The impact of research independence on PhD students' careers: Large-scale evidence from France

Sofia Patsali[1,3] Michele Pezzoni[1,4,5,6] Fabiana Visentin[2]

[1]Université Côte d'Azur, CNRS, GREDEG, France
[2]Maastricht University, UNU-MERIT, The Netherlands
[3]Université de Strasbourg, CNRS, BETA, France
[4]Observatoire des Sciences et Techniques, HCERES, France
[5]OFCE, Sciences Po, France
[6]ICRIOS, Bocconi University, Italy

Abstract: This study investigates the effect of research independence during the PhD period on students' career outcomes. We use a unique and detailed dataset on the French population of STEM PhD students who graduated between 1995 and 2013. To measure research independence, we compare the PhD thesis content with the supervisor's research. We employ advanced neural network text analysis techniques evaluating the similarity between student's thesis abstract and supervisor's publications during the PhD period. After exploring which characteristics of the PhD training experience and supervisor explain the level of research similarity, we estimate how similarity associates with the likelihood of pursuing a research career. We find that the student thesis's similarity with her supervisor's research work is negatively associated with starting a career in academia and patenting probability. Increasing the PhD-supervisor similarity score by one standard deviation is associated with a 2.1 percentage point decrease in the probability of obtaining an academic position and a 0.57 percentage point decrease in the probability of patenting. However, conditional on starting an academic career, PhD-supervisor similarity is associated with a higher student's productivity after graduation as measured by citations received, network size, and probability of moving to a foreign or US-based affiliation.

Keywords: Research independence, Early career researchers, Scientific career outcomes, Neural network text analysis

JEL Codes: D22, O30, O33, O38

Acknowledgments: This work has been supported by the French government, through the UCA^{JEDI} Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. M.P. and F.V. also acknowledge the financial support of GIGA (ANR-19-CE26-0014).

1. Introduction

“Fellows are selected on the basis of their independent research accomplishments, creativity, and potential to become leaders in the scientific community through their contribution to their field” (SLOAN Foundation Research Fellowships)

“*Researchers should demonstrate how their existing professional experience, talents and the proposed research will contribute to their development as independent/mature researchers, during the fellowship”* (Marie Skłodowska-Curie Actions Individual Fellowship, EU Framework Program for Research and Innovation)

A common thread through the quotes reported above describing the selection criteria of two leading grant programs is the funding applicant's requirement to demonstrate a certain level of research independence. Like the Alfred P. Sloan Foundation, funding agencies include independent research achievements of young investigators as one of the most prominent characteristics revealing young candidates’ aptitude to make a significant scientific contribution. Moreover, independence is a selection criterion and one of the target features and outcomes aimed by funding opportunities like the Marie Skłodowska-Curie Actions Individual Fellowship supported by the EU Framework Program for Research and Innovation. In this respect, the EU insists that candidates should clearly state how their training and experience relate to their leadership and independent thinking skills.

Although being a topic of concern for academics and policymakers, the research independence of PhD students and early career researchers presents many unexplored issues and questions. This paper aims to tackle some of them. We explore how student training experience and supervisor’s characteristics during the PhD period explain research independence. Then, we investigate the impact of research independence on the likelihood of pursuing an academic career or being involved in R&D activities after graduating. Finally, conditional on starting an academic career, we estimate how independence affects career outcomes. Our analyses explore a unique large-scale dataset on the entire population of 46,774 STEM students in France who graduated between 1995 and 2013. For those students, we compare students’ thesis and supervisors’ work. A high level of similarity identifies students dependent from their supervisors, while low level of similarity identifies independent students.

Looking at the factors that lead a PhD student to develop an independent research profile, we find that students with productive supervisors show lower levels of independence. Moreover, female students pairing with female supervisors are the least inclined to conduct independent research. A female student pairing with a female supervisor shows a thesis 0.12 standard deviations more similar to her supervisor's work than a male student pairing with a male supervisor. Among the STEM disciplines, students in math show the lowest level of independence. When considering the association between being independent during the PhD and career outcomes, the student's research similarity with the supervisor negatively associates with starting a career in academia and patenting. Our results show that increasing the student's similarity by one standard deviation is associated with a 2.1 percentage point decrease in the probability of obtaining an academic position in the five years after graduation and a 0.57 percentage point decrease in the probability of patenting at least once within the same time frame. However, conditional on starting an academic career, research similarity is associated with a higher student's productivity measured by the average number of yearly citations received, network size, and probability of moving to a foreign or US-based affiliation.

Our work relates to a growing body of contributions in economics and sociology of science on early career researchers and scientific productivity. Several studies have analysed the productivity determinants for experienced scientists investigating factors such as age (Zuckerman and Merton, 1972, Diamond, 1986, Stephan and Levin, 1997), cohort (Weiss, 1981), training (Garcia-Romero and Modrego, 2001) and gender (Levin and Stephan, 1998). More recently, a complementary body of studies has focused on the determinants of young scholars' productivity. These works have either tackled the PhD productivity through factors such as the supervisor's gender (Tenenbaum *et al.* 2001, Pezzoni *et al.* 2016, Gaule and Piacentini 2018), funding (Horta *et al.* 2018), faculty quality (Waldinger 2010), or a more general set of social-environmental characteristics (Corsini *et al.* 2021). Previous studies have remained silent on an essential factor in young scholars' careers, such as the nature of their relations of independence with their supervisor during the PhD period and how the degree of research independence affects researchers' later academic productivity.

A few studies have analysed qualitatively the process throw which scholars develop independence in their research and careers. For instance, Laudel and Glaser (2008) looked at the transitional phases that lead scholars from the apprentice to becoming responsible for setting their own research agendas and running their own research teams. In a further study, Laudel and Glaser (2018) described the process through which early career researchers develop their research programs. Other studies have explored the impact of PhD training on later career outcomes. For instance, Horta and Santos (2016), on a sample of Portuguese students, find that students who publish during their PhD are more likely to publish single-authored papers and publish with foreign peers later in their careers. Drawing on survey and bibliometric data, Yoshioka-Kobayashi and Shibayama (2020) explore the effect of different types of research training on later academic independence at Japanese universities in the life sciences field. They show that higher autonomy leads to greater organizational independence. Moreover, encouragement to deviate from conventional research topics during early-career training is associated with greater cognitive independence in students' later careers.

Our study advances the extant literature on the effect of research independence in two different ways. First, our study analyses the independence of a PhD student from her supervisor looking at the content of the research conducted. Productivity and career studies usually overlook the research content focusing on traditional bibliometrics indicators such as publication or citation count. By employing advanced neural network text analysis techniques, we conduct a fine-grained content analysis and evaluate the similarity between students' thesis and supervisors' publication abstracts. Importantly, using the thesis abstract, we include in our study also students who did not publish during the PhD training. The lack of publications for calculating productivity indexes often excludes these students from studies relying on bibliometric data. Second, while previous works focused on specific disciplines and relatively small samples of students affiliated to a limited number of prestigious universities, our work provides accurate analyses on what enhances the quality and effectiveness of doctoral training across disciplines and institutions. Doing so, we provide the empirical evidence needed to make discipline-specific informed policy decisions without discriminating between top-tier and low-rank universities.

The rest of the paper proceeds as follows. The next section discusses the relevance of the PhD academic training in shaping researchers' careers. Section 3 describes the data, and section 4 presents the empirical strategy. We discuss our results in Section 5, and, in the same section, we dig into the moderating effects of gender and discipline on independence. Section 6 concludes with some policy implications, limitations, and possible extensions of this work.

2. PhD training and research independence

Young researchers are confronted with the expectation of becoming intellectually independent members of their scientific communities since the very beginning of their careers (NRC 2005). According to an early definition advanced by Becher (1990:10), independence consists of "*learning to work on their own and discussing questions with senior colleagues rather than merely following their advice*". The PhD training is part and parcel of the journey towards academic independence as young scholars acquire knowledge and nurture their skills to become autonomous researchers (Campbell 2003, Laudel and Glaser 2008, Gardner 2008, Stephan 2012). PhD students undertake their first research projects under the supervision of established scientists (Delamont and Atkinson 2001, Latour and Woolgar 1979). The student-advisor relation is so influential in shaping students' profiles in the doctoral education process that supervisors can make or break a PhD student (Lee 2008, Mangematin 2000).

For the most part, young researchers begin their careers by contributing to their advisors' research agenda (Campbell 2003; Delamont and Atkinson 2001; Kam 1997), and their research independence is limited. Supervisors exert strong scientific authority in orienting students' areas of interest, while they also dispose of considerable administrative and political power to boost students' careers (Laudel 2008). Thus, PhD students' autonomy falls within a set of implicit contractual relations with their supervisors (Stephan and Levin 1997) that are governed by the delayed nature of reward after graduation and the effect of the supervisor's reputation: students' contribute to supervisor's work during her training period, and gain support later when entering the job market.

Driven by the delayed nature of rewards, PhD students might be incentivized to embrace the research lines suggested by their supervisors to leverage supervisors' resources and capabilities, avoiding the risks entailed in exploring new research lines during the thesis. At the same time, young researchers are subject to their respective communities' expectations to signal their ability to launch independent research programs as leaders rather than protégé of their supervisors (Merton 1973, Stephan 1996, Liénard *et al.* 2018). Thereby, PhD students are also encouraged to identify their original research agenda and explore new areas beyond supervisors' expertise (Kam 1997, Lee 2008, Lee *et al.* 2007, Mainhard *et al.* 2009, Shibayama 2019).

Due to the path dependency that affects research topic choices (Austin 2002, Campbell 2003, Laudel and Glaser 2008), the research topics selected in the thesis period can influence those investigated at later career stages. On that account, the student's approach to position their work during the PhD training can have long-lasting career effects.

In this paper, we focus our attention on the level of research independence reached by PhD students during their training period. We measure this level of independence by comparing the PhD thesis content with the papers published by the supervisor during the PhD training period. In the first set of analyses, we investigate the factors leading to a higher (lower) level of independence. Then, we look at how the research independence level associates with researchers' career outcomes.

3. Data and variables

3.1 Empirical context

The empirical context of our study is the population of French students who obtained their PhD in the STEM field between 1995 and 2013. French universities are well-known for their excellence in the STEM disciplines such as biology, chemistry, medicine, and mathematics and enumerate notable researchers such as Laplace (mathematics), Picardet (literature and chemistry), Calmette (medicine), Curie (physics), and Pasteur (microbiology). The country ranks 4th in the world in the number of Nobel prize awards (70) after the USA, UK, and Germany[1]. French universities are also well placed according to several international rankings such as Shanghai ranking, ARWU global ranking, and Nature's Lens score.

