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 de 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 young researchers' career outcomes. We use a unique and detailed dataset on the French population of STEM PhD students who graduated between 1995 - 2013. To measure research independence, 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 experience and supervisor explain the level of research similarity, we estimate the impact of similarity on the likelihood of pursuing a research career and its effect on a set of productivity measures. We find that research similarity has a negative effect both on starting a career in academia and on patenting. Increasing student's similarity by one standard deviation is associated with a 2% decrease in the probability of obtaining an academic position and a 0.6% decrease in the probability of patenting. However, conditional on starting an academic career, research similarity is associated with a higher student's productivity measured by publications, citations received, network size, and probability of having a foreign and US-based affiliation.

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

JEL Codes: D22, O30, O33, O38

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)

"Postdoc Mobility fellowships are aimed at researchers who have done a doctorate and who wish to pursue an academic career in Switzerland. A research stay abroad enables such researchers to acquire in-depth knowledge, increase their scientific independence and enhance their research profile" (Swiss National Science Foundation Post-doc mobility grants).

"R01 R03 R21 R34 Investigator: Are the PD/PIs, collaborators, and other researchers well suited for to the project? If Early Stage Investigators or those in early stages of independent career, do they have appropriate experience and training?" (National Institutes of Health Research grants)

"*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"* (EU Framework Program for Research and Innovation).

Common characteristic among the above grant programs is the requirement of presenting a certain level of research independence by funding applicants. Grant awarding funding agencies, such as the SLOAN Foundation, the NIH and the Swiss National Science Foundation 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 not only a selection criterion but also one of the targets features and outcomes aimed by funders such as the Marie Skłodowska-Curie Actions Individual Fellowship supported by EU Framework Program for Research and Innovation. In this respect, the EU insists that candidates should spell out how their training and actual experience relate to their leadership and independent thinking skills.

Although being a topic of concern for academics and policy makers, research independence of PhD students and early career researchers presents many un-explored issues and questions. This paper aims to tackle some of them by using a unique large-scale dataset on the entire population of STEM students in France that graduated between 1995 and 2013.

In line with the above considerations, we explore how student experience and advisors' characteristics during the PhD period relate to research independence. Secondly, we investigate the impact of research independence on the likelihood of pursuing a scientific career after graduating. Finally, conditional on starting an academic career, we estimate how independence affects later career outcomes.

Baseline estimation shows that pairing between female advisor and student is associated with the highest positive effect on similarity. The impact of research similarity with the PhD advisor is negative for both starting a career in academia and patenting. Our results yield that increasing student's similarity by one standard deviation is associated with a 2% decrease in the probability of getting an academic position and a 0.6% decrease in the probability of

patenting at least once. Conditional on starting an academic career, research similarity is associated with a higher student's productivity measured by publications, average number of yearly citations received, network size, and probability of having a foreign and US-based affiliation.

Our work is related to a growing body of contributions in economics and sociology of science on early career researchers and scientific productivity. While some studies look at scientific research production at the laboratory level (Carayol and Matt, 2004), most contributions concentrate on the effects of individual determinants of academic productivity such as age (Zuckerman and Merton, 1972, Diamond, 1986, Stephan and Levin, 1997), cohort (Weiss, 1981), training (Garcia-Romero and Modrego, 2001) and gender (Stephan, 1998). Carayol and Matt (2006) combine both individual and collective determinants to explain individual research productivity. However, for the most part, these studies treated the productivity of experienced scientists.

A complementary body of studies has slowly been consolidating around the determinants of young scholars' productivity. These works either tackle PhD productivity through the impact of supervisor' gender (Tenenbaum et al. 2001, Pezzoni et al. 2016, Gaule and Piacentini 2018), the effects of PhD funding on research performance (Horta et al. 2018), the influence of faculty quality on PhD student outcomes (Waldinger, 2010). In a recent contribution, Corsini et al. (2021) provide a comprehensive picture of doctorate student productivity determinants by considering the entire set of characteristics of the social environment in which the student is trained.

Academic independence has been explored mainly in the field of sociology of science. Laudel and Glaser (2008) look at the transitions from the apprentice into an independent scholar, in which young researchers are responsible for setting their own research agendas and running their own research teams. In a further study, Laudel and Glaser (2018) study the actual process through which early career researchers develop their individual research programs. Hemmings and Kay (2010) investigate the relationship between self-efficacy beliefs about research and the publishing outputs of freshly appointed lecturers and show the more self-efficacious in terms of research the more likely an academic would produce refereed publications.

A related topic of inquiry is the study of the effect of post-doc or particular fellowship on young scholars' later careers. Su (2013) evaluates how postdoctoral training affects scientists' academic employment. Using NIH administrative records on applicants and awardees of NRSA F32 fellowships between 1996 and 2008, Heggeness et al. (2018) provide evidence that an F32 fellowship substantially increases the likelihood of obtaining subsequent NIH research funding, including the probability of an NIH R01 grant, which is a major indicator of transition to an independent biomedical research career. In a subsequent study, employing data from a prestigious training program at the National Institute of Health (NIH), Ginther and Heggeness (2020) show that the NIH peer-review system works as a predictor of high-quality science and future research independence. Exploring a sample of PhD students in Portugal, Horta and Santos (2016) reveal that students who publish during their PhD are also more able to publish single-authored papers and publish with foreign peers, suggesting higher autonomy and more international collaborations later in their career. Drawing on

survey and bibliometric data, Yoshioka-Kobayashi and Shibayama (2020) explore the effect of different types of research training on academic independence in life sciences at Japanese universities and show that higher autonomy leads to greater organizational independence and encouragement to deviate from conventional research topics during early-career training is associated to greater cognitive independence in students later career.

