EXPLORATION IN RESEARCH TEAMS: BUILDING ON THE SHOULDERS OF PHD STUDENTS

Documents de travail GREDEG
GREDEG Working Papers Series

RAFFAELE MINIACI
MICHELE PEZZONI
SOTARO SHIBAYAMA

GREDEG WP No. 2025-49
https://ideas.repec.org/s/gre/wpaper.html

Exploration in Research Teams:

Building on the Shoulders of PhD Students

Raffaele Miniaci
University of Brescia, Italy
raffaele.miniaci@unibs.it

Michele Pezzoni
Université Côte d'Azur, CNRS, GREDEG, France
Observatoire des Sciences et Techniques, HCERES, Paris, France
michele.pezzoni@univ-cotedazur.fr

Sotaro Shibayama
Institute for Future Initiatives,
The University of Tokyo, Tokyo, Japan
shibayama@ifi.u-tokyo.ac.jp

GREDEG Working Paper No. 2025-49

Exploration is a critical input for creativity and innovation. This paper aims to investigate how the innovator and her team's exploration activities boost the innovator's performance. In our empirical context, the innovator is a French professor at the university, and her team consists of her PhD students. We study 14,978 research teams, led by an equivalent number of supervisors. Supervisors and students can explore by investigating research subjects that the supervisor has not previously investigated. Moreover, the direction of their exploration can be more or less aligned. We measure exploration by assessing the similarity of students' and supervisors' research documents using text analysis. Our regression analyses find that both supervisors' and students' exploration activities play a role in determining the supervisors' performance, as measured by publication quantity, impact, and novelty. We show that an optimal combination of exploration activities and alignment yields considerably higher supervisor performance compared to the average. Our results support the idea that PhD students' exploration activities are of paramount importance to their supervisors' performance, and that supervisors should pay close attention when assigning students' thesis subjects.

Keywords: Research teams; Student exploration; Supervisor exploration; Scientific performance; Text analysis algorithm; Science of science

JEL codes: I20; O30

1. Introduction

Exploration constitutes a fundamental driver of creativity and innovation, through which individuals and organizations are exposed to new ideas and unlock opportunities that would be inaccessible through mere refinement of existing practices (Gupta et al., 2006; March, 1991). Exploration is especially vital in scientific research, where advancing beyond the existing knowledge frontier and demonstrating originality constitute foundational norms (Bourdieu, 1975; Kuhn, 1977; Merton, 1973). Science is characterized by a high degree of autonomy, a dynamic and uncertain environment, and limited bureaucratic constraints, which are considered favorable conditions for exploration (Jansen et al., 2006; Mcgrath, 2001). Exploration involves a trial-and-error process, which often yields little or no immediate payoff, and the likelihood of failure is substantial (Arts and Fleming, 2018; He and Wong, 2004; March, 1991). Indeed, several studies have drawn attention to the potential adverse consequences of scientific exploration. Wang et al. (2017) demonstrated that novel scientific contributions face a higher risk of limited recognition, both in terms of reduced citation uptake and delayed acknowledgment by the scientific community. Hill et al. (2025) also suggest that transitioning into new research topics often disrupts existing knowledge trajectories and networks, resulting in lower scholarly impact.

Although exploration in science has been widely studied, what is critically overlooked in the literature is that science is a team activity (Wuchty et al., 2007). Team leaders can assign exploratory and non-exploratory (i.e., exploitative) roles to different members, potentially generating outcomes unattainable individually (He and Wong, 2004; Tushman and O'reilly, 1996). It is especially important to recognize the distinction between experienced and inexperienced members, as research teams led by experienced members (supervisors) involve inexperienced members (students). As trainees, students are often assigned tasks by their supervisor (Delamont and Atkinson, 2001; Shibayama et al., 2015). At the same time, students are not merely passive learners, but they are expected to develop into independent scientists (Shibayama et al., 2025; Wang and Shibayama, 2022). Therefore, the supervisor might shape the team exploration strategy by asking students to be effective agents of exploration.