Universities' PhD programs are nowadays the pillars of the French educational system to train tomorrow's researchers. Doctorates in STEM disciplines represent about 70% of the doctorates yearly granted[2]. To access a PhD program, students must hold a national master's degree or an equivalent diploma, certifying their aptitude to undertake research. A committee of professors selects students through a formal interview process. Once they are hired, doctoral students obtain a fixed-term working contract for the entire duration of their studies. According to the

[1] The figure refers to 2021 data.
[2] https://www.campusfrance.org/en/what-involved-Doctorate-France

French *Ministry of Education, Higher Education & Research*, 9 out of 10 theses in STEM are completed in four years[3].

After completing the PhD, students have three possible career outcomes. First, they can undertake an academic career. The academic career often starts being employed as a postdoc in a French or foreign university. After the postdoc period, researchers apply for a tenure track position in France or abroad. The second possible career outcome for a PhD student is to be hired in the research division of a company and continue her research activity in the private sector. Finally, the third outcome is to quit the research career (either public or private) and find another non-research occupation.

3.2 Data sources

We combine different sources of data to create a unique dataset on French PhD students in STEM. We obtained fine-grained data on PhD theses from the nation-wide French repository of *Electronic Doctoral Theses* and combined those data with students' and supervisors' publication records collected from the Elsevier's SCOPUS database.

With special permission from the *Agence Bibliographique de l'Enseignement Supérieur*, the French public Institute in charge of maintaining the bibliographic archive of French universities, we collected the entire universe of STEM theses in France between 1995-2013. By accessing the *Electronic Doctoral Theses* repository, for each thesis record, we gathered information on the author, the university of graduation, defense date, supervisor's name, co-supervisor's name (if any), the field of study, and the abstract of the thesis in English and French.

Concerning publication data, we retrieved students' and supervisors' publications from Elsevier's SCOPUS database. For each publication, we collected basic metadata such as title, abstract, journal, year of publication, citations, and keywords. In addition, publication data also provides authors' affiliations.

Our initial sample counted 69,466 PhD students who graduated in STEM disciplines between 1995 and 2013, reporting a thesis abstract written in English. To implement the text analysis aiming to compare the content of the thesis abstracts with the supervisors' paper abstracts, we

[3] https://publication.enseignementsup-recherche.gouv.fr/eesr/FR/T744/le_doctorat_et_les_docteurs/

required supervisors to have at least one publication written in English during the PhD training period. We also dropped students and supervisors with common names, for whom it was unlikely to attribute publications correctly due to homonymy issues. The restrictions applied to the initial sample leave us with a study sample of 46,774 PhD students.

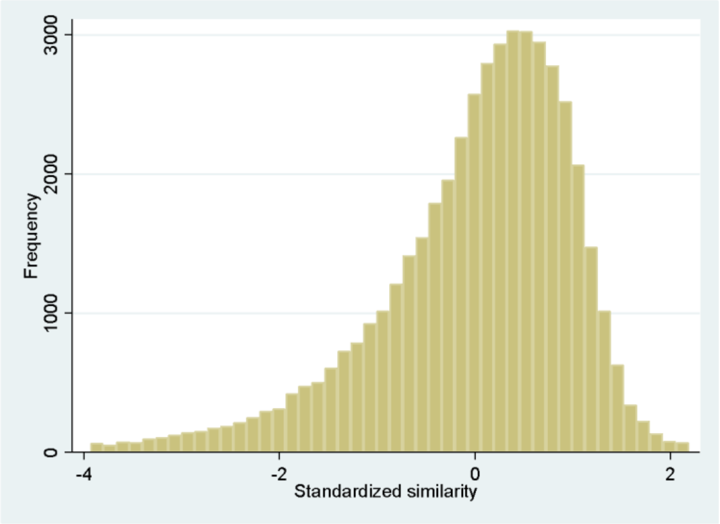
3.3 Variables

Research independence proxy

A crucial aspect of our empirical analysis is identifying a reliable proxy for the student's research independence from her supervisor. To do so, we compare the content of the student's PhD thesis with the content of the supervisor's research activity during the training of the PhD student. Specifically, we compute a similarity score between the thesis abstract and the abstracts of the papers published by the supervisor during the PhD period. A higher similarity score between the thesis and the supervisor's publications indicates lower students' independence from the supervisor. On the contrary, a lower level of similarity indicates higher independence. To operationalize our measure, we use an advanced neural network algorithm for text analysis that transforms thesis and publication documents into vectors according to the semantic meaning of the words appearing in their texts (Mikolov *et al.* 2013). Once we identified the vectorial representation of theses and publications, we calculated for each thesis-publication pair a cosine similarity value. Overall, we calculated cosine similarity values pairing the text of 46,774 theses abstracts with 411,128 supervisors' publication abstracts. A cosine similarity score can range from -1 to 1, where 1 means the thesis content is identical to the paired supervisor's publication. The similarity scores observed in our sample show only positive values. Indeed, for scientific documents such as thesis and publication abstracts, levels of cosine similarity below 0.75 indicate already substantial differences, e.g., values around 0.75 result from a comparison between a thesis in mathematics and a publication in biology. After calculating the cosine similarity scores pairing the thesis abstract with all the supervisor's publications, we calculate the maximum among those scores. We choose to calculate the maximum to identify if the student's thesis subject is close to at least one supervisor research line, in case supervisors' publications cover multiple research lines. Since most of the maximum cosine similarity values range between 0.75 and 1, we define our *Similarity* variable as the standardized[4] maximum cosine similarity score. This later variable ranges from -3.93 standard deviations to 2.18 standard deviations.

Figure 1 shows the distribution of our *Similarity* variable. As expected, the distribution is asymmetric left-skewed since many theses show content coherence with supervisors' research work, although a non-negligible part of our sample has a divergent content. In the figure, the long tail of the distribution illustrates this heterogeneity in *Similarity* values.

Figure 1: Similarity score distribution



The interpretation of the similarity score values might be challenging. In Appendix 1, we present an example to clarify the meaning of the values observed. As an example, we consider the case of a professor (B) with two students, A and Y. Student A graduated in 2007 from the University of Nantes and shows a very low level of *Similarity* with Professor's B work (*Similarity* = -3.4). During the PhD period, A did not publish any paper with his supervisor and had a co-supervisor (Professor C) whose research was highly similar to student's A thesis (*Similarity* = 0.42). This example shows that Professor B was not an expert in student A's thesis topic, and he invited a colleague to co-supervise her. Probably, student A was mentored more by her co-supervisor (Professor C) than her supervisor (Professor B). In contrast, the second student supervised by Professor B, Student Y, shows a high similarity with B's work (*Similarity*

[4] To calculate *Similarity* we subtracted to the similarity score its sample average and we divided by its standard deviation.

= 0.38). During the PhD period, student Y published two papers co-authored with her supervisor B. In this case, no co-supervisors were involved.

Our dependent variables: Similarity and students' career outcomes

In our study, we are interested in assessing the determinants of student's independence and the association between independence and student's career path. Therefore, our proxy for independence, i.e., the PhD-supervisor similarity score, plays the dual role of dependent and independent variable in our econometric models. Specifically, the PhD-supervisor similarity score is the dependent variable in the econometric model estimating the determinants of student's independence, while it is the independent variable in the econometric models explaining the student's career path. As career path, we look at obtaining an academic position (*Academic career*) or patenting (*At least one patent*), the latter one as a signal of being involved in an R&D activity. Conditional of staying in academia, we investigate closely the career outcomes by looking at the quantity and quality of the students' scientific productivity (*Number of publications* and *Average citations per article*), the dimension of their co-authorship network (*Number of distinct co-authors*), and if they start an international research career (*Having a foreign affiliation* and *Having a US affiliation*).

To identify those students who pursue an academic career, we collected the student's publications in the five years after the defense year t, from t+1 to t+6, and gathered the affiliations appearing in those publications. We construct a dummy variable, *Academic career*, that takes value 1 if we observe at least one academic affiliation, i.e., a university or a public research organization, denoting that the student has started an academic career, 0 otherwise. Regardless of staying in academia, the student might remain active in R&D activities. The dummy variable *At least one patent* indicates if the student has patented at least once in the 5 years after her defense. To identify patenting students, we consider patent applications at the European Patent Office.

For those who stay in academia, we characterize their research activity by counting the number of publications in the 5 years after the PhD defense (*Number of publications*). We calculate the yearly average number of citations received per paper (*Average citations per article*). We counted the number of distinct co-authors appearing in the publications (*Number of distinct co-authors*) as a proxy of the research network. Finally, we used the publication data to scrutinize the PhD students' affiliations and trace their mobility. The dummy *Having a foreign affiliation* equals 1 if the PhD student locates in at least one non-French affiliation in the 5 years after her graduation, 0 if she stays for the entire period in France. The dummy *Having a US affiliation* equals 1 if the PhD student locates in at least one US affiliation in the 5 years after her graduation, 0 otherwise.

Table 1 reports the descriptive statistics of our dependent variables. Regarding our overall sample of 46,774 students, 51% of them start an academic career and 8% result as the inventor of at least one patent in the 5 years following the PhD thesis defense. Students who stay in academia have, on average, 7.2 publications and receive 3.04 yearly citations per publication. They count on a network of 28.40 distinct co-authors. Almost half of those students staying in academia move outside France and 11% are affiliated with a US university or research center.

Table 2: Descriptive statistics of the dependent variables

Entire sample: 46,774 observations				
	Mean	SD	Min	Max
Similarity	0.00	1.00	-3.94	2.19
Academic career	0.51	0.50	0.00	1.00
At least one patent	0.08	0.28	0.00	1.00

Subsample of students who stay in academia after graduation:24,088 observations				
	Mean	SD	Min	Max
Number of publications	7.20	8.52	1.00	256.00
Average citations per article	3.04	4.88	0.00	223.57
Number of distinct co-authors	28.40	44.11	0.00	1425.00
Having a foreign affiliation	0.47	0.50	0.00	1.00
Having a US affiliation	0.11	0.31	0.00	1.00

In all our econometric models, we include a set of regressors for the student’s and supervisor’s characteristics.

Among the student’s characteristics, we consider the student’s academic profile. We flag those students who published at least one paper during their PhD period (*At least one pub. during the PhD period*= 1, 0 otherwise), and among those students, we distinguish those who publish with their supervisor (*At least one pub. co-authored with the supervisor during the PhD period*=1, 0 otherwise). We control for being co-supervised during the PhD (*Having a co-supervisor*=1, 0 otherwise) and for the field of study (*Math*, *Med-bio-chem*, *Physics*, and *Engineering*). Table 3 shows that in our sample the 39% of the students are female, most of the students has published at least one paper during the PhD period (72%), and more than half of the student count a publication co-authored with her supervisor (64%). While having only one supervisor remains common practice for PhD students, 37% of them have a co-supervisor. Looking at the share of students by field, a large part of our students belongs to the Medicine, Biology and

Chemistry field (42%), 26% obtains a PhD in Engineering, 17% in Physics and the remaining in Math (14%).