Although the existent literature has provided evidence concerning various elements of early career researcher's productivity determinants, altogether the above studies display three important shortcomings. First, previous studies have remained silent on an essential factor in young scholars' career, 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. Second, extant works focus on specific disciplines and relatively small samples of students affiliated to limited number of prestigious universities or associated to a particular research funding or award. Therefore, none of these contributions have been able to conduct empirical work on the entire population of PhD students at a national scale, across all disciplines without discriminating between top-tier and low-rank universities. The lack of available data on early-career researchers has been pointed as main obstacle to provide accurate counts and statistics on the research workforce and makes it difficult to formulate the kind of targeted and rigorous analysis needed for making informed policy decisions. We draw from unique data on the entire student population in France across all disciplines. More precisely, we focus on a sample of 46,774 STEM students who defended their thesis between 2005 and 2013 in France. Third, the phenomenon of becoming independent involves more than the ability to autonomously develop one's own individual research questions and it is closely linked to the content of student's research. Research content is usually overlooked by career studies that focus on publications counts. Our work goes beyond as we measure independence by looking at the actual research content. We employ advanced neural network text analysis techniques to evaluate the similarity between student's thesis abstract and director publications during the PhD period.

At a more systemic level, the independence of early career scientists is an essential feature of the vitality of research enterprise and its workforce dynamics. Therefore, we suggest that our study can have significant policy impact as it might deliver insights about the mechanisms enhancing the quality and effectiveness of doctoral training, the ability of young scientists to receive independent research funding, and the ability to establish a stable research program. Bridges to independence report by the National Research Council. (2005) emphasize several distressing trends in the US biomedical research workforce, such as a decline in the share of key research grants going to younger scientists, as well as a steady rise in the age at which investigators receive their first funding.

Despite long history of concern on the issues of independence of young scholars and the way they manage to establish their stable research careers, progress has been slow. A decade after the NRC report, scholars stress that the trajectory of NIH funding system evolution away from young researchers has only deteriorated (Daniels, 2015). These trends have been accentuated by a shift in the conditions underlying young researchers' careers in the last thirty years. Among the most significant changes are the common role of casual and project-based employment combined with long insecure career spans, increasing reliance on external funding as well as the postponement of tenure (NRC 2005; Stephan and Ma 2005; Akerlind

2005; Lam and de Campos 2015; NAS 2014). Against this context, our study can provide insights helping to grasp the factors at play influencing the above trends as research independence is closely related to a broader range of issues of recruiting, retaining, and supporting young scholars in STEM.

The rest of the paper proceeds as follows. The next section discusses the previous literature and presents the theoretical framework of our analysis. Section 3 describes the data, and section 4 presents the empirical setup. We discuss our preliminary results in Section 5. Section 6 summarizes the results and concludes with some policy implications, limitations, and possible extensions of this work.

2. Research independence: a central tenet of the implicit contract between student and advisor

The aspect of PhD students' autonomy falls within a set of implicit contractual relations between young researchers and their supervisors (Stephan and Levin 1997). These implicit bindings among young and established scholars are at the heart of the dynamics of scientific production across STEM disciplines. Stephan and Levin (1997) identify two mechanisms of influence underlying these unspoken rules.

The first one stems from the nature of delayed rewards in academia. Given the short-term and unattractive nature of the PhD status, the reward for the work during this period comes only after conclusion of the PhD dissertation. During and after the PhD period, there are reciprocal influences within the pair. Entering academia depends on the quality of the student's work, its coherence with the supervisor and work ethics. If the student is good according to these conditions, the senior will support him as he is going to enhance supervisors' reputation in his field and among his colleagues. This in turn will motivate the supervisor to introduce the student to his own research networks. The mechanisms are self-reinforcing, since the value of the supervisor support and inclusion depend on his own reputation built by research and previous students. The complexity in this implicit contract comes from the fact that eventually, the students' prestige will also depend on that of the advisor.

The second mechanism at work comes from the senior reputation effect. If the supervisor fails to respect his promises, he will lose credibility among good PhD students. In this framework, the quality of the students' dissertation depends on the supervisor limited amount of attention and time and the latest optimizes his visibility through co-authored papers in highly ranked peer-reviewed journals and academic positions for his PhDs, by choosing the best possible students.

Against this background, students' ability to conduct independent research with respect to their supervisors plays an important role in two instances. First, before entering the academia, during their dissertation, students need to show certain dedication and hard-working ethics to be perceived as worthwhile candidates by supervisors. Working on a subject that is very far from the one of the supervisors would not be sufficient to secure the positive effects for the supervisor's network such as co-authored publications and other collaborations. Working on subject that is also too similar to the one of the supervisors would also not be optimal as

it might signal a lack of personal research interests and capabilities and might fail to generate desired outcomes for the supervisor's existing research team and network. Based on the concept of implicit contracts between student and advisor, we expect to observe empirically some of the above-described dynamics. Concerning the relations among supervisor reputation and similarity, we expect that the stronger the research record of the supervisor, the higher the student motivation to stay close in order to benefit from her advisor's network. Consequently, we believe that the likelihood of getting an academic position and advisors' research profile will be positively associated.

