-0.0397**	-0.802**	-0.0779	-0.00730	0.240	-0.0207	0.0108	-0.708	-0.00462	-0.0315*	0.172	0.0847
	(0.0163)	(0.400)	(0.0655)	(0.0178)	(0.830)	(0.158)	(0.0107)	(0.829)	(0.0342)	(0.0178)	(0.420)	(0.0672)
Female X Having 1 female supervisor	-0.0130	0.0448	-0.0151	-0.0599***	0.353	-0.00705	0.00682	-0.960	0.00947	-0.00354	-0.0421	0.0361
	(0.00900)	(0.275)	(0.0450)	(0.0134)	(0.515)	(0.0981)	(0.0119)	(0.961)	(0.0396)	(0.0124)	(0.324)	(0.0517)
Female X Having 2 female supervisors	0.0380	0.0469	0.0909	-0.0106	-1.927	-0.0839	0.0142	-0.420	-0.0500	0.0246	-3.363**	-0.0775
	(0.0540)	(1.260)	(0.207)	(0.0536)	(2.171)	(0.414)	(0.0429)	(3.218)	(0.133)	(0.0600)	(1.309)	(0.209)
Female X Having 1 female	0.00906	0.361	0.117	-0.0276	-1.582	-0.244	-0.00151	2.868**	0.0621	0.00483	-0.859	-0.0210
and 1 male supervisors	(0.0274)	(0.660)	(0.108)	(0.0280)	(1.222)	(0.233)	(0.0185)	(1.412)	(0.0582)	(0.0299)	(0.662)	(0.106)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
University FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	44189	6616	6616	38247	14642	14642	96433	43895	43895	36759	8442	8442

Standard errors in parentheses. All regressions include also the variables *female, Total Impact of the supervisors*, and the interaction between the two.
Number of publications and *average impact* are computed conditionally of *any publication equal to 1*
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A12: Economics - Effect of supervision's gender composition

	Before t+4			Between t+5 and t+10		
	(1)	(2)	(3)	(4)	(5)	(6)
	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact
Female PhD	-0.0377***	-0.794***	-0.0218	-0.0469***	-1.359***	-0.0449
	(0.00919)	(0.150)	(0.0592)	(0.0105)	(0.298)	(0.0624)
Total Impact Supervisors	0.0235***	0.0378	0.118***	0.0331***	0.133*	0.131***
	(0.00321)	(0.0345)	(0.0136)	(0.00433)	(0.0770)	(0.0161)
Having 2 supervisors	0.0485***	0.500**	-0.0309	0.101***	0.299	0.0216
	(0.0163)	(0.229)	(0.0899)	(0.0216)	(0.510)	(0.107)
Having 1 female supervisor	0.0460***	-0.0189	0.0508	0.0122	-0.388	0.0207
	(0.0159)	(0.239)	(0.0939)	(0.0191)	(0.525)	(0.110)
Having 2 female supervisors	0.0240	-0.738	-0.102	0.0490	-1.693	-0.338
	(0.0556)	(0.764)	(0.300)	(0.105)	(2.372)	(0.496)
Having 1 female supervisor and 1 male supervisor	0.0123	-0.0414	-0.101	0.00469	0.118	-0.167
	(0.0263)	(0.360)	(0.142)	(0.0400)	(0.927)	(0.194)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
University FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9141	2383	2383	7149	1793	1793

Standard errors in parentheses. All regressions include the variables *female, Total Impact of the supervisors*, and the interaction between the two.
Number of publications and *average impact* are computed conditionally of *any publication equal to 1*
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A13: Economics - Effect of supervision's gender composition