Table 3: Descriptive statistics of the regressors

	Entire sample (46,774)		Subsample of students who start an academic career (24,088)	
	Mean	SD	Mean	SD
Student's characteristics				
Female student	0.39	0.49	0.39	0.49
At least one pub. during the PhD period	0.72	0.45	0.91	0.28
At least one pub. co-authored with the supervisor during the PhD period	0.64	0.48	0.83	0.38
Having a co-supervisor	0.37	0.48	0.37	0.48
Engineering	0.26	0.44	0.2	0.4
Math	0.14	0.35	0.13	0.33
Med-bio-chem	0.42	0.49	0.49	0.5
Physics	0.17	0.38	0.17	0.38
Supervisor's characteristics				
Female supervisor	0.22	0.42	0.24	0.43
Number of publications during the PhD period	21.94	20.62	22.46	20.78
Average yearly citations received per article	7.2	3.26	7.5	3.25
Number of distinct co-authors during the PhD period	58.16	76.55	62.51	79.29
Supervisor-Student pair's characteristics				
Female supervisor - Female student	0.11	0.31	0.11	0.31
Male supervisor - Female student	0.29	0.45	0.28	0.45
Female supervisor - Male student	0.12	0.32	0.13	0.33
Male supervisor - Male student	0.49	0.50	0.49	0.50

Note: The table presents the regressors' statistics for the entire sample of 46,774 PhD students and for the subsample of 24,088 PhD students who starting an academic career.

As supervisor's characteristics, we control for the supervisor's productivity and academic quality. We counted the number of papers published during the students' PhD period (*Number of publications during the PhD period*), the visibility of those publications in terms of yearly citations received (*Average yearly citations received per article*), and the dimension of the scientific network (*Number of distinct co-authors during the PhD period*). Including among regressors, the variable *Number of publications during the PhD period* is particularly important for obtaining unbiased coefficients in our econometric analysis explaining the PhD-supervisor similarity score. Indeed, a higher number of publications increases the likelihood of finding one publication similar to the student's PhD thesis content. Table 3 summarizes the profiles of the supervisors in our sample. 22% of them are female. Supervisors are established and well-reputed scholars who publish on average 21.94 papers that receive 7.2 yearly citations. Their scientific networks are quite large, having on average 58.16 co-authors each.

Following previous studies considering the role of mentor's gender in coaching students (Pezzoni *et al.* 2016, Gaule and Piacentini 2018), we include in our analyses four dummy variables representing all the possible combinations in the gender paring supervisor-student[5]. Specifically, the dummy variable *Female supervisor-Female student* equals one if both supervisor and student are female researchers, zero otherwise. We define the three dummy variables *Male supervisor-Female student*, *Female supervisor-Male student,* and *Male supervisor-Male student* similarly.

As additional controls, we include the defense year and PhD university dummies. Defense year dummy variables allow controlling for the scientists' labor market cycle[6]. We also include PhD university dummy variables to control for the unobservable characteristics of the university where the PhD is enrolled. Specifically, we define a dummy variable for all the universities with the largest PhD programs, such as those located in *Paris* (22.6% of the observations), *Toulouse, Lyon*, *Grenoble*, *Marseille*, *Strasbourg*, *Bordeaux*, *Montpellier, Rennes,* and *Lille*. Then, we define a residual category including all the remaining French universities (*Other universities*).

4. Empirical strategy

In this study, we conduct two sets of analyses. First, we analyze the factors driving research independence during the PhD period. Then, we explore how the independence level reached during the PhD associates with students' career paths. Finally, conditional on starting an academic career, we look at the effect of independence on further academic career outcomes.

To analyze the factors associated with research independence, we estimate an Ordinary Least Squares (OLS) regression having *Similarity* as the dependent variable and three vectors

[5]We associate each student's given name to the corresponding gender using a multiple-iteration matching strategy (Gaule and Piacentini, 2018; OECD, 2012). First, we match the student's given name with the names listed in the official French gender-name dataset (https://www.insee.fr/). Then, for the non-matched students, we repeated the matching procedure using the U.S. Census Bureau gender-name dataset (https://www.ssa.gov/) and the WIPO gender-name dataset (https://www.wipo.int/). We apply a similar procedure to identify the gender of the supervisors.

[6] Including the year dummies allow us also to control for the cohort unbalanced composition of our final study sample. Due to the restriction of selecting only theses with an English abstract, we observe a higher proportion of students belonging to the cohorts 2005-2013 than in the earlier years. To respond to this concern, in Appendix 2, we also run an additional analysis on a subset of theses defended between 2005 and 2013. The results are consistent with those reported in the main text.

including student's, supervisor's, and student-supervisor pair's characteristics as independent variables. Equation 1 represents the estimated model:

$$\begin{aligned} Similarity_i = \beta_0 &+ \beta_1 Student's\ characteristics_i + \beta_2 Supervisor's\ characteristics_i \\ &+ \beta_3 Supervisor\ Student\ pair's\ characteristics_i + \gamma\ Defence\ year_i \\ &+ \delta\ University_i + \varepsilon_i \end{aligned}$$

(Equation 1)

The vector of student's characteristics includes the variables *At least one publication during the PhD period, At least one publication co-authored with the supervisor during the PhD period, Having a co-supervisor,* and field of study (*Engineering, Math, Medicine-Biology-Chemistry,* and *Physics*). The vector of supervisor's characteristics includes the variables *Number of publications during the PhD period, the Average yearly citations per article, and the Number of distinct co-authors during the PhD period*. The vector student-supervisor pair's characteristics includes the variables *Female supervisor-Female student, Male supervisor-Female student, Female supervisor-Male student,* and *Male supervisor-Male student*. Finally, we add to our model a set of *Defence year* dummy variables and *University* dummy variables.

In the second set of analyses, we explore how independence associates with students' career paths. We estimate two Logistic regressions relating *Similarity* to the probability of pursuing an academic career and being involved in patenting activity, respectively. Equation 2 shows the estimated model. The dependent variable, *Career path*, takes the values of either *Academic career* or *At least one patent* according to the type of career considered. The vectors of controls in Equation 2 are the same as in Equation 1.

$$\begin{gathered} Career\ path_i = \beta_0 + \beta_1 Similarity_i + \beta_2 Student's\ characteristics_i + \\ \beta_3 Supervisor's\ characteristics_i + \beta_4 Supervisor\ Student\ pair's\ characteristics_i + \\ \gamma\ Defence\ year_i + \delta\ University_i + \varepsilon_i \end{gathered}$$

(Equation 2)

Finally, we further explore the career outcomes for those students who stay in academia. For this sub-sample, we explore the association between *Similarity* and five different career outcomes, the *Number of publications,* the *Average citations per article,* the *Number of distinct co-authors, Having a foreign affiliation* and, *Having a US affiliation*. Equation 3 shows the estimated model. The dependent variable, *Academic career outcome*, takes, in turn, the values of the five features considered while the vectors of the controls are the same as the ones presented in Equation 1. We use the OLS estimator when the dependent variable is continuous, while we use the Logit estimator when the dependent variable is a dummy variable.

$$Academic\ career\ outcome_i = \beta_0 + \beta_1 Similarity_i + \beta_2 Student's\ characteristics_i + \beta_3 Supervisor's\ characteristics_i + \beta_4 Supervisor\ Student\ pair's\ characteristics_i + \gamma\ Defence\ year_i + \delta\ University_i + \varepsilon_i\ |_{Academic\ career_i=1}$$

(Equation 3)

5. Results

Similarity determinants

Table 5 reports the OLS estimates of the model described in Equation 1. Results show how PhD student and supervisor's basic characteristics relate to the degree of similarity between the student's thesis and her supervisor's research work.

Looking at the student's characteristics, we find that publishing at least one article during the PhD period is associated with higher independence from the supervisor's research work. Students who publish at least one article are associated with a lower similarity score than those not publishing (0.46 fewer standard deviations). However, if the article published is co-authored with the supervisor, it depicts the opposite scenario. Indeed, having at least one article co-authored with the supervisor is associated with a 0.56 standard deviations higher similarity score (-0.46+1.02)[7]. Not surprisingly, these results show that student's publications outcomes during the PhD highly correlate with the independence level reached by the student. Students who publish without their supervisors are more independent than colleagues who co-author with their supervisors. Having a co-supervisor is negatively associated with the similarity between student and supervisor's work. Specifically, having a co-supervisor is associated with a similarity score 0.027 standard deviations lower than not having a co-supervisor. The negative sign is explained by the practice of supervisors lacking competencies on their student's thesis subjects of delegating the supervision to colleagues[8]. Concerning disciplines, our estimates show heterogeneity across fields. We find that students in mathematics tend to

[7] The linear combination of the coefficients of the variables *At least one pub. during the PhD period* and *At least one pub. co-authored with the supervisor during the PhD period* is statistically significant at 1% level (t-statistic =57.41).

[8] In the regression exercise presented in Table 5, we calculate the similarity between student's thesis and supervisor's work, not considering the similarity with the co-supervisor. In Appendix 3, we run a robustness check where we compute similarity both to the supervisor and co-supervisor's work and we assign to the variable *Similarity* the highest value between the two scores. Results are aligned with the ones presented in Table 5.

have theses more similar to their supervisors' work, while students in medicine, biology, and chemistry tend to have theses less similar to their supervisors' work.

All supervisor's productivity proxies (number of publications, average citations, and co-authors) are associated with a higher similarity of the student's thesis to the supervisor's work. Specifically, one additional publication is associated with a 0.0053 standard deviations higher similarity score, one additional citation with +0.013 standard deviations, and one additional co-author with +0.00023 standard deviations. A positive correlation between the number of publications and similarity was expected since the likelihood of finding one publication similar to the thesis increases when we compare the thesis with a larger number of supervisors' publications. However, there might be a more substantial reason explaining the positive correlation. Students might be more likely to choose a thesis subject close to their supervisors' research work when supervisors are highly productive scientists with a high reputation and visibility within the discipline. In this case, students can benefit from the supervisor's prestige within the scientific community. This latter explanation is consistent with the positive signs observed for the coefficients of supervisors' citations and network size.

Our estimates show heterogeneous associations between gender pairing and the similarity level between student and supervisor's work. The pair *Female supervisor – Female student* is associated with a 0.12 standard deviations higher similarity score, if compared with the baseline *Male supervisor – Male student* pair. The latter corresponds to the largest similarity variation among the four possible gender pairs.