However, concerning similarity, we expect that being close to one's advisor will play out rather negatively in the early stages of young research career since students are expected to demonstrate a more independent approach in order to gain their own credibility within the academic community and obtain fixed position. However, once entered in academia, the newly appointed fellow can only benefit from vicinity to her PhD advisor and team.
We consider two types of control variables related to student's discipline and gender.

Previous studies have reported productivity and citation patterns that significantly differ between the two large scientific fields of *applied* and *social* sciences (Jung 2012; Shin and Cummings 2010). Beyond this, among disciplines in applied sciences there are important differences in theory and methods of research as well as basic paradigms (Biglan, 1973; Becher and Trowler, 1989). Laudel and Bielick (2018) provide a case study evidence about the ways in which new knowledge is produced and its effect on the intellectual and material bases of independence. They establish two characteristic categories for the emergence of individual research agenda: *early IRP emergence* as the emergence of IRPs during the PhD phase and *extended IRP emergence* as processes that stretched out over a longer time period after the defence. As we control for student's discipline, we anticipate a broad heterogeneity among research fields.

Concerning gender pairing as control, the sustained trend of women's lower productivity in academia suggests that supervisors' gender and gender pairing between supervisor and student could be crucial to take into consideration in the study of independence and young scholars' productivity.

Although recent contributions (Cech et al. 2011, Pezzoni et al. 2016, Gaule and Piacentini, 2018) present a consistent evidence about the role of female advisors in PhD performance, they belong to a broader set of studies that have so far offered conflicting pieces of evidence on the matter. Based on these elements, we expect to observe quite a heterogeneity among the different gender-pairing groups with a special focus on female-female pairs.

In what follows, we present our data and the empirical approach.

3. Data

3.1 Empirical context

We test our research conjectures in a unique empirical setting that provides comprehensive information on the whole population of French Ph.D. students who defended their dissertations in 1990-2017 across all disciplines. French universities are known for their excellence, especially in STEM disciplines such as biology, chemistry, medicine, and mathematics, wherein France has a long list of Nobel laureates. Beyond the most iconic French heavyweight in science are Laplace (mathematics), Picardet (literature and chemistry), Calmette (medicine), Curie (physics), and Pasteur (microbiology), France is the 4th country in the world in the number of Nobel prize awards (69) after USA, UK and Germany. French research universities are well placed according to many international rankings such as Shanghai ranking, 2015 ARWU global ranking or Nature's 2017 Lens score.

Studying the relationship between PhD students' independence and their chance to get in academia and career outcomes is especially relevant in the French context. French PhD contracts are standard in all disciplines regarding their key attributes for instance earnings (scholarships are of the same amount) and duration as students are hired for a fixed term of 3 to 4 years. On top of this, according to stated rules, PhD students cannot be employed in the same lab in which they graduated. Therefore, as Mangematin (2000) explains, in order to complete their PhD and start their academic career, other factors related to the supervisor provide such stimulus. In other words, since the PhD administrative package is a fix variable in France, it is to the supervisor to make the difference. In terms of an implicit contract between the student and advisor, the latest provides incentives for the youngsters to produce research by supporting them through their career path and eventually helping them to get a position. After graduation, the young doctors across all disciplines need to establish a record for national qualification to be qualified to present themselves at any job market competition in France. In the last decade, one of the important criteria to obtain the qualification is to have at least one working paper or publication as sole author. Independence from the supervisor and other senior researchers is a key element to obtain a permanent position.

3.2 Data sources

This section describes in detail the process by which the data used in the econometric analysis were assembled. Our study draws on two data sources: (i) fine-grained data on PhD students from the national-wide French repository of Electronic Doctoral Theses (EDT) and (ii) students' and supervisors' publication records from Elsevier's SCOPUS database.

PhD data

The primary data is from the nation-wide French repository of Electronic Doctoral Theses (EDT). By special permission, we obtained access to the whole universe of STEM thesis records collected by the Agence Bibliographique de l'Enseignement Supérieur (ABES) that is managing the repository since 1985. For each thesis record, we have information on the author, the university of graduation, the defence date, the supervisor's name, the co-supervisor name (if any), and the field of study. As fields, we distinguished between Mathematics, Engineering, Physics, Medicine-biology-chemistry as well as Social and Human Sciences.

Publication data

We retrieve students' and supervisors' publication records from Elsevier's SCOPUS database. The Scopus dataset is world's biggest abstract and citation database of peer-reviewed literature and quality web sources. The Scopus dataset covers publications of all French authors for the period until 2019. For each publication, we have its title, journal, year of publication, citations, and keywords. In addition, the publication data also provides authors' affiliations.

Information on the students' and supervisors' gender stems from a multiple-iteration matching strategy (Gaule and Piacentini, 2018; OECD, 2012). First, we match the given names with the official French gender-name dataset. Then, for the non-matched names, we repeated the matching procedure with the U.S. Census Bureau gender-name dataset and with WIPO gender-name dataset, respectively.

To create our study sample, we merged all of the above datasets using the student's full name as the merging criterion. After some manual cleaning, we obtain a study sample of 46 774 Ph.D. students who graduated between 1995 and 2013.

3.3 Academic career and research independence

To explore the role played by research independence in young researchers' careers, we carry out our analysis in three stages. Firstly, we explore how a set of student and supervisor characteristics explain the degree of research independence. Secondly, we estimate the effect of independence on students' probability of starting an academic career. Finally, we deepen our analysis by looking at the effect of research independence on further scientific productivity for those students who continued their academic careers.