	Before t+4			Between t+5 and t+10		
	(1)	(2)	(3)	(4)	(5)	(6)
	Any Publi	# Publi	Av. Impact	Any Publi	# Publi	Av. Impact
Female PhD	-0.0394***	-0.776***	-0.0975	-0.0331***	-1.160***	-0.0705
	(0.0106)	(0.188)	(0.0739)	(0.0116)	(0.343)	(0.0717)
Total Impact Supervisors	0.0227***	0.0548	0.0976***	0.0278***	0.189**	0.147***
	(0.00396)	(0.0407)	(0.0160)	(0.00531)	(0.0932)	(0.0195)
Having 2 supervisors	0.0565***	0.444	-0.0664	0.133***	0.373	-0.137
	(0.0210)	(0.287)	(0.113)	(0.0281)	(0.637)	(0.133)
Having 1 female supervisor	0.0475**	-0.219	-0.0590	0.0493**	-0.236	0.0151
	(0.0211)	(0.311)	(0.122)	(0.0252)	(0.644)	(0.135)
Having 2 female supervisors	-0.0207	-0.209	-0.00562	0.0192	0.450	0.0944
	(0.0814)	(1.076)	(0.425)	(0.164)	(3.416)	(0.736)
Having 1 female	-0.0315	0.648	-0.0687	0.0236	0.843	-0.0141
and 1 male supervisors	(0.0366)	(0.503)	(0.198)	(0.0565)	(1.202)	(0.251)
Female X Total Impact Supervisors	0.00206	-0.0592	0.0630**	0.0148*	-0.158	-0.0443
	(0.00615)	(0.0675)	(0.0266)	(0.00834)	(0.142)	(0.0296)
Female X Having 2 supervisors	-0.0257	-1.160	-0.0669	0.00772	0.688	0.0699
	(0.0795)	(1.111)	(0.437)	(0.162)	(3.462)	(0.724)
Female X Having 1 female supervisor	-0.00338	0.470	0.262	-0.0870**	-0.431	0.0248
	(0.0312)	(0.471)	(0.185)	(0.0378)	(1.093)	(0.229)
Female X Having 2 female supervisors	0.100	0.695	-0.0274	0.0903	-4.440	-0.895
	(0.111)	(1.522)	(0.599)	(0.211)	(4.739)	(0.991)
Female X Having 1 female	0.0888*	-1.379*	-0.0430	-0.0244	-1.746	-0.418
and 1 male supervisors	(0.0523)	(0.725)	(0.285)	(0.0797)	(1.843)	(0.386)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
University FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9141	2383	2383	7149	1793	1793

Standard errors in parentheses. All regressions include also the variables *female, Total Impact of the supervisors*, and the interaction between the two.
Number of publications and *average impact* are computed conditionally of *any publication equal to 1*
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B Table 2 by Fields

Table B14: Gender Gap Intensive and Extensive Margin - Humanities and Law

All Disciplines	*Before t+4*				*Between t+5 and t+10*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Any Publication				
Female PhD	-0.0136***	-0.0267***	-0.0190***	-0.0316***	-0.0182***	-0.0265***	-0.0239***	-0.0333***
	(0.00289)	(0.00285)	(0.00286)	(0.00282)	(0.00335)	(0.00334)	(0.00332)	(0.00331)
Observations	58008	58008	58003	58003	45430	45430	45426	45426
				# Publication if Any Publication = 1				
Female PhD	-0.559***	-0.601***	-0.539***	-0.580***	-0.724***	-0.776***	-0.689***	-0.749***
	(0.0723)	(0.0725)	(0.0728)	(0.0730)	(0.101)	(0.101)	(0.102)	(0.102)
				Av. Impact if Any Publication = 1				
Female PhD	-0.0133	-0.00592	-0.0127	-0.00497	-0.0109	-0.00496	-0.00723	-0.000844
	(0.0125)	(0.0125)	(0.0126)	(0.0126)	(0.0169)	(0.0169)	(0.0169)	(0.0169)
Observations	8161	8161	8154	8154	6791	6791	6782	6782
Controls								
University FE			Yes	Yes			Yes	Yes
Year FE		Yes		Yes		Yes		Yes

Table B15: Gender Gap Intensive and Extensive Margin - Biological and Earth Sciences