Table 5: Regression explaining Similarity

	(1) OLS Standardized Similarity
Student's characteristics	
At least one pub. during the PhD period	-0.46***
	(0.017)
At least one pub. co-authored with the supervisor during the PhD period	1.02***
	(0.016)
Having a co-supervisor	-0.027***
	(0.0091)
Math	0.027*
	(0.014)
Med-bio-chem	-0.33***
	(0.012)
Physics	-0.25***
	(0.013)
Engineering	Ref.
Supervisor's characteristics	
Number of publications during the PhD period	0.0053***
	(0.00028)
Average yearly citations received per article	0.013***
	(0.0016)
Number of distinct co-authors during the PhD period	0.00023***
	(0.000079)
Supervisor-Student pair's characteristics	
Female supervisor - Female student	0.12***
	(0.015)
Male supervisor - Female student	0.039***
	(0.010)
Female supervisor - Male student	0.030**
	(0.014)
Male supervisor - Male student	Ref.
Defense year dummies	Yes
University dummies	Yes
Constant	-0.82***
	(0.089)
Observations	46,774
R-squared	0.178

NOTE: The table presents OLS estimates of Equation 1. Standard errors are reported in parentheses. Significance level are defined as follows: *** p<0.01, ** p<0.05, * p<0.1

Similarity and career paths

Table 6 shows that the PhD-supervisor similarity score negatively associates with the probability of starting a career in academia and patenting. We find that one standard deviation increase in the similarity score is associated with a 2.1 percentage point lower probability of starting an academic career. Likewise, one standard deviation increase in the similarity score is associated with a 0.57 percentage point lower probability of observing at least one patent application by the students after the thesis defense. Although the latter result appears to have a limited economic impact, we have to compare the value with the unconditional probability of observing a patent application to understand its relevance. A value of 0.57 percentage point variation represents 7% of the unconditional probability of observing a parent application (8 percentage points as shown in Table 2).

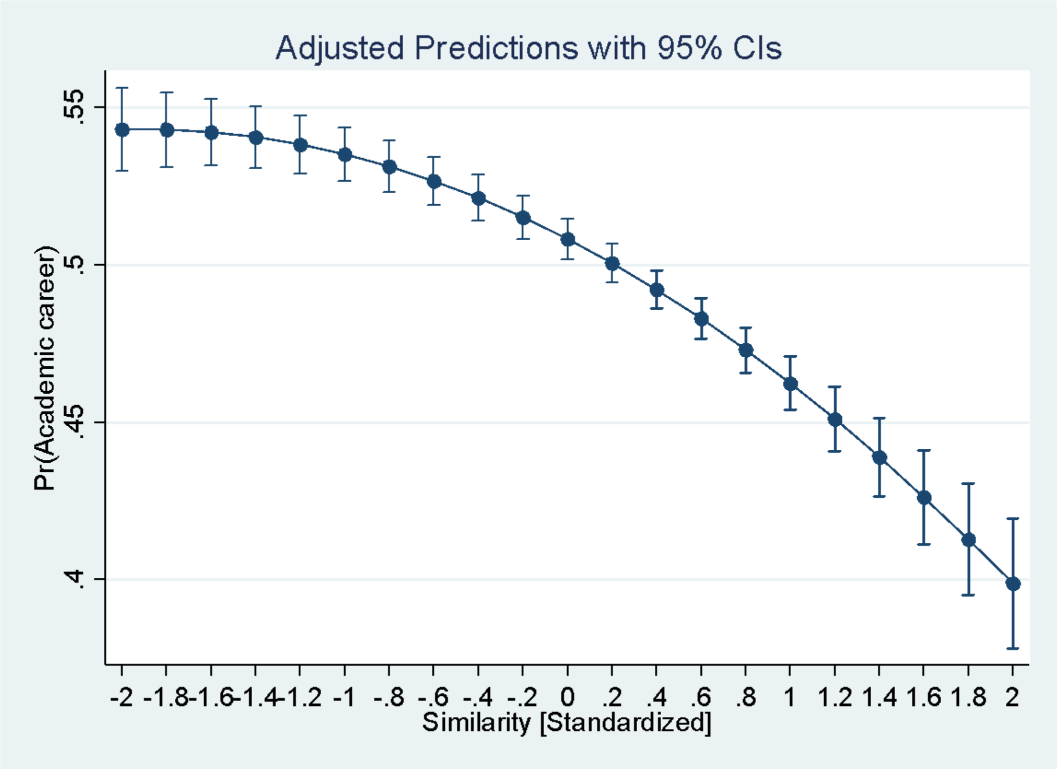
Table 6: Regression estimating the impact of similarity on starting an academic career and patenting

	(1) Logit Academic career	(2) Logit At least one patent
Similarity	-0.021***	-0.0057***
	(0.0023)	(0.0014)
Student's characteristics		
At least one pub. during the PhD period	0.38***	0.044***
	(0.0075)	(0.0052)
At least one pub. co-authored with the supervisor during the PhD period	0.094***	-0.0060
	(0.0074)	(0.0048)
Having a co-supervisor	0.023***	-0.0049*
	(0.0044)	(0.0028)
Math	0.036***	-0.058***
	(0.0067)	(0.0047)
Med-bio-chem	0.12***	-0.012***
	(0.0058)	(0.0035)
Physics	0.096***	-0.025***
	(0.0065)	(0.0040)
Engineering	Ref.	Ref.
Supervisor's characteristics		
Number of publications during the PhD period	-0.00057***	0.0010***
	(0.00013)	(0.000071)
Average yearly citations received per article	0.0086***	0.0011**
	(0.00084)	(0.00050)
Number of distinct co-authors during the PhD period	0.00019***	-0.00017***
	(0.000039)	(0.000026)
Supervisor-Student pair's characteristics		
Female supervisor - Female student	-0.019***	-0.055***
	(0.0072)	(0.0054)
Male supervisor - Female student	-0.029***	-0.045***
	(0.0049)	(0.0034)
Female supervisor - Male student	0.013**	0.0046
	(0.0066)	(0.0038)
Male supervisor - Male student	Ref.	Ref.
Defense year dummies	Yes	Yes
University dummies	Yes	Yes
Constant	-	-
Observations	46,774	46,774
Pseudo R-squared	0.177	0.039

NOTE: The table presents Logit estimates of Equation 2. Marginal effects are reported. Standard errors are reported in parentheses. Significance level are defined as follows: *** p<0.01, ** p<0.05, * p<0.1

To further investigate the association between the PhD-supervisor similarity score and the probability of starting an academic career and patenting, we analyze its curvilinear effect by including the *Similarity* squared term in the model specification of Equation 2. Figures 2 and 3 report the predicted probability of an average student starting an academic career and patenting, respectively[9]. Figure 2 shows that, when considering a wide range of variation of the similarity score from -2 to +2 standard deviations, keeping all the other characteristics of the student/supervisor constant at their average level, the probability of starting academic career decreases by more than 10 percentage points. This result confirms the non-negligible impact of the PhD-supervisor similarity score. Likewise, in Figure 3, we observe that for the same range of variation from -2 to +2 standard deviations, the probability of observing at least one patent varies more than 2 percentage points, representing a non-negligible share (about 25%) of the unconditional probability of observing at least one patent application.

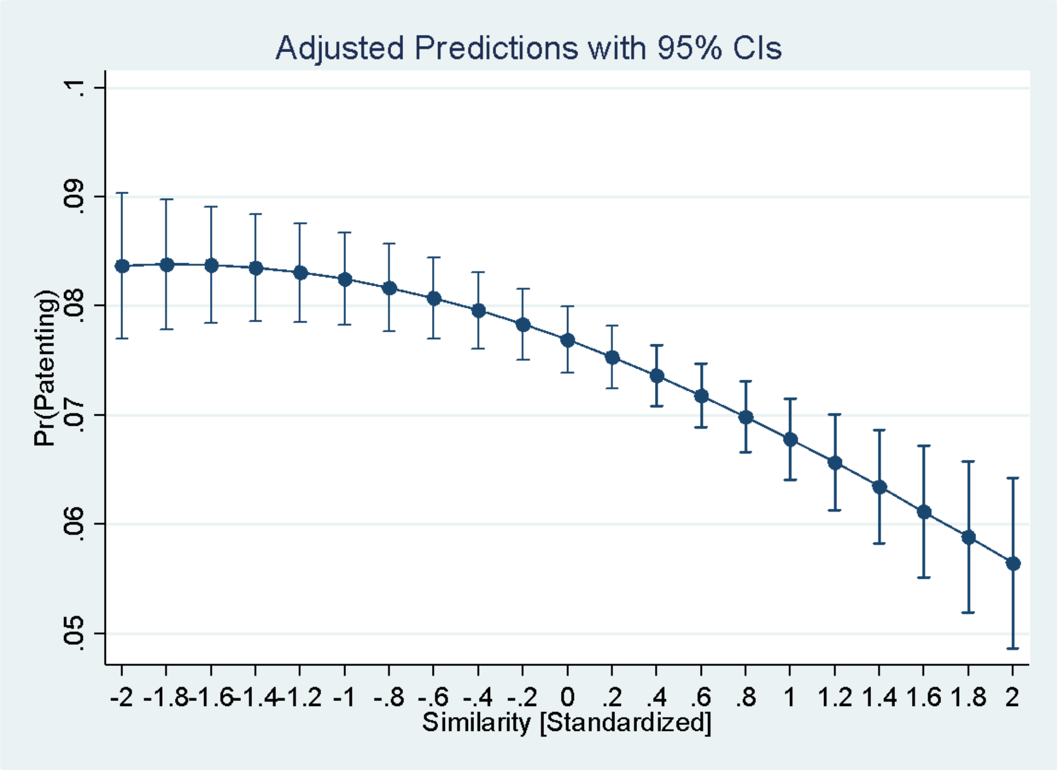
Figure 2: Predicted probability of starting an academic career for the average PhD student



NOTE: The figure shows the predicted probability of starting an academic career for an average student. The predictions are based on the model estimated in Appendix 4, Table A4.1, column 1. The average student is characterized by the average values of the covariates reported in Table 2.

[9] Appendix 4 reports the detailed model estimations.

Figure 3: Predicted probability of patenting for the average PhD student



NOTE: The figure shows the predicted probability of patenting for an average student. The predictions are based on the model estimated in Appendix 4, Table A4.1, column 2. The average student is characterized by the average values of the covariates reported in Table 2.

We find results in line with our expectations concerning the student and supervisor's characteristics included as controls. Publishing at least one paper during the PhD is associated with a 38 percentage point higher probability of starting an academic career and a 4.4 percentage point higher probability of patenting. We also observe that co-publishing with the supervisor during the thesis further increases the associated probability of starting an academic career by 9.4 percentage points. In contrast, it does not increase the probability of patenting. Co-authoring a paper with the supervisor is likely to secure the student's access into the supervisor's academic network and community favoring her career. Having a co-supervisor is positively associated with the probability of staying in academia, while it is negatively associated with patenting. Discipline dummies are in line with expectations. Engineers show a lower probability of staying in academia due to their high employability in alternative remunerative jobs. Students in the engineering field also associate with the highest probability of observing patenting activity after the PhD due to the intrinsic patentability nature of their activities.

Concerning the supervisor's characteristics included as controls, surprisingly, the number of publications during the PhD period negatively associates with the student's probability of starting an academic career. On the contrary, in line with our expectations, citations received by the supervisor's work and supervisors' professional network size are positively associated with the probability of starting an academic career. When considering the probability that the student patents after the PhD, we observe a positive association with the supervisor's publications and citations received. On the contrary, network size negatively associates with the probability of patenting.