Compared with inexperienced members, experienced members have clear advantages in exploration. Their richer knowledge base enables them to more quickly comprehend the knowledge landscape in new areas and detect emergent opportunities (Shibayama et al., 2015). Their stable employment status may afford a greater tolerance for failures and inefficiencies inherent to exploratory work (Wang et al., 2018). Yet inexperienced members may also bring unique strengths. With less commitment to established paradigms, they may be more open to accepting unorthodox ideas that experienced scientists might dismiss prematurely (Stephan and Levin, 1992; Zuckerman, 1988). Also, they typically have more concentrated time available for research, with no or limited burden of teaching, administrative, and funding acquisition responsibilities, and the prolonged effort can be crucial for exploration (Mattsson and Shibayama, 2025). Given these mixed arguments, we examine how the allocation of exploration among supervisors and students in research teams associates with the supervisor's research outcome.

To these ends, we draw on a sample of 14,978 senior scientists across all disciplines in France and examine their team members, particularly PhD students they have supervised. Considering a senior scientist and her supervisees as a team, we measure the team's exploration strategy. Drawing on supervisors' publication trajectories and PhD students' dissertations, we quantify supervisors' exploration as the distance between the content of their current and past publications and students' exploration as the distance between the content of their theses and the supervisors' publications before the students' enrollment in the PhD program. We further measure the alignment between supervisors' and students' exploration as the similarity between the content of the supervisors' current publications and the students' theses. We then investigate how student and supervisor exploration is associated with different dimensions of the supervisor's research outcomes, including quantity, impact, and novelty.

Our results show that supervisors' exploration and students' exploration are associated with the research outcome differently. For example, supervisors' exploration is positively associated with the number of articles published but negatively associated with their impact. On the contrary, students' exploration is negatively associated with the number of articles published but positively associated with their impact. High alignment of the research subjects explored by students and supervisors is negatively associated with the number of articles published, but increases their impact.

Based on the results of our econometric exercises, we analyze the optimal allocation of exploration between supervisors and students and find that maximizing each supervisor's outcome requires a distinct strategy. Specifically, we show that both supervisors and students should explore, but in different directions, to maximize the supervisor's number of publications. On the contrary, strategies that maximize the impact of the research outcome are more diverse according to the impact measure considered.

2. Data, variables, and methodology

This section describes our study sample, the variables considered, and the specifications of our econometric models.

2.1 Data

2.1.1 Sources

In this study, we use two main data sources. The first data source is the French doctoral thesis repository provided by the *Agence bibliographique de l'enseignement supérieur*[1]. This dataset includes information about all PhD theses defended in France over the last three decades, across all disciplines. It allows us to retrieve information on the name and surname of the student who authored the thesis, her supervisor, the university affiliation, the title and abstract, the defence date, and a fine-grained classification of the thesis's discipline according to the *Dewey Decimal Classification*[2] standard. The second data source is the *OpenAlex* bibliometric dataset (Priem et al., 2022). From *OpenAlex*, we retrieve metadata on the supervisor's scientific articles[3], including the article title, abstract, publication year, and journal.

2.1.2 Study sample

To construct our study sample, we consider only supervisors in STEM disciplines. The reason for restricting to STEM is twofold. First, publication of research articles is the main scientific outcome for researchers in STEM disciplines, rather than books, architectural, design, or artistic works, which are more common in non-STEM disciplines. Second, the time elapsed between the result of the research activity and its publication is shorter in STEM disciplines than in non-STEM disciplines. Since our proxies for the supervisor's exploration activities are largely based on publications, the short time lag between obtaining the research results and their publication allows us to promptly trace the supervisor's exploration. In our study, we define supervisors as active in STEM disciplines if they supervise at least one thesis in a STEM field.