All Disciplines	*Before t+4*				*Between t+5 and t+10*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Any Publication				
Female PhD	0.0680***	0.0573***	0.0573***	0.0494***	0.0161***	0.00894*	0.00769	0.000930
	(0.00446)	(0.00434)	(0.00434)	(0.00433)	(0.00490)	(0.00490)	(0.00481)	(0.00481)
Observations	49859	49852	49852	49852	39323	39323	39316	39316
				# Publication if Any Publication = 1				
Female PhD	-2.517***	-2.509***	-2.575***	-2.552***	-3.832***	-3.876***	-3.909***	-3.949***
	(0.149)	(0.149)	(0.149)	(0.149)	(0.198)	(0.198)	(0.198)	(0.198)
				Av. Impact if Any Publication = 1				
Female PhD	-0.0474*	-0.0900***	-0.0597**	-0.0967***	-0.0704*	-0.114***	-0.0829**	-0.130***
	(0.0258)	(0.0257)	(0.0252)	(0.0250)	(0.0389)	(0.0387)	(0.0384)	(0.0382)
Observations	23289	23289	23283	23283	14991	14991	14984	14984
Controls								
University FE			Yes	Yes			Yes	Yes
Year FE		Yes		Yes		Yes		Yes

Table B16: Gender Gap Intensive and Extensive Margin - STEM

All Disciplines	*Before t+4*				*Between t+5 and t+10*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Any Publication				
Female PhD	-0.0425***	-0.0520***	-0.0395***	-0.0470***	-0.0972***	-0.100***	-0.0942***	-0.0973***
	(0.00301)	(0.00293)	(0.00299)	(0.00292)	(0.00353)	(0.00351)	(0.00353)	(0.00351)
Observations	129402	129402	129400	129400	98854	98854	98853	98853
				# Publication if Any Publication = 1				
Female PhD	-1.954***	-2.138***	-2.016***	-2.160***	-3.396***	-3.523***	-3.457***	-3.583***
	(0.179)	(0.179)	(0.179)	(0.178)	(0.294)	(0.293)	(0.293)	(0.292)
				Av. Impact if Any Publication = 1				
Female PhD	0.0616***	0.0525***	0.0451***	0.0382***	0.0806***	0.0711***	0.0631***	0.0541***
	(0.00808)	(0.00806)	(0.00786)	(0.00784)	(0.0125)	(0.0124)	(0.0122)	(0.0122)
Observations	78991	78991	78988	78988	44636	44636	44632	44632
Controls								
University FE			Yes	Yes			Yes	Yes
Year FE		Yes		Yes		Yes		Yes

Table B17: Gender Gap Intensive and Extensive Margin - Social Sciences

All Disciplines	*Before t+4*				*Between t+5 and t+10*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Any Publication				
Female PhD	0.0134***	-0.00805**	0.00886**	-0.0112***	-0.00653	-0.0211***	-0.00946**	-0.0250***
	(0.00382)	(0.00377)	(0.00380)	(0.00375)	(0.00435)	(0.00434)	(0.00434)	(0.00433)
Observations	49545	49545	49543	49543	37521	37521	37520	37520
				# Publication if Any Publication = 1				
Female PhD	-0.505***	-0.585***	-0.544***	-0.616***	-0.886***	-1.038***	-0.891***	-1.045***
	(0.0732)	(0.0732)	(0.0739)	(0.0739)	(0.115)	(0.115)	(0.116)	(0.116)
				Av. Impact if Any Publication = 1				
Female PhD	-0.0673***	-0.0613***	-0.0567***	-0.0525***	-0.0654***	-0.0605***	-0.0580***	-0.0535***
	(0.0159)	(0.0159)	(0.0157)	(0.0158)	(0.0193)	(0.0194)	(0.0190)	(0.0191)
Observations	11745	11745	11742	11742	8592	8592	8587	8587
Controls								
University FE			Yes	Yes			Yes	Yes
Year FE		Yes		Yes		Yes		Yes

Table B18: Gender Gap Intensive and Extensive Margin - Economics

All Disciplines	*Before t+4*				*Between t+5 and t+10*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Any Publication				
Female PhD	-0.0119	-0.0329***	-0.0172*	-0.0354***	-0.0270**	-0.0422***	-0.0324***	-0.0477***
	(0.00945)	(0.00924)	(0.00933)	(0.00914)	(0.0106)	(0.0106)	(0.0106)	(0.0105)
Observations	9303	9303	9293	9293	7274	7274	7266	7266
				# Publication if Any Publication = 1				
Female PhD	-0.763***	-0.794***	-0.744***	-0.764***	-1.245***	-1.371***	-1.184***	-1.326***
	(0.146)	(0.146)	(0.147)	(0.148)	(0.285)	(0.286)	(0.293)	(0.293)
				Av. Impact if Any Publication = 1				
Female PhD	-0.0547	-0.0366	-0.0358	-0.0221	-0.0550	-0.0410	-0.0447	-0.0389
	(0.0606)	(0.0609)	(0.0601)	(0.0605)	(0.0654)	(0.0655)	(0.0636)	(0.0637)
Observations	2453	2453	2442	2442	1852	1852	1842	1842
Controls								
University FE			Yes	Yes			Yes	Yes
Year FE		Yes		Yes		Yes		Yes