As for the student-supervisor pair characteristics, Female students show lower probabilities than male students of starting an academic career and patenting either if a female or a male supervisor supervises them.

Similarity and academic career outcomes

Conditional on starting an academic career, Table 7 shows that the PhD-supervisor similarity score associates with academic careers characterized by a higher number of citations received, a larger network size, and a higher probability of experiencing international mobility. Increasing the PhD-supervisor similarity score by one standard deviation associates with 0.23 additional citations (0.75% of the sample average), 0.52 additional co-authors (1.83% of the sample average), 1.7 percentage points higher probability of having a foreign affiliation, and 1.4 percentage points higher probability of having a US-based affiliation. Only the number of publications authored by the student after graduation does not associate with similarity. Our results show a substantial advantage for less independent students on their follow-on academic careers. Indeed, students who choose thesis topics close to their supervisor's research subjects during the PhD experience more successful academic careers according to the standard bibliometric indicators observed. Comparing these results with the ones obtained in Table 6 shows a multifaceted effect of independence on career achievements of PhD students after graduation. Although higher independence increases the probability of starting an academic career path, it is detrimental to academic career outcomes as measured by the standard bibliometric indicators.

Concerning the control variables, our results show that students publishing a paper during the PhD period are more productive in their follow-on academic careers and are more likely to experience international mobility. On the contrary, having a co-supervisor and the discipline to which the student belongs show heterogeneous results. For instance, having a co-supervisor is associated with a higher publication score in the student’s follow-on academic career and the probability of having an affiliation with a foreign university. On the other hand, having a co-supervisor is associated negatively with the student's citations and the probability of being affiliated with a US university. Supervisor’s characteristics (articles published, citations received, and network size) are positively associated with almost all the student’s career features considered. Finally, for the student-supervisor pair characteristics, we observe that female students show lower publication scores, citations, and network size than their male counterparts during their follow-on academic careers. This result holds both if a female or a male supervisor supervises them. Female students also show a lower probability of having a foreign affiliation or a US affiliation after graduation. These results are coherent with those presented in Table 6, highlighting a lower probability of starting a research career for female students.

Table 7: Regression estimating the impact of similarity on academic career outcomes

	(1)	(2)	(3)	(4)	(5)
	Conditional on starting an academic career				
	OLS	OLS	OLS	Logit	Logit
		Avg cit	N. distinct	Foreign	US
	N. pubs	per pub	co-authors	affiliation	affiliation
Similarity	0.093	0.23***	0.52*	0.017***	0.014***
	(0.061)	(0.035)	(0.30)	(0.0037)	(0.0024)
Student's characteristics					
At least one pub. during the PhD period	2.60***	0.33**	10.4***	0.065***	0.057***
	(0.25)	(0.15)	(1.25)	(0.015)	(0.012)
At least one pub. co-authored with the supervisor during the PhD period	0.064	-0.13	-1.17	0.013	-0.0041
	(0.20)	(0.12)	(1.00)	(0.012)	(0.0080)
Having a co-supervisor	0.32***	-0.16**	-0.0075	0.047***	-0.015***
	(0.11)	(0.066)	(0.56)	(0.0068)	(0.0045)
Math	0.13	-0.29***	-2.67***	0.075***	0.021**
	(0.19)	(0.11)	(0.95)	(0.012)	(0.0085)
Med-bio-chem	-1.16***	1.36***	4.34***	0.065***	0.049***
	(0.16)	(0.090)	(0.78)	(0.0095)	(0.0067)
Physics	1.17***	0.64***	10.4***	0.14***	0.060***
	(0.18)	(0.10)	(0.88)	(0.011)	(0.0072)
Engineering	Ref.	Ref.	Ref.	Ref.	Ref.
Supervisor's characteristics					
Number of publications during the PhD period	0.044***	-0.0024	-0.27***	0.00091***	0.00012
	(0.0035)	(0.0020)	(0.017)	(0.00021)	(0.00012)
Average yearly citations received per article	0.0082	0.32***	0.37***	0.011***	0.0057***
	(0.021)	(0.012)	(0.10)	(0.0013)	(0.00068)
Number of distinct co-authors during the PhD period	0.011***	0.0020***	0.24***	0.000012	0.000098***
	(0.00097)	(0.00055)	(0.0048)	(0.000059)	(0.000030)
Supervisor-Student pair's characteristics					
Female supervisor - Female student	-1.91***	-0.19*	-5.34***	-0.095***	-0.036***
	(0.18)	(0.11)	(0.91)	(0.011)	(0.0073)
Male supervisor - Female student	-2.23***	-0.10	-6.79***	-0.076***	-0.019***
	(0.13)	(0.074)	(0.64)	(0.0077)	(0.0049)
Female supervisor - Male student	-0.087	0.17*	1.71**	-0.0011	0.0037
	(0.17)	(0.096)	(0.82)	(0.0099)	(0.0061)
Male supervisor - Male student	Ref.	Ref.	Ref.	Ref.	Ref.
Defense year dummies	Yes	Yes	Yes	Yes	Yes
University dummies	Yes	Yes	Yes	Yes	Yes
Constant	2.52*	0.19	-1.70	-	-
	(1.29)	(0.74)	(6.36)		
Observations	24,088	24,088	24,088	24,088	24,088
R-squared / Pseudo R-squared	0.084	0.084	0.170	0.028	0.035

NOTE: The table presents estimates of Equation 3. OLS estimates are reported in columns 1, 2, and 3 and Logit estimates (marginal effects) in columns 4 and 5. Standard errors are reported in parentheses. Significance level are defined as follows: *** p<0.01, ** p<0.05, * p<0.1

5.1 Further analysis: The moderating effects of gender and discipline

This section explores the possible moderating effects of gender and discipline on the PhD-supervisor similarity score.

Appendix 5 assesses the supervisor-student gender pairing moderating effect. Interaction terms between gender and *Similarity* in Column 1 of Table A5.1, show that the PhD-supervisor similarity score negatively associates with the probability of starting an academic career for all the possible gender pairs. On the contrary, only male students show a negative association between similarity and probability of patenting. When we look at the career features for those students who start an academic career, Table A5.2 shows homogenous similarity effects across gender pairs when it explains the average number of citations received by the student after graduating and the probability of undertaking an international career. However, columns 1 and 3 of Table A5.2 show that an increase of *Similarity* associates with more publications only for male supervisor-male student pairs, while it associates with a higher number of coauthors only for female supervisor-male supervisor pairs.

Appendix 6 assesses the discipline moderating effect. Interaction terms between discipline dummy variables and *Similarity* in Column 1 of Table A6.1, show that *Similarity* negatively associates with the probability of starting an academic career in all disciplines except physics. One possible explanation for this result might rely on how research work is organized in physics, especially applied physics. Indeed, it might be difficult for a PhD student to propose independent research since physics research has a prevalent team dimension, and the student's work must be integrated into the lab research plan. The difficulty of proposing independent research subjects for PhD students in physics might be considered by the recruiting committees which overlook this aspect when deciding for hiring. When looking at the probability of patenting, *Similarity* negatively associates with patenting in Engineering and Med-Bio-Chem disciplines. At the same time, it shows no association and a positive association in Physics and Math, respectively. When we look at the career outcomes for those students who start an academic career, Table A6.2 shows limited effects of *Similarity* for the career features of students who start an academic career in engineering and physics. On the contrary, in Med-Bio-Chem and Math, *Similarity* correlates with several of the career outcomes considered. For instance, in Med-Bio-Chem and Math disciplines, *Similarity* is associated positively with the citations received during the academic career and the probability of observing US affiliation.

6. Conclusions

Understanding the determinants of students' research independence from the supervisor and the effect of independence on career outcomes is crucial for designing appropriate science and technology policies shaping the research workforce's career paths. By investigating the association between research independence and the likelihood of starting an academic career and patenting, this study sheds light on how PhD training shapes career outcomes.

We conducted our empirical analysis by relying on unique and detailed information on PhD students' thesis abstracts from the national-wide French repository of Electronic Doctoral Theses and students' and supervisors' publication records from Elsevier's SCOPUS database. Our data cover the entire student population in France for all the STEM fields.

In exploring the factors leading PhD students to develop their independence, we observe that students working with highly productive supervisors have higher incentives to follow supervisors' lines of research. Interestingly, also the gender pairing supervisor-PhD student matters. Our results show that female students having female supervisors have the highest level of similarity between the student's thesis and the supervisor's work. Looking at the students' career paths, we find that a higher level of research independence reached during the PhD training is associated with a higher probability of starting an academic career or being involved in R&D activities. This latter result is in line with the general expectations promoted in the job market. Often, candidates are asked to prove their independence as young scholars by presenting solo-author job market papers or encouraged by funding agencies to submit projects showing their individual contributions to the field. However, when examining the career outcomes of students who pursue an academic career, we observe that a greater level of independence with supervisors is negatively associated with students' work visibility, scientific network size, and probability of experiencing international mobility. Students with a higher level of independence tend to receive fewer citations, have a higher likelihood of remaining in France, and have fewer co-authors. These later results show the existence of a trade-off between entering academia and succeeding in academia in the first five years after graduation observed. It might be the case that more independent students have to pay the price for their independence. In the short run, i.e., 5 years after the PhD graduation, it seems more challenging for independent students to be productive and internationally mobile than students following the supervisor research avenue and, therefore, leveraging on supervisor's reputation and network.

Our results have implications for the policy approach towards the independence of young researchers. On one side, independence appears as a relevant dimension in the academic selection process. On the other side, our results show a potential warning for employers and funding agencies when drafting expectations for newly hired young scholars and funding awardees. Specifically, young scholars who are required to develop a high independence level to be selected for an academic career are also required to achieve high-quality research standards in the short run during their tenure track periods or when reporting the outcomes of their projects. Our study shows that for students with a high level of independence during the thesis is challenging to achieve those results. Should researchers who have shown independent work during the early stages of their career benefit from a discount when evaluated for tenured positions or when reporting the outcomes of their projects? This is a question that recruiters and policymakers should address.

Our study shed light on the factor leading a young scholar to become independent, and it also opens a call for further studies on the mechanisms behind some of our results. In particular, the mechanisms driving the decisions to undertake an independent research line during the PhD are unexplored. For instance, we document that female students are more reluctant to develop independent research lines than their male counterparts, especially when supervised by female researchers. What are the mechanisms driving this behavior? This is a question that deserves further reflection within the scientific community.

References

Austin Ann, E. (2002). Preparing the next generation of faculty. *The Journal of Higher Education*, *73*(1).

Becher, T. (1990). Physicists on physics. *Studies in Higher Education*, *15*(1), 3-20.

Campbell, R. A. (2003). Preparing the next generation of scientists: The social process of managing students. *Social studies of Science*, *33*(6), 897-927.

Corsini, A., Pezzoni, M., & Visentin, F. (2021). *What makes a productive Ph. D. student?* (No. 2021-11). Groupe de REcherche en Droit, Economie, Gestion (GREDEG CNRS), Université Côte d'Azur, France.