Measure of research independence

To measure student' research independence vis-a-vis their supervisors during the PhD, we introduce a similarity measure among the dissertation and the advisor publications during the period. Using advanced neural network text analysis techniques, we compare the content of students' thesis abstracts with the content of supervisors' publications. To do so, we first transform the text of all the documents into vectors, using the Word2vec algorithm (Mikolov et al., 2013). We then use the vectors to compute a cosine similarity score between the thesis and each supervisor publication. Specifically, we extract from those publications the abstract texts and pair each abstract with the thesis abstract. Overall, to compute the similarity for our final sample, we use the text of 46 774 PhD thesis abstracts and 411 128 advisors' publication abstracts. In our regression exercise, we standardize the similarity values subtracting the sample average similarity and dividing by the sample standard deviation.

To illustrate our independence measure, in the appendix, we present a detailed example of Professor B and two of his students A and Y. Student A graduated in 2007 at the university of Nantes and shows a very low level of Similarity with Professor's B work (0.908). During the Ph.D. period, A does not publish any paper with his supervisor and has a co-supervisor (Professor C) whose research was highly similar to student's A thesis (0.971). This case shows that in practice, Professor B was not expert in the topic of the thesis, and he invited a colleague

of his to co-supervise student's A dissertation. Naturally, student A gravitated towards the co-supervisor Professor C. In contrast, Student Y is another Ph.D. student of Professor B, and her work displays a very high level of Similarity with her supervisor work (0.97). During the Ph.D. period, student Y publishes two papers co-authored with her supervisor. In the latest case, professor B is only advisor of the student Y.

Academic career proxies

To proxy for starting an academic career, we look at the student's publications in the 5 years (t+1 t+6) after the defence t. Having at least one publication showing an academic affiliation, such as a university or a public research organization, indicates that the student has started an academic career. Then, we consider a different kind of research career looking at the student's inventive activity. Specifically, we calculate a dependent variable indicating whether the student was the inventor of at least one patent in the 5 years after the defence. These two measures allow us to estimate the effect of independence on the probability of having two types of research careers.

Scientific productivity measures

We evaluate the impact of independence on five dimensions of later research productivity as measured by the number of peer-reviewed papers published by the student (Publications). Publication quality is proxied by counting the number of yearly citations received on average by the student's papers (Average citations per publication). We proxy the student's research network size as the number of the student's distinct co-authors (Co-authors). Mobility is proxied by two dummy variables indicating whether the student had at least one foreign (Foreign affiliation) or US affiliation (US affiliation). The five productivity variables are calculated for the five years after the thesis defence, i.e., from t+1 to t+6.

Control variables

Finally, we consider a set of student and supervisor characteristics. To control for further student attributes, we include a dummy variable indicating if the student has published at least once during the thesis (Publication Ph.D.); a dummy showing if the student has at least one co-authored work with her supervisor (Student-Supervisor Publication) and a dummy that equals one in presence of co-supervision (Co-supervision). Our analysis includes three further variables related to the advisor activity: advisor's average yearly citations on articles published during the Ph.D. period (Supervisor average citations); advisor's number of co-authors (Supervisor Co-authors) and advisor's number of publications during the Ph.D. (Supervisor Publications). Additional controls include gender interaction terms, discipline variables, Mathematics, Medicine-biology-chemistry and Physics, defence year dummies to control for graduation year cohort effects, university, and discipline dummies to control for the non-observable heterogeneity across universities and thesis research fields. Table 1 below lists all the variables included in our analysis with a short description for each of them.

Table 1: List of variables and description

List of variables	Description
Dependent variables	
Student's academic career	
Academic career	Dummy variable that equals one if the student has at least one publication reporting a university or public research organization affiliation in the period t+1 and t+6
Patent activity	Dummy variable that equals one if the student is the inventor of at least one patent in the period t+1 and t+6
Student's academic productivity	
Publications	Ph.D. student's number of papers published in the period t+1 - t+6
Average citations	Average yearly citations received by the student's papers published between t+1 and t+6
Co-authors	Number of distinct co-authors of the student between t+1 and t+6
Foreign affiliation	Dummy variable that equals one if the student has at least one foreign affiliation between t+1 and t+6
US affiliation	Dummy variable that equals one if the student has at least one US affiliation between t+1 – t+6
Independent variables	
Student independence	
Similarity	Cosine similarity distance between students' thesis abstract and directors' publications during the period t-1 t+4. The similarity values are standardized, subtracting the sample average and dividing by the sample standard deviation.
Other student characteristics	
Publication Ph.D.	Dummy variable that equals one if the student has at least one publication during the thesis period
Student-Supervisor Publication	Dummy variable that equals one if the student has at least one publication co-authored with the supervisor during the thesis period
Co-supervisor	Dummy variable that equals one in the presence of a co-supervisor
Supervisor characteristics	
Supervisor average citations	Average yearly citations received by the supervisor's articles published during the PhD period
Supervisor Co-authors	Supervisor's number of distinct co-authors during the PhD period
Supervisor Publications	Supervisor's number of papers published during the PhD period
Other controls	
F-student M-supervisor	Gender interaction term
M-student F-supervisor	Gender interaction term
F-student F-supervisor	Gender interaction term
Mathematics	Dummy variable that equals one if the Ph.D. dissertation is in Mathematics
Medicine-biology-chemistry	Dummy variable that equals one if the Ph.D. dissertation is in Medicine-biology-chemistry
Physics	Dummy variable that equals one if the Ph.D. dissertation is in Physics
Medicine-biology-chemistry	Dummy variable that equals one if the Ph.D. dissertation is in Medicine, Biology or Chemistry
Defence year dummies	The student PhD defence year t
University dummies	The student university
Discipline dummies	The student discipline