C Data Thèses.fr - Detailed Procedure

We construct our dataset using data from *Theses.fr*, which provides records of all PhD theses defended in French universities between 1988 and 2021. *Theses.fr* is a centralized public platform that systematically compiles data from university catalogs across France, sourced through library and documentation services within higher education and research institutions, establishing it as the most comprehensive and reliable platform for French PhD graduation.

The dataset is not immune to limitations. Data entry occurs manually at various stages, which introduces the potential for spelling inconsistencies. Furthermore, certain theses may go unreported due to a lack of submission by graduates, loss, or failure to meet quality control standards, which we estimate affects approximately 5% of theses each year. In addition, the processing of records is time-intensive, making the data for 2022 potentially incomplete. Additionally, an observed scarcity of records prior to 1988 suggests further underreporting. Consequently, we restrict our sample to the period from 1988 to 2021.

From an initial sample of 407,260 theses recorded between 1988 and 2021, we impose a series of exclusions to ensure data reliability. Theses supervised by more than two advisors—constituting roughly 2% of the dataset—are excluded, yielding a refined dataset of 399,118 observations. Additional filters are applied to exclude records with incomplete names for PhD candidates or supervisors, as well as cases with missing discipline information, resulting in a final dataset of 397,536 theses. At this stage, we exclude theses in medicine due to reliability concerns, which we discuss in detail in Section C.2, leaving a total of 340,073 observations.

For each thesis, we gathered information on the research discipline, defense year, university affiliation, and full names of the PhD student and supervisor(s). In the sections that follow, we detail the data-cleaning procedures applied to discipline and university affiliation, explain the exclusion of health and medical sciences, and outline our methodology for associating gender with first names.

C.1 Gender association

In this study, we determine the gender of both PhD students and supervisors based on first names. Our primary source is the INSEE database, which compiles first names assigned in France from 1900 to 2020, including the gender distribution for each name over the period 1940–2020. We focus on this range, assuming that the majority of PhD students in our dataset were born after 1940. For names associated with both genders, we establish a reliable gender ratio and retain only those names where one gender represents at least 95% of total occurrences; names below this threshold are treated as indeterminate. This process allows us to identify the gender for 305,187 out of 340,073 PhD student first names. Recognizing the limitations posed by foreign names, we supplement INSEE data with governmental databases from Australia, Canada, Spain, Sweden, the UK, and the US.

Through additional data collection from these international sources, we resolve the gender of an additional 9,246 PhD students. We further employ the methodology of Benveniste (2023), which classifies names based on the last two letters and the associated gender probability, allowing us to identify the gender of 3,004 more PhD students. In total, we successfully identify the gender of 317,437 doctoral students, covering 93% of the sample. Of the remaining 7%, 3% (8,166 names) represent names used by both genders without a clear distributional majority (e.g., Camille, Claude). Using the same approach, we successfully associate a gender for 95% of PhD supervisors.

C.2 Disciplines

The categorization of discipline fields in *Thèses.fr* is imprecise, partly due to manual data entry. The database originally contained around 22,000 unique entries for the discipline variable, which we grouped into twenty-two subcategories and further into four broader categories based on the Australian and New Zealand Standard Research Classification (ANZSRC). To classify these entries, we adopted a keyword-based approach, manually associating each entry with relevant

discipline categories. We began by filtering with specific keywords unique to each category, as illustrated in the following examples:

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

Following this, we applied progressively broader keywords, carefully verifying that each association was accurate to avoid misclassification. For example, general keywords like "MAGNETISME," "LANGUES," and "VEGETAL" were used, corresponding to "Physical Sciences," "Language and Culture," and "Biological Sciences," respectively.