Diamond, A. M. (1986). The life-cycle research productivity of mathematicians and scientists. Journal of gerontology, 41(4), 520-525.

Delamont, S., & Atkinson, P. (2001). Doctoring uncertainty: Mastering craft knowledge. *Social studies of science*, *31*(1), 87-107.

Gardner, Susan K. 2008. "What's Too Much and What's Too Little?" Journal of Higher Education 79 (3): 326–50

Garcia-Romero, A., Modrego, A., 2001. Research training in Spain: an assessment exercise. In: Proceedings of the Conference "The contribution of socio-economic research to the benchmarking of RTD policies in Europe". Brussels, March 15–16.

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

Horta, H., Cattaneo, M., & Meoli, M. (2018). PhD funding as a determinant of PhD and career research performance. *Studies in Higher Education*, *43*(3), 542-570.

Horta, H., & Santos, J. M. (2016). The impact of publishing during PhD studies on career research publication, visibility, and collaborations. *Research in Higher Education*, *57*(1), 28-50.

Kam, B. H. (1997). Style and quality in research supervision: the supervisor dependency factor. *Higher education*, *34*(1), 81-103.

Laudel, G., & Bielick, J. (2018). The emergence of individual research programs in the early career phase of academics. *Science, Technology, & Human Values*, *43*(6), 972-1010.

Laudel, G., & Gläser, J. (2008). From apprentice to colleague: The metamorphosis of early career researchers. *Higher education*, *55*(3), 387-406.

Latour, B., & Woolgar, S. (2013). *Laboratory life*. Princeton University Press.

Lee, A. (2008). How are doctoral students supervised? Concepts of doctoral research supervision. *Studies in higher education*, *33*(3), 267-281.

Lee, A., Dennis, C., & Campbell, P. (2007). Nature's guide for mentors. *Nature*, *447*(7146), 791-797.

Levin, S. G., & Stephan, P. E. (1998). Gender Differences in the Rewards to Publishing in Academe: Science in the 1970s. *Sex Roles*, *38*(11), 1049-1064.

Liénard, J. F., Achakulvisut, T., Acuna, D. E., & David, S. V. (2018). Intellectual synthesis in mentorship determines success in academic careers. *Nature communications*, *9*(1), 1-13.

Mainhard, T., Van Der Rijst, R., Van Tartwijk, J., & Wubbels, T. (2009). A model for the supervisor–doctoral student relationship. *Higher education*, *58*(3), 359-373.

Mangematin, V. (2000). PhD job market: professional trajectories and incentives during the PhD. *Research policy*, *29*(6), 741-756.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

National Research Council. (2005). Bridges to independence: Fostering the independence of new investigators in biomedical research.

OECD, 2012. Closing the gender gap: Act now. OECD Publishing Paris.

Pezzoni, M., Mairesse, J., Stephan, P., & Lane, J. (2016). Gender and the publication output of graduate students: A case study. *PLoS One*, *11*(1), e0145146.

Shibayama, S. (2019). Sustainable development of science and scientists: Academic training in life science labs. *Research Policy*, *48*(3), 676-692.

Stephan, P. E., & Levin, S. G. (1997). The critical importance of careers in collaborative scientific research. *Revue d'économie industrielle*, *79*(1), 45-61.

Stephan, P. E. (1996). The economics of science. *Journal of Economic literature*, *34*(3), 1199-1235.

Stephan, P. (2012). *How economics shapes science*. Harvard University Press.

Tenenbaum, H. R., Crosby, F. J., & Gliner, M. D. (2001). Mentoring relationships in graduate school. *Journal of vocational behavior*, *59*(3), 326-341.

Yoshioka-Kobayashi, T., & Shibayama, S. (2020). Early career training and development of academic independence: a case of life sciences in Japan. *Studies in Higher Education*, 1-23.

Waldinger, F. (2010). Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany. *Journal of political economy*, *118*(4), 787-831.

Weiss, Y. (1981), Output variability, academic labour contracts and waiting times for promotion", Technical Report

Zuckerman, H., & Merton, R. K. (1972). Age, aging, and age structure in science. *Higher Education*, *4*(2), 1-4.

Appendix 1: Calculating the similarity variable, an example

In this section, we present an example to clarify the interpretation of the similarity score values. Student A is a Ph.D. student of Professor B. He graduated in 2007 from the University of Nantes. His thesis abstract reads as follows:

"When designing or re-assessing an offshore structure, one of the most delicate stages relates to the calculation of the solicitations : actions exerted by the swell, the wind and the currents. It comes partly from the randomness or uncertainties that concern the marine environment as well as the modelling of loading. Presence of marine growth makes these questions more sensitive. The generic term of marine growth includes the vegetable species (algae ...) and animal (mussel, anemones, corals ...). Indeed very quickly, the structures are covered of a multitude of marine organisms. It remains particularly difficult to quantify this phenomenon by taking into account the diversity of the organism, the seasonal conditions and the competition to which the various species for their survival are delivered. Its influence on an offshore structure can be measured on several levels : obstruct or prohibits a visual inspection of the subjacent support, expensive procedures of cleaning for oil industries, increase in the hydrodynamic efforts on the level of the structure. This work aims to provide a probabilistic modelling of marine growth evolution in five regions in the Gulf of Guinea. A physical matrix response surface is then built in view to provide a probabilistic modelling of the environmental loading on Jacket offshore structures in presence of marine growth. A study case allows performing sensitivity and uncertainty studies in view to improve, supplement, integrate and make more operational the methods and tools for structural reassessment."

Student A shows a very low level of *Similarity* with his supervisor work (-3.4). During the Ph.D. period, A did not publish any paper with his supervisor and had a co-supervisor (Professor C) whose research was similar to A's thesis (0.42)[10].

Student Y is also a Ph.D. student of Professor B. She graduated in 2005 from the University of Nantes. Her thesis abstract reads as follows:

"The materials of inflatable structures have a behavior of the membrane type in general. Because of their lack of bending stiffness, these membrane structures are often partially wrinkled at each time they are in compression. One develops a pure finite element membrane without bending stiffness to study such structures. The modeling of the folds is done under the pure bifurcation theory. For this purpose, one modifies the standard arclength method by adding a procedure to treat the presence of the complex roots in the quadratic equation of the second degree. In another hand, one studies the buckling of the pressurized tubes through a finite element inflatable beam. The construction of this element is based on the assumptions of Timoshenko and finite rotation. From simplifying assumptions , one establishes the linearized equations and a symmetric stifness matrix which present the effect of pressure. This finite element inflatable beam is compared with the 3D finite element membrane through a series of

[10] The title of the paper published by the co-supervisor during the PhD period is "Sensitivity approach for modelling the environmental loading of marine structures through a matrix response surface". It is a sole author publication.

examples. Finally, experiments were conducted to obtain the buckling of the pressurized tubes. The experimental results obtained are compared with the two previous numerical models."

Student Y shows a high level of Similarity with his supervisor work (0.38). During the Ph.D. period, Y published two papers co-authored with her supervisor[11].

Professor B published 7 and 11 papers during the Ph.D. period of student A and student Y, respectively.

[11] The titles of the two papers co-authored by professor B and student A are "Buckling and wrinkling of prestressed membranes" and "Modélisation du plissage dans les structures membranaires".

Appendix 2: Restricting our sample to the theses defended between 2005 and 2013

Table A2.1: Regression explaining Similarity

	(1) OLS Standardized Similarity
Student's characteristics	
At least one pub. during the PhD period	-0.47***
	(0.018)
At least one pub. co-authored with the supervisor during the PhD period	1.02***
	(0.017)
Having a co-supervisor	-0.030***
	(0.0095)
Math	0.038**
	(0.015)
Med-bio-chem	-0.33***
	(0.013)
Physics	-0.25***
	(0.014)
Engineering	Ref.
Supervisor's characteristics	
Number of publications during the PhD period	0.0050***
	(0.00029)
Average yearly citations received per article	0.012***
	(0.0017)
Number of distinct co-authors during the PhD period	0.00026***
	(0.000080)
Supervisor-Student pair's characteristics	
Female supervisor - Female student	0.12***
	(0.016)
Male supervisor - Female student	0.039***
	(0.011)
Female supervisor - Male student	0.033**
	(0.014)
Male supervisor - Male student	Ref.
Defense year dummies	Yes
University dummies	Yes
Constant	-0.54***
	(0.032)
Observations	46,774
R-squared	0.167

Table A2.2: Regression estimating the impact of similarity on starting academic career and patenting

	(1) Logit Academic career	(2) Logit At least one patent
Similarity	-0.023***	-0.0060***
	(0.0025)	(0.0016)
Student's characteristics		
At least one pub. during the PhD period	0.37***	0.041***
	(0.0084)	(0.0058)
At least one pub. co-authored with the supervisor during the PhD period	0.100***	-0.0031
	(0.0081)	(0.0053)
Having a co-supervisor	0.024***	-0.0042
	(0.0047)	(0.0029)
Math	0.039***	-0.058***
	(0.0072)	(0.0051)
Med-bio-chem	0.13***	-0.011***
	(0.0063)	(0.0038)
Physics	0.10***	-0.026***
	(0.0071)	(0.0043)
Engineering	Ref.	Ref.
Supervisor's characteristics		
Number of publications during the PhD period	-0.00055***	0.00100***
	(0.00014)	(0.000075)
Average yearly citations received per article	0.0085***	0.0011**
	(0.00088)	(0.00052)
Number of distinct co-authors during the PhD period	0.00018***	-0.00017***
	(0.000040)	(0.000027)
Supervisor-Student pair's characteristics		
Female supervisor - Female student	-0.018**	-0.055***
	(0.0077)	(0.0058)
Male supervisor - Female student	-0.026***	-0.045***
	(0.0054)	(0.0036)
Female supervisor - Male student	0.015**	0.0030
	(0.0071)	(0.0040)
Male supervisor - Male student	Ref.	Ref.
Defense year dummies	Yes	Yes
University dummies	Yes	Yes
Constant	-	-
Observations	46,774	46,774
Pseudo R-squared	0.170	0.039

Table A2.3: Regression estimating the impact of similarity on academic career outcomes