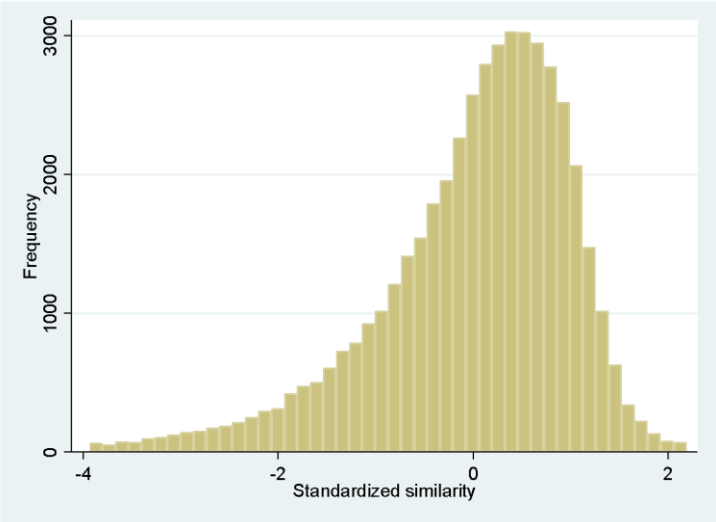
Descriptive statistics

Table 1 reports the descriptive statistics of the dependent variables. With regards to our overall sample of 46 774 students, 51% of them start an academic career and 8% are inventors of at least one patent in the 5 years following their thesis defence. Young researchers that stay in academia have 5 publications on average and receive 2.5 citations per publication. By and large, young scholars network counts 23 distinct cowriters. 38% of students staying in academia have at least once a foreign affiliation as their home institution and 7% sign from an US based university or public research organization. The average level of similarity between students' thesis abstract and supervisors' publications, for students that start academic career in our sample is 0.05 standard deviations larger than the mean for the whole sample.

Table 2: Descriptive statistics

All		Mean	SD	Min	Max
Academic career	46,774	0.51	0.50	0.00	1.00
At least one patent	46,774	0.08	0.28	0.00	1.00
Standardized similarity	46,774	0.00	1.00	-3.94	2.19
Conditional on starting an academic career					
N. pubs	23,725	5.44	6.63	1.00	234.00
Avg cit per pub	23,725	2.53	4.46	0.00	223.57
N. distinct co-authors	23,725	23.33	38.66	0.00	975.00
Foreign affiliation	23,725	0.38	0.49	0.00	1.00
US affiliation	23,725	0.07	0.26	0.00	1.00
Standardized similarity	23,725	0.05	0.96	-3.94	2.18

Figure 1: Similarity distribution



Our main variable of interest similarity presents an asymmetrical left-skewed distribution with the presence of some observations that are very dissimilar, and their value reaches 4 standard deviations from the mean. Heterogeneity among students is illustrated by the presence of an important tail of observations that are very different from their PhD advisors.

By discipline, our sample counts the highest share of students in Medicine-Biology-Chemistry and Engineering followed by Physics and Mathematics. Medicine and physics show the highest share of students staying in academia, 60% and 51% respectively, while around 40% of engineering and mathematics doctors stay in university. As expected, the largest share of engineering students are patent inventors (11%), while only 5% of mathematicians show an

inventive activity. Concerning patents, medicine and physics students are situated around the mean at 8%. The levels of research similarity are quite heterogeneous among the four disciplines. Students in engineering have the most similar dissertations vis-à-vis their advisors work with a similarity index of 0.13, followed by students in mathematics with a similarity index equal to 0.07. Most independent are physics students with -0.08 of average similarity and medicine with -0.07.

Table 3: Dependent variable and similarity average values, by discipline

Discipline	Academic career	At least one patent	Standardized Similarity	Obs.
Engineering	0.40	0.11	0.13	12258
Math	0.45	0.05	0.07	6620
Med-bio-chem	0.60	0.08	-0.07	19741
Physics	0.51	0.08	-0.08	8155

When considering differences among gender pairs, we observe less heterogeneity among the four gender pairs with respect to starting academic career: it is interesting to note that female supervised students have a slightly higher probability of starting an academic career (53% and 55% for female and male students), compared to 50% for students supervised by men. Concerning patenting, male students exhibit a higher probability to patent no matter their advisor gender (10%), while 6% of female students patent with male supervisor and 5% with same-gender supervisor. With regards to similarity, there are differences among male and female advised students. Students with male supervisor, have dissertations that are less similar than female-advised students. Female students supervised by male directors write the most dissimilar theses, while female students supervised by same-gender advisors stick closest to their supervisor research subjects.