In cases of ambiguous or unknown disciplines, we examined thesis titles and applied the same keyword methodology. Despite these efforts, discipline association may still contain errors, especially for multidisciplinary theses that we must assign to a single category. To account for this, we created four overarching categories to group similar subjects: Humanities and Law, Biological and Earth Sciences, Sciences, Technology and Engineering, and Social Sciences.

Drop *Health and Medical Sciences* discipline. In this section, we discuss the unreliability of Health and Medical Sciences thesis data prior to the 2000s. Our analysis identified notable irregularities in medical theses data, particularly around 1994. We traced the origin of these discrepancies to the data selection mechanism in *Thèses.fr*, which automatically selects defended doctoral theses and excludes documents not categorized as such. However, in the French health sciences domain, "*thèses d'exercice*" - theses defended to obtain a State Diploma of Doctor required for medical practice—are often included. These are distinct from doctoral theses intended to confer the national diploma of doctor (*diplôme national de doctorat*). Unfortunately, during data import into *Thèses.fr*, a substantial number of *thèses d'exercice* were incorrectly labeled as doctoral theses, introducing bias.

Figure C2 displays the number of theses defended in health and medical sciences since 1988, showing that institutions began systematically distinguishing between doctoral theses and *thèses d'exercice* around the early 2000s. As we aim to focus on theses from before 2000, we must exclude medical theses from our sample to avoid biasing our study.

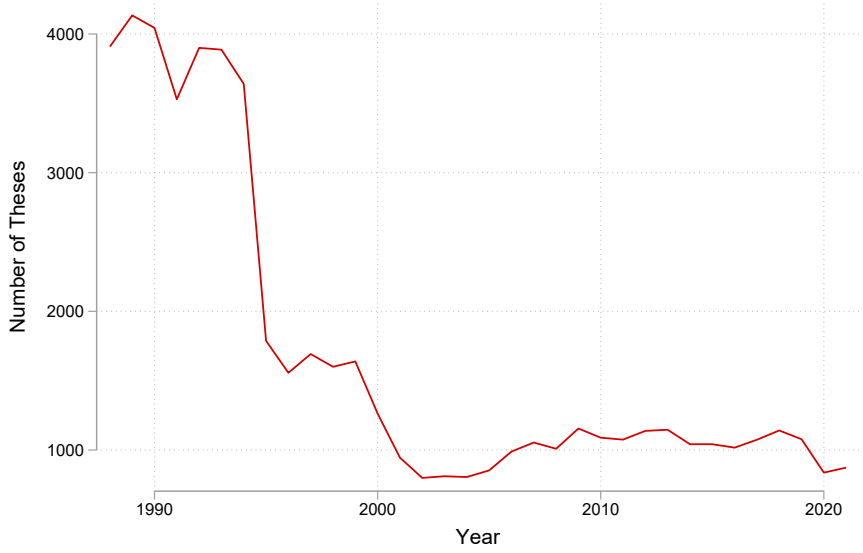


Figure C2: Number of Thesis by Year of Defense in Health and Medical Sciences

C.3 University

In recent years, French universities have been undergoing a series of institutional mergers, intended to enhance their international visibility and competitiveness[10]. To ensure consistency in our analysis, we standardized university codes following the documentation provided by Thèses.fr[11] and tracked changes in institutional names over time. Between 2007 and 2020, 26 new universities were established through the consolidation of 76 existing institutions. For example, in 2013, Aix-Marseille University was formed by merging Aix-Marseille 1, Aix-Marseille 2, and Aix-Marseille 3.

In certain cases, however, institutions have subsequently split, complicating the distinction between former codes. In such instances, it is more practical to apply a single code for universities that have separated, even at the cost of some specificity. For example, the University of Paris-Saclay was initially formed in 2015 as a merger of 11 institutions, only to divide into two distinct entities by the end of 2019.

Table C19, C20, and C21 provide a detailed list of all universities and their coding changes, while Table C22 covers the National Institutes of Polytechnics, and Table C23 presents the Higher Education Establishments. Each institution is listed with its associated code and any historical coding changes from 1988 to 2021. Any changes or codes appearing before or after

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

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

this period are not documented. A blank description indicates no changes during the specified timeframe.

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 C19: 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 2018
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 C20: 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 C21: 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 C22: 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 C23: 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