	(1)	(2)	(3)	(4)	(5)
	Conditional on starting an academic career				
	OLS	OLS	OLS	Logit	Logit
		Avg cit	N. distinct	Foreign	US
	N. pubs	per pub	co-authors	affiliation	affiliation
Similarity	0.12*	0.25***	0.52	0.015***	0.011***
	(0.068)	(0.040)	(0.34)	(0.0040)	(0.0027)
Student's characteristics					
At least one pub. during the PhD period	2.62***	0.39**	11.2***	0.063***	0.051***
	(0.29)	(0.17)	(1.45)	(0.017)	(0.013)
At least one pub. co-authored with the supervisor during the PhD period	-0.013	-0.15	-1.64	0.0097	0.0036
	(0.23)	(0.13)	(1.15)	(0.013)	(0.0090)
Having a co-supervisor	0.32***	-0.16**	0.10	0.048***	-0.017***
	(0.12)	(0.071)	(0.62)	(0.0072)	(0.0047)
Math	0.23	-0.31**	-2.58**	0.084***	0.022**
	(0.21)	(0.12)	(1.06)	(0.012)	(0.0091)
Med-bio-chem	-1.00***	1.34***	4.65***	0.053***	0.045***
	(0.17)	(0.10)	(0.88)	(0.010)	(0.0073)
Physics	1.15***	0.62***	10.3***	0.13***	0.062***
	(0.20)	(0.11)	(0.99)	(0.012)	(0.0078)
Engineering	Ref.	Ref.	Ref.	Ref.	Ref.
Supervisor's characteristics					
Number of publications during the PhD period	0.044***	-0.0018	-0.27***	0.00073***	0.000069
	(0.0038)	(0.0022)	(0.019)	(0.00023)	(0.00013)
Average yearly citations received per article	0.0076	0.32***	0.40***	0.011***	0.0054***
	(0.022)	(0.013)	(0.11)	(0.0014)	(0.00071)
Number of distinct co-authors during the PhD period	0.011***	0.0017***	0.23***	0.000040	0.00010***
	(0.0010)	(0.00059)	(0.0051)	(0.000061)	(0.000031)
Supervisor-Student pair's characteristics					
Female supervisor - Female student	-1.89***	-0.18	-5.49***	-0.095***	-0.034***
	(0.20)	(0.12)	(1.00)	(0.012)	(0.0077)
Male supervisor - Female student	-2.19***	-0.13	-7.02***	-0.078***	-0.017***
	(0.14)	(0.082)	(0.72)	(0.0083)	(0.0053)
Female supervisor - Male student	-0.12	0.17	1.62*	-0.0027	0.0014
	(0.18)	(0.10)	(0.91)	(0.011)	(0.0065)
Male supervisor - Male student	Ref.	Ref.	Ref.	Ref.	Ref.
Defense year dummies	Yes	Yes	Yes	Yes	Yes
University dummies	Yes	Yes	Yes	Yes	Yes
Constant	5.62***	0.56**	9.95***	-	-
	(0.45)	(0.26)	(2.30)		
Observations	24,088	24,088	24,088	24,088	24,088
R-squared / Pseudo R-squared	0.084	0.084	0.170	0.025	0.033

Appendix 3: Calculating similarity including the similarity score of the co-supervisor

Table A3.1: Regression explaining Similarity

	(1) OLS Standardized Similarity
Student's characteristics	
At least one pub. during the PhD period	-0.27***
	(0.017)
At least one pub. co-authored with the supervisor during the PhD period	0.81***
	(0.016)
Having a co-supervisor	0.12***
	(0.0092)
Math	0.021
	(0.014)
Med-bio-chem	-0.34***
	(0.012)
Physics	-0.25***
	(0.014)
Engineering	Ref.
Supervisor's characteristics	
Number of publications during the PhD period	0.0046***
	(0.00028)
Average yearly citations received per article	0.013***
	(0.0017)
Number of distinct co-authors during the PhD period	0.00034***
	(0.000079)
Supervisor-Student pair's characteristics	
Female supervisor - Female student	0.12***
	(0.015)
Male supervisor - Female student	0.042***
	(0.010)
Female supervisor - Male student	0.032**
	(0.014)
Male supervisor - Male student	Ref.
Defense year dummies	Yes
University dummies	Yes
Constant	-0.90***
	(0.090)
Observations	46,774
R-squared	0.159

Table A3.2: Regression estimating the impact of similarity on starting academic career and patenting

	(1)	(2)
	Logit Academic career	Logit At least one patent
Similarity	-0.020***	-0.0056***
	(0.0023)	(0.0014)
Student's characteristics		
At least one pub. during the PhD period	0.38***	0.045***
	(0.0074)	(0.0052)
At least one pub. co-authored with the supervisor during the PhD period	0.088***	-0.0072
	(0.0073)	(0.0047)
Having a co-supervisor	0.026***	-0.0041
	(0.0044)	(0.0028)
Math	0.036***	-0.058***
	(0.0067)	(0.0047)
Med-bio-chem	0.12***	-0.012***
	(0.0058)	(0.0035)
Physics	0.096***	-0.025***
	(0.0065)	(0.0040)
Engineering	Ref.	Ref.
Supervisor's characteristics		
Number of publications during the PhD period	-0.00058***	0.0010***
	(0.00013)	(0.000071)
Average yearly citations received per article	0.0086***	0.0011**
	(0.00084)	(0.00050)
Number of distinct co-authors during the PhD period	0.00019***	-0.00017***
	(0.000039)	(0.000026)
Supervisor-Student pair's characteristics		
Female supervisor - Female student	-0.019***	-0.055***
	(0.0072)	(0.0054)
Male supervisor - Female student	-0.029***	-0.045***
	(0.0049)	(0.0034)
Female supervisor - Male student	0.013**	0.0046
	(0.0066)	(0.0038)
Male supervisor - Male student	Ref.	Ref.
Defense year dummies	Yes	Yes
University dummies	Yes	Yes
Constant	-	-
Observations	46,774	46,774
Pseudo R-squared	0.176	0.039

Table A3.3: Regression estimating the impact of similarity on academic career outcomes

	(1)	(2)	(3)	(4)	(5)
	Conditional on starting an academic career				
	OLS	OLS	OLS	Logit	Logit
	N. pubs	Avg cit per pub	N. distinct co-authors	Foreign affiliation	US affiliation
Similarity	0.11*	0.24***	0.57*	0.018***	0.014***
	(0.060)	(0.035)	(0.30)	(0.0036)	(0.0024)
Student's characteristics					
At least one pub. during the PhD period	2.58***	0.28*	10.3***	0.061***	0.054***
	(0.25)	(0.15)	(1.25)	(0.015)	(0.012)
At least one pub. co-authored with the supervisor during the PhD period	0.068	-0.085	-1.10	0.016	-0.0013
	(0.20)	(0.11)	(0.98)	(0.012)	(0.0079)
Having a co-supervisor	0.31***	-0.19***	-0.086	0.045***	-0.016***
	(0.11)	(0.066)	(0.56)	(0.0068)	(0.0045)
Math	0.13	-0.29***	-2.67***	0.075***	0.021**
	(0.19)	(0.11)	(0.95)	(0.012)	(0.0084)
Med-bio-chem	-1.15***	1.37***	4.36***	0.065***	0.049***
	(0.16)	(0.090)	(0.78)	(0.0095)	(0.0067)
Physics	1.17***	0.64***	10.4***	0.14***	0.060***
	(0.18)	(0.10)	(0.88)	(0.011)	(0.0072)
Engineering	Ref.	Ref.	Ref.	Ref.	Ref.
Supervisor's characteristics					
Number of publications during the PhD period	0.044***	-0.0023	-0.27***	0.00092***	0.00013
	(0.0035)	(0.0020)	(0.017)	(0.00021)	(0.00012)
Average yearly citations received per article	0.0078	0.32***	0.37***	0.011***	0.0057***
	(0.021)	(0.012)	(0.10)	(0.0013)	(0.00068)
Number of distinct co-authors during the PhD period	0.011***	0.0020***	0.24***	0.000010	0.000097***
	(0.00097)	(0.00055)	(0.0048)	(0.000059)	(0.000030)
Supervisor-Student pair's characteristics					
Female supervisor - Female student	-1.91***	-0.19*	-5.35***	-0.095***	-0.036***
	(0.18)	(0.11)	(0.91)	(0.011)	(0.0073)
Male supervisor - Female student	-2.23***	-0.10	-6.79***	-0.076***	-0.019***
	(0.13)	(0.074)	(0.64)	(0.0077)	(0.0049)
Female supervisor - Male student	-0.087	0.17*	1.71**	-0.0011	0.0037
	(0.17)	(0.095)	(0.82)	(0.0099)	(0.0061)
Male supervisor - Male student	Ref.	Ref.	Ref.	Ref.	Ref.
Defense year dummies	Yes	Yes	Yes	Yes	Yes
University dummies	Yes	Yes	Yes	Yes	Yes
Constant	2.55**	0.22	-1.60	-	-
	(1.29)	(0.74)	(6.36)		
Observations	24,088	24,088	24,088	24,088	24,088
R-squared / Pseudo R-squared	0.084	0.084	0.170	0.028	0.035

Appendix 4: Econometric model including similarity squared term

Table A4.1: Regression estimating the impact of similarity on starting academic career and patenting.

	(1) Logit Academic career	(2) Logit At least one patent
Similarity	-0.15***	-0.11***
	(0.014)	(0.023)
Similarity 2	-0.038***	-0.030**
	(0.0072)	(0.012)
Student's characteristics		
At least one pub. during the PhD period	1.94***	0.59***
	(0.042)	(0.069)
At least one pub. co-authored with the supervisor during the PhD period	0.48***	-0.083
	(0.039)	(0.064)
Having a co-supervisor	0.12***	-0.066*
	(0.023)	(0.037)
Math	0.20***	-0.76***
	(0.035)	(0.063)
Med-bio-chem	0.64***	-0.16***
	(0.030)	(0.047)
Physics	0.49***	-0.34***
	(0.034)	(0.053)
Engineering	Ref.	Ref.
Supervisor's characteristics		
Number of publications during the PhD period	-0.0030***	0.014***
	(0.00069)	(0.00095)
Average yearly citations received per article	0.045***	0.015**
	(0.0043)	(0.0067)
Number of distinct co-authors during the PhD period	0.00099***	-0.0023***
	(0.00020)	(0.00035)
Supervisor-Student pair's characteristics		
Female supervisor - Female student	-0.098***	-0.73***
	(0.037)	(0.072)
Male supervisor - Female student	-0.15***	-0.61***
	(0.026)	(0.045)
Female supervisor - Male student	0.069**	0.062
	(0.034)	(0.050)
Male supervisor - Male student	Ref.	Ref.
Defense year dummies	Yes	Yes
University dummies	Yes	Yes
Constant	-1.82***	-3.13***
	(0.24)	(0.41)
Observations	46,774	46,774
Pseudo R-squared	0.177	0.039

Appendix 5: Moderating effect of the supervisor-student gender pairing

Table A5.1: Regression estimating the impact of similarity on starting academic career and patenting