Table 4: Dependent variable and similarity average values, by gender pairs

Gender pairs	Academic career	At least one patent	Standardized Similarity	Obs.
F sup-F stud	0.53	0.05	0.05	4,914
F sup-M stud	0.55	0.10	0.04	5,504
M sup-F stud	0.49	0.06	-0.03	13,379
M sup-M stud	0.50	0.10	0.00	22,977

4. **Empirical strategy and main results**

Exploring research similarity: estimating the effect of discipline and gender

To explore the impact of research independence on young researchers' productivity, we proceed in two steps. First, we are interested in the way discipline, gender-pairing, student and advisor characteristics explain the research independence of the student. Therefore, we estimate an Ordinary Least Squares regression, where the dependent variable is Similarity and independent variables are set of gender pairing variables, student, supervisor characteristics, and discipline controls (Equation 1). The level of analysis, as represented by the subscript *i*, is the student.

$$Student's\ Similarity_i = \beta_0 + \beta_1 Gender + \beta_2 Student's\ characteristics_i + \beta_3 Supervisor's\ characteristics_i + \beta_4 Controls_i + \varepsilon_i$$

(Equation 1)

Explaining the impact of similarity for starting a research career

As a second stage, we run a Logistic regression and estimate the impact of student's similarity on the probability to start an academic career (Equation 2) and to file at least one patent application (Equation 3) during the 5 years following the thesis defence. We aim to evaluate the effect of research independence during the thesis on the later probability of staying in academia or patenting. The two different dependent variables indicate the two most significant research paths among which young doctors choose once they accomplish their dissertation: stay in academia or switch for a research job in the industry.

$$Student's\ academic\ career_i = \beta_0 + \beta_1 Similarity + \beta_2 Gender + \beta_3 Student's\ characteristics_i + \beta_4 Supervisor's\ characteristics_i + \beta_5 Controls_i + \varepsilon_i$$

(Equation 2)

$$Student's\ patent\ activity_i = \beta_0 + \beta_1 Similarity + \beta_2 Gender + \beta_3 Student's\ characteristics_i + \beta_4 Supervisor's\ characteristics_i + \beta_5 Controls_i + \varepsilon_i$$

(Equation 3)

Explaining the impact of similarity for further scientific productivity

Finally, conditional on starting academic career, we explore the impact of similarity on later scientific career outcomes as proxied by five different variables that capture various dimensions of the scientific activity such as Number of publications, Average citations per paper, Number of co-authors (OLS estimations), and Foreign and US affiliations (Logit estimations) (Equation 4).

$$Student's\ productivity_i = \beta_0 + \beta_1 Similarity + \beta_2 Gender + \beta_3 Student's\ characteristics_i + \beta_4 Supervisor's\ characteristics_i + \beta_5 Controls_i + \varepsilon_i$$

(Equation 4)

5. Results

Baseline Results

Table 5 below reports the OLS estimates of the model described in Equation 1.

Table 5: Regression explaining Similarity

	(1) OLS Standardized Similarity
Female sup. - Female Stud.	0.12***
Male sup. - Female Stud.	0.039***
Female sup. - Male Stud.	0.030**
Co-supervisor	-0.027***
At least one pub. co-authored sup-stud during the PhD period	1.02***
The stud has at least one pub. during the PhD period	-0.46***
Average yearly citations received by the sup articles	0.013***
Number of distinct co-authors of sup during the PhD period	0.00023***
Number of sup publications during the PhD period	0.0053***
Math	0.027*
Med-bio-chem	-0.33***
Physics	-0.25***
Constant	-0.82***
Observations	46,774
R-squared	0.178
Defence year dummies	Yes
University dummies	Yes

The OLS estimation shows a heterogeneous effect of gender pairing on the similarity level between students and advisors. While all gender pairs have a significant and positive effect on similarity, the pair "female supervisor – female student" is associated with the highest positive effect on similarity. Interestingly, the pairing among female advisors and male students has a smaller effect on similarity. Having a co-supervisor, is associated negatively and significantly to the similarity between student and the main advisor. The sign of the coefficient is expected, since often the initial advisor of the student invites a colleague that might have more competencies in the thesis subject or disposes of more time to co-supervise the student. Unsurprisingly, having at least one co-authored work with the advisor is associated with a significant increase of student's similarity. In contrast, having published other works in peer-reviewed journals (not co-authored) with the supervisor indicates higher independence. All supervisors' characteristics average citations, co-authors and publications during the thesis are associated with higher similarity. With respect to disciplines, our estimations show heterogeneity across fields. Doing a thesis in mathematics is associated to higher similarity while students in Medicine-Biology-Chemistry and Physics are more independent.

Overall, estimation of Equation 1 gives us a first clue of how some of the basic PhD student and advisor characteristics relate to the degree of similarity between the young researcher and her supervisor. We observed important differences among advisors' genders with same-gender women pairs strongly associated with higher similarity. The better the supervisor in terms of scientific visibility, the higher the similarity level. Finally, this basic estimation confirms our theoretical discussion of the existence of different patterns of similarity and, by consequence establishment of independent research programs among scientific fields.

Next, we discuss the results of our Equations 2 and 3, focusing on the impact of similarity for starting an academic career and patenting.

Similarity and starting an academic career

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

	(2)	(3)
	Logit	Logit
	Academic career	At least one patent
Standardized similarity	-0.021***	-0.0057***
Female sup. - Female Stud.	-0.019***	-0.055***
Male sup. - Female Stud.	-0.029***	-0.045***
Female sup. - Male Stud.	0.013**	0.0046
Co-supervisor	0.023***	-0.0049*
At least one pub. co-authored sup-stud during the PhD period	0.094***	-0.0060
The stud has at least one pub. during the PhD period	0.38***	0.044***
Average yearly citations received by the sup articles	0.0086***	0.0011**
Number of distinct co-authors of sup during the PhD period	0.00019***	-0.00017***
Number of sup publications during the PhD period	-0.00057***	0.0010***
Math	0.036***	-0.058***
Med-bio-chem	0.12***	-0.012***
Physics	0.096***	-0.025***
Constant		
Observations	46,774	46,774
Defence year dummies	Yes	Yes
University dummies	Yes	Yes
Discipline dummies	Yes	Yes
R-squared		

The impact of research similarity with the PhD advisor is negative for the probability of starting a career in academia and patenting. We find a significant and negative coefficient for both estimations. Increasing student's similarity by one standard deviation is associated with a 2% decrease in the probability of getting an academic position and a 0.6% decrease in the probability of patenting at least once.