	(1) Logit Academic career	(2) Logit At least one patent
Female supervisor - Female student * Similarity	-0.037***	-0.0071
	(0.0064)	(0.0049)
Male supervisor - Female student * Similarity	-0.019***	-0.0045
	(0.0039)	(0.0028)
Female supervisor - Male student * Similarity	-0.022***	-0.0058*
	(0.0061)	(0.0034)
Male supervisor - Male student * Similarity	-0.019***	-0.0059***
	(0.0032)	(0.0018)
Student's characteristics		
At least one pub. during the PhD period	0.38***	0.044***
	(0.0075)	(0.0052)
At least one pub. co-authored with the supervisor during the PhD period	0.093***	-0.0060
	(0.0074)	(0.0048)
Having a co-supervisor	0.023***	-0.0049*
	(0.0044)	(0.0028)
Math	0.037***	-0.058***
	(0.0067)	(0.0047)
Med-bio-chem	0.12***	-0.012***
	(0.0058)	(0.0035)
Physics	0.096***	-0.025***
	(0.0065)	(0.0040)
Engineering	Ref.	Ref.
Supervisor's characteristics		
Number of publications during the PhD period	-0.00057***	0.0010***
	(0.00013)	(0.000071)
Average yearly citations received per article	0.0086***	0.0011**
	(0.00084)	(0.00050)
Number of distinct co-authors during the PhD period	0.00019***	-0.00017***
	(0.000039)	(0.000026)
Supervisor-Student pair's characteristics		
Female supervisor - Female student	-0.018**	-0.055***
	(0.0072)	(0.0054)
Male supervisor - Female student	-0.029***	-0.045***
	(0.0049)	(0.0034)
Female supervisor - Male student	0.013**	0.0046
	(0.0067)	(0.0038)
Male supervisor - Male student	Ref.	Ref.
Defense year dummies	Yes	Yes
University dummies	Yes	Yes
Constant	-	-
Observations	46,774	46,774
Pseudo R-squared	0.177	0.039

Table A5.2: Regression estimating the impact of similarity on academic career outcomes

	(1)	(2)	(3)	(4)	(5)
	Conditional on starting an academic career				
	OLS	OLS	OLS	Logit	Logit
		Avg cit	N. distinct	Foreign	US
	N. pubs	per pub	co-authors	affiliation	affiliation
Female supervisor - Female student * Similarity	-0.067	0.40***	0.24	0.024**	0.015**
	(0.17)	(0.095)	(0.82)	(0.010)	(0.0073)
Male supervisor - Female student * Similarity	0.010	0.26***	0.20	0.017***	0.012***
	(0.10)	(0.060)	(0.52)	(0.0064)	(0.0043)
Female supervisor - Male student * Similarity	0.11	0.22**	1.83**	0.025***	0.018***
	(0.16)	(0.089)	(0.77)	(0.0094)	(0.0061)
Male supervisor - Male student * Similarity	0.18**	0.18***	0.43	0.012**	0.013***
	(0.085)	(0.049)	(0.42)	(0.0051)	(0.0034)
Student's characteristics					
At least one pub. during the PhD period	2.61***	0.33**	10.4***	0.064***	0.056***
	(0.25)	(0.15)	(1.25)	(0.015)	(0.012)
At least one pub. co-authored with the supervisor during the PhD period	0.057	-0.13	-1.16	0.013	-0.0040
	(0.20)	(0.12)	(1.00)	(0.012)	(0.0080)
Having a co-supervisor	0.32***	-0.15**	-0.013	0.047***	-0.015***
	(0.11)	(0.066)	(0.56)	(0.0068)	(0.0045)
Math	0.14	-0.29***	-2.70***	0.075***	0.021**
	(0.19)	(0.11)	(0.95)	(0.012)	(0.0085)
Med-bio-chem	-1.15***	1.35***	4.33***	0.064***	0.049***
	(0.16)	(0.090)	(0.78)	(0.0095)	(0.0067)
Physics	1.17***	0.63***	10.4***	0.14***	0.060***
	(0.18)	(0.10)	(0.88)	(0.011)	(0.0072)
Engineering	Ref.	Ref.	Ref.	Ref.	Ref.
Supervisor's characteristics					
Number of publications during the PhD period	0.044***	-0.0022	-0.27***	0.00092***	0.00012
	(0.0035)	(0.0020)	(0.017)	(0.00021)	(0.00012)
Average yearly citations received per article	0.0086	0.32***	0.37***	0.011***	0.0057***
	(0.021)	(0.012)	(0.10)	(0.0013)	(0.00068)
Number of distinct co-authors during the PhD period	0.011***	0.0020***	0.24***	0.000010	0.000098***
	(0.00097)	(0.00055)	(0.0048)	(0.000059)	(0.000030)
Supervisor-Student pair's characteristics					
Female supervisor - Female student	-1.90***	-0.20*	-5.33***	-0.096***	-0.037***
	(0.18)	(0.11)	(0.91)	(0.011)	(0.0075)
Male supervisor - Female student	-2.22***	-0.11	-6.78***	-0.076***	-0.018***
	(0.13)	(0.074)	(0.64)	(0.0077)	(0.0050)
Female supervisor - Male student	-0.086	0.17*	1.61**	-0.0019	0.0028
	(0.17)	(0.096)	(0.82)	(0.0100)	(0.0062)
Male supervisor - Male student	Ref.	Ref.	Ref.	Ref.	Ref.
Defense year dummies	Yes	Yes	Yes	Yes	Yes
University dummies	Yes	Yes	Yes	Yes	Yes
Constant	2.51*	0.20	-1.65	-	-
	(1.29)	(0.74)	(6.36)		
Observations	24,088	24,088	24,088	24,088	24,088
R-squared / Pseudo R-squared	0.084	0.084	0.170	0.028	0.035

Appendix 6: Moderating effect of the discipline dummy variables

Table A6.1: Regression estimating the impact of similarity on starting an academic career and patenting

	(1)	(2)
	Logit Academic career	Logit At least one patent
Math * Similarity	-0.051***	0.0074*
	(0.0053)	(0.0045)
Med-bio-chem * Similarity	-0.019***	-0.0088***
	(0.0031)	(0.0019)
Physics * Similarity	-0.0018	-0.0017
	(0.0052)	(0.0034)
Engineering * Similarity	-0.021***	-0.0063**
	(0.0047)	(0.0026)
Student's characteristics		
At least one pub. during the PhD period	0.37***	0.045***
	(0.0075)	(0.0052)
At least one pub. co-authored with the supervisor during the PhD period	0.097***	-0.0076
	(0.0074)	(0.0048)
Having a co-supervisor	0.023***	-0.0051*
	(0.0044)	(0.0028)
Math	0.041***	-0.060***
	(0.0068)	(0.0049)
Med-bio-chem	0.12***	-0.013***
	(0.0058)	(0.0035)
Physics	0.097***	-0.025***
	(0.0065)	(0.0040)
Engineering	Ref.	Ref.
Supervisor's characteristics		
Number of publications during the PhD period	-0.00052***	0.00099***
	(0.00013)	(0.000072)
Average yearly citations received per article	0.0086***	0.0012**
	(0.00084)	(0.00050)
Number of distinct co-authors during the PhD period	0.00018***	-0.00017***
	(0.000039)	(0.000026)
Supervisor-Student pair's characteristics		
Female supervisor - Female student	-0.019***	-0.054***
	(0.0072)	(0.0054)
Male supervisor - Female student	-0.028***	-0.045***
	(0.0049)	(0.0034)
Female supervisor - Male student	0.014**	0.0047
	(0.0066)	(0.0038)
Male supervisor - Male student	Ref.	Ref.
Defense year dummies	Yes	Yes
University dummies	Yes	Yes
Constant	-	-
Observations	46,774	46,774
Pseudo R-squared	0.177	0.039

Table A6.2: Regression estimating the impact of similarity on academic career outcomes

	(1)	(2)	(3)	(4)	(5)
	Conditional on starting an academic career				
	OLS	OLS	OLS	Logit	Logit
		Avg cit	N. distinct	Foreign	US
	N. pubs	per pub	co-authors	affiliation	affiliation
Math * Similarity	0.36**	0.26***	1.63**	-0.0065	0.013*
	(0.16)	(0.090)	(0.78)	(0.0094)	(0.0073)
Med-bio-chem * Similarity	-0.0044	0.33***	0.55	0.022***	0.015***
	(0.077)	(0.044)	(0.38)	(0.0046)	(0.0030)
Physics * Similarity	0.19	0.068	0.51	0.016*	0.0097*
	(0.14)	(0.082)	(0.71)	(0.0086)	(0.0053)
Engineering * Similarity	0.17	-0.0099	-0.46	0.013	0.015**
	(0.14)	(0.083)	(0.71)	(0.0089)	(0.0072)
Student's characteristics					
At least one pub. during the PhD period	2.62***	0.33**	10.5***	0.063***	0.056***
	(0.25)	(0.15)	(1.25)	(0.015)	(0.012)
At least one pub. co-authored with the supervisor during the PhD period	0.0071	-0.10	-1.29	0.017	-0.0036
	(0.20)	(0.12)	(1.01)	(0.012)	(0.0081)
Having a co-supervisor	0.31***	-0.15**	-0.014	0.048***	-0.015***
	(0.11)	(0.066)	(0.56)	(0.0068)	(0.0045)
Math	0.11	-0.34***	-3.03***	0.078***	0.022**
	(0.20)	(0.11)	(0.96)	(0.012)	(0.0089)
Med-bio-chem	-1.16***	1.33***	4.14***	0.065***	0.049***
	(0.16)	(0.091)	(0.78)	(0.0096)	(0.0070)
Physics	1.17***	0.60***	10.2***	0.14***	0.061***
	(0.18)	(0.10)	(0.89)	(0.011)	(0.0076)
Engineering	Ref.	Ref.	Ref.	Ref.	Ref.
Supervisor's characteristics					
Number of publications during the PhD period	0.043***	-0.0016	-0.27***	0.00097***	0.00012
	(0.0036)	(0.0020)	(0.018)	(0.00021)	(0.00012)
Average yearly citations received per article	0.010	0.32***	0.37***	0.011***	0.0057***
	(0.021)	(0.012)	(0.10)	(0.0013)	(0.00068)
Number of distinct co-authors during the PhD period	0.011***	0.0018***	0.23***	-3.9e-07	0.000097***
	(0.00097)	(0.00056)	(0.0048)	(0.000059)	(0.000030)
Supervisor-Student pair's characteristics					
Female supervisor - Female student	-1.90***	-0.20*	-5.38***	-0.095***	-0.036***
	(0.18)	(0.11)	(0.91)	(0.011)	(0.0073)
Male supervisor - Female student	-2.23***	-0.11	-6.80***	-0.076***	-0.019***
	(0.13)	(0.074)	(0.64)	(0.0077)	(0.0049)
Female supervisor - Male student	-0.088	0.16*	1.68**	-0.00089	0.0037
	(0.17)	(0.095)	(0.82)	(0.0099)	(0.0061)
Male supervisor - Male student	Ref.	Ref.	Ref.	Ref.	Ref.
Defense year dummies	Yes	Yes	Yes	Yes	Yes
University dummies	Yes	Yes	Yes	Yes	Yes
Constant	2.57**	0.14	-1.78	-	-
	(1.29)	(0.74)	(6.36)		
Observations	24,088	24,088	24,088	24,088	24,088
R-squared / Pseudo R-squared	0.084	0.085	0.170	0.028	0.035