By gender pair, consistent with our previous estimation being in same-gender female pair has a significant and negative impact on starting an academic career and patenting. However, the pair "female supervisor-male student" has a positive and significant effect on staying in academia, while no effect is observed for patent activity. Having a co-supervisor significantly increases the probability of staying in academia. On the contrary, co-supervision has a negative impact on patenting. Co-publishing with the supervisor during the thesis has a positive effect on academic career, increasing the probability of getting a position by 9.4%. Co-authoring a paper with the advisor is likely to give to the student a credit with respect to her supervisor on which is based advisors' support for the student's further career. An additional effect is that by writing a paper with the advisor, the student secures his entry into a certain academic network and community favouring the student's career.

In the same spirit, publishing at least one paper during the PhD period increases the chances of getting an academic career by 38%. The coefficient is positive and significant for the patent activity; however, the effect is weaker 4% increase in the probability of patenting for each standard deviation increase. Along the same line, the number of distinct co-authors during the thesis, apart from the supervisor, brings a significant and positive increase of the probability of getting a position, while the coefficient is negative for patenting.
The number of supervisor publications during the dissertation period has a significant negative effect on student's career, while the visibility of supervisor's work, as proxied by the average yearly citations received by her articles, has a positive and significant effect on student's probability of starting an academic career and patenting.
To summarize, Table 6 shows that working on a thesis similar to the advisor research work has a significant and negative effect on student's probability to embark in an academic career and to patent. In our final analysis, we explore the effect of similarity on further academic output, conditional on starting an academic career.

The impact of similarity for further scientific productivity

Table 7 shows that, conditional on starting an academic career, research similarity is associated with a higher student's productivity according to all the five research outcomes considered in our analysis. In four out of five estimations, the coefficient of the similarity variable is statistically significant. Specifically, our results show a positive but not statistically significant effect of similarity on the number of publications authored by the student after graduation. However, we find a positive and significant impact of similarity on the average number of yearly citations received by the student's papers. Increasing student's similarity by one standard deviation is associated with 0.23 additional citations (0.75% of the sample average). Similarly, similarity exerts a statistically significant (at the 10% level) and positive effect on young scholars' network size. Increasing by one standard deviation student similarity is associated with 0.52 additional co-authors (1.83% of the sample average). Finally, similarity has a significant positive effect on student affiliation quality and international mobility. Increasing student's similarity by one standard deviation is associated with 1.7% increase in probability of having a foreign affiliation and 1.4% increase in probability of having a US-based affiliation.
Concerning the controls, our current results show that being a female student supervised by a female advisor has a significant negative impact on all the academic outcome variables. The picture is opposite for male students advised by women, with most of the coefficient being positive, whereas average citations per paper and number of co-authors indicate a statistically significant effect.

Table 7: Regression estimating the impact on productivity

	(4)	(5)	(6)	(7)	(8)
	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
Standardized similarity	0.093	0.23***	0.52*	0.017***	0.014***
Female sup. - Female Stud.	-1.91***	-0.19*	-5.34***	-0.095***	-0.036***
Male sup. - Female Stud.	-2.23***	-0.10	-6.79***	-0.076***	-0.019***
Female sup. - Male Stud.	-0.087	0.17*	1.71**	-0.0011	0.0037
Co-supervisor	0.32***	-0.16**	-0.0075	0.047***	-0.015***
At least one pub. co-authored sup-stud during the PhD period	0.064	-0.13	-1.17	0.013	-0.0041
The stud has at least one pub. during the PhD period	2.60***	0.33**	10.4***	0.065***	0.057***
Average yearly citations received by the sup articles	0.0082	0.32***	0.37***	0.011***	0.0057***
Number of distinct co-authors of sup during the PhD period	0.011***	0.0020***	0.24***	0.000012	0.000098***
Number of sup publications during the PhD period	0.044***	-0.0024	-0.27***	0.00091***	0.00012
Math	0.13	-0.29***	-2.67***	0.075***	0.021**
Med-bio-chem	-1.16***	1.36***	4.34***	0.065***	0.049***
Physics	1.17***	0.64***	10.4***	0.14***	0.060***
Constant	2.52*	0.19	-1.70		
Observations	24,088	24,088	24,088	24,088	24,088
Defence year dummies	Yes	Yes	Yes	Yes	Yes
University dummies	Yes	Yes	Yes	Yes	Yes
Discipline dummies	Yes	Yes	Yes	Yes	Yes
R-squared	0.084	0.084	0.170		

6. Conclusions

Understanding the determinants of the student's research independence from the supervisor and the effect of the student's research independence on career outcomes is crucial for designing appropriate science and technology policies shaping the research workforce career paths. By investigating the impact of research independence on the likelihood of starting an academic career and patenting, this study contributes to the research on the effects of PhD training on career outcomes.

We conduct 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 (EDT) 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.

Our results show that gender pairing between a female advisor and a female student is associated with a high level of similarity between the student's thesis and the supervisor's work, while having a thesis co-supervision decreases the similarity level. Unsurprisingly, having at least one co-authored work with the advisor is associated with a significant increase of the similarity. In contrast, having published other works in peer-reviewed journals (not co-authored) with the supervisor indicates higher independence. Supervisors' productivity positively relates to the thesis similarity level. We find similarity heterogeneity across research fields.

When we consider the impact of similarity on the student's career after graduation, we find a significant negative effect of research similarity on the probability of starting a career in academia and patenting. Increasing student's similarity by one standard deviation is associated with a 2% decrease in the probability of obtaining an academic position and a 0.6% decrease in the probability of patenting at least once. Finally, we find that newly appointed academics benefit significantly from research similarity as measured by the average number of yearly citations received, network size, and international mobility.

Our results have implications for the policy approach towards the independence of young researchers. On the one hand, independence is important in the early stage of the research career. A PhD student that is more independent from her research supervisor is a student who has a higher probability of obtaining an academic position. On the other hand, conditional on starting an academic career, the productivity of the researchers is fostered by high similarity values with their PhD supervisor research work.

The present study could be extended in several ways. First, we could assess how similarity impact varies across STEM disciplines. Second, we could explore more in-depth the role of gender by including interaction terms among student-supervisor gender pairs and similarity.

References

Åkerlind*, G. S. (2005). Postdoctoral researchers: roles, functions and career prospects. *Higher Education Research & Development*, *24*(1), 21-40.

Becher, T., & Trowler, P. R. (1989). Academic tribes and territories. SRHE.

Biglan, A. (1973). Relationships between subject matter characteristics and the structure and output of university departments. *Journal of applied psychology*, *57*(3), 204.

Carayol, N., & Matt, M. (2004). Does research organization influence academic production? : Laboratory level evidence from a large European university. *Research Policy*, *33*(8), 1081-1102.

Carayol, N., & Matt, M. (2006). Individual and collective determinants of academic scientists' productivity. *Information Economics and Policy*, *18*(1), 55-72.

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.

Daniels, R. J. (2015). A generation at risk: young investigators and the future of the biomedical workforce. *Proceedings of the National Academy of Sciences*, *112*(2), 313-318.

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

Glaser, J. (2001). Macrostructures, careers and knowledge production: a neo institutionalist approach. International Journal of Technology Management, 22(7-8), 698-715.

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.

Ginther, D. K., & Heggeness, M. L. (2020). Administrative discretion in scientific funding: Evidence from a prestigious postdoctoral training program☆. *Research policy*, *49*(4), 103953.

Jung, J. (2012). Faculty Research Productivity in Hong Kong across Academic Discipline. *Higher education studies*, *2*(4), 1-13.

Heggeness, M. L., Ginther, D. K., Larenas, M. I., & Carter-Johnson, F. D. (2018). *The impact of postdoctoral fellowships on a future independent career in federally funded biomedical research* (No. w24508). National Bureau of Economic Research.

Hemmings, B., & Kay, R. (2010). Research self-efficacy, publication output, and early career development. *International Journal of Educational Management.*

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.

Lam, A., & de Campos, A. (2015). 'Content to be sad'or 'runaway apprentice'? The psychological contract and career agency of young scientists in the entrepreneurial university. *Human Relations*, *68*(5), 811-841.

Laudel, G. (2017). How do national career systems promote or hinder the emergence of new research lines?. *Minerva*, *55*(3), 341-369.

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.

Laudel, G., & Gläser, J. (2014). Beyond breakthrough research: Epistemic properties of research and their consequences for research funding. *Research Policy*, *43*(7), 1204-1216.

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.

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*.

Müller, R. (2014). Postdoctoral life scientists and supervision work in the contemporary university: A case study of changes in the cultural norms of science. *Minerva*, *52*(3), 329-349.

National Research Council. (2014). *Convergence: Facilitating transdisciplinary integration of life sciences, physical sciences, engineering, and beyond*. National Academies 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.

Shin, J., & Cummings, W. (2010). Multilevel analysis of academic publishing across disciplines: Research preference, collaboration, and time on research. *Scientometrics*, *85*(2), 581-594.

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., & Ma, J. (2005). The increased frequency and duration of the postdoctorate career stage. *American Economic Review*, *95*(2), 71-75.

Su, X. (2013). International doctoral science and engineering students: Impact on cohorts' career prospects. *Journal of Studies in International Education*, *17*(5), 590-606.

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.

Wanner, R. A., Lewis, L. S., & Gregorio, D. I. (1981). Research productivity in academia: A comparative study of the sciences, social sciences and humanities. *Sociology of Education*, 238-253.

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: Similarity check

Student A is a Ph.D. student of Professor B. He graduated in 2007 at 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 (0.908). During the Ph.D. period, A does not publish any paper with his supervisor and has a co-supervisor (Professor C) whose research was highly similar to A's thesis (0.971)[1].

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 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 very high level of Similarity with his supervisor work (0.97). During the Ph.D. period, Adama publishes two papers co-authored with her supervisor[2].

[1] "Sensitivity approach for modelling the environmental loading of marine structures through a matrix response surface" (sole author publication)

[2] "Buckling and wrinkling of prestressed membranes" and "Modélisation du plissage dans les structures membranaires" https://www.scopus.com/authid/detail.uri?authorId=13610487700

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