After selecting supervisors in STEM disciplines, we attribute to them the publications they authored. To do so, we parse the names of supervisors listed in the French doctoral thesis repository and the names of authors with a French affiliation in *OpenAlex* by removing punctuation, middle names, and titles. Then we match supervisors and authors by their full

[1] Detailed information about *Agence bibliographique de l'enseignement supérieur* is available at https://abes.fr/

[2] Detailed information about the *Dewey Decimal Classification* is available at https://www.oclc.org/en/dewey.html

[3] In our analysis we consider only scientific articles defined as those documents labelled as "articles" in *OpenAlex* and published on a review with an ISSN code.

names. Among the resulting supervisor-author matches[4], we removed those whose authors' publication timing was incompatible with the role of supervisor. Specifically, we removed those supervisor-author matches in which the matched author published her first article less than 5 years before the matched supervisor's first student graduated[5]. Following a conservative strategy, among the remaining supervisor-author matches, we considered only those in which the supervisor is associated with only one author.

To define our study sample, we apply additional filters on the language of the abstracts and titles of the students' theses and the supervisors' publications. Specifically, we select only those theses and publications having their abstracts and titles in English. Having abstracts and titles in English is required to calculate our proxies for supervisors' and students' exploration based on text content.

Among the remaining supervisors, we select those whose careers overlap, partially or completely, the time window considered in our study, i.e., 2009-2022. We define the supervisor's career as starting when she publishes her first article appearing in *OpenAlex* and ending when she publishes her last article. Finally, we filter the remaining supervisor-year pairs to consider only those that allow us to calculate our exploration and alignment measures. Specifically, we select pairs that satisfy two conditions: (1) the supervisor must have at least one PhD student active in the last 5 years, between t-4 and t, and (2) the supervisor i must have at least one publication in year t. We define a PhD student as active from 3 years before the defence year (d-3), i.e., the enrollment year, to the defence year (d), i.e., from d-3 to d.

After applying the matching and filtering procedures described above, we are left with a study sample including 14,978 supervisors observed between 2009 and 2022. Each supervisor is observed on average for 8.07 years, resulting in a total of 120,935 supervisor-year pairs. The panel is unbalanced, with 5.51% of the supervisors observed for the entire period, from 2009 to 2022.

[4] One supervisor can be matched to multiple authors, for instance, due to cases of homonymy. In these cases, one supervisor and multiple authors (who are different individuals) generate multiple matches. Applying filtering criteria after the matching procedure allows to retain the correct matches and drop the false ones.

[5] We set the threshold to 5 years because, in the French system, supervisors need a habilitation to supervise PhD students. Habilitation can be obtained after the recruitment as an associate professor and consists of an evaluation of the supervisor's research performance by a scientific committee. If the professor is deemed scientifically mature enough to have her own research team, the habilitation is awarded. At least 5 years between the first publication and the first PhD student graduation seems to be the minimum time required to secure a tenured position as an associate professor, complete the habilitation process, and supervise the first student until graduation.

2.2 Variables

2.2.1 Supervisor exploration, student exploration, and alignment

We measure exploration activities in year *t* using three variables: *Supervisor exploration*, *Student exploration,* and *Alignment*.

Supervisor exploration in year *t* is calculated in four steps. First, as shown in Figure 1, we select the supervisor's publications in year *t* (*current* publications) and her publications in the five years preceding *t*, from *t-5* to *t-1* (*past* publications). Second, we use a machine learning algorithm to transform documents into vectors based on their content,[6] and we compute the cosine similarity between each "*current* publication-*past* publication" pair of vectors, i.e., $cos(\gamma)$ in Figure 2. Third, we convert the cosine similarity measure into a distance measure by calculating $1 - cos(\gamma)$. Finally, we define *Supervisor exploration* as the average distance value of all the pairs, i.e., $\overline{1 - cos(\gamma)}$.

The variable *Student exploration* in year *t* is calculated as the average content distance between the theses of the students active in the last 5 years, between *t-4* to *t* in Figure 1, and the publications of the supervisor in the 5 years preceding the enrollment of each student (*Before enrollment* publications), from *d-8* to *d-4* in Figure 1. Specifically, we calculate the similarity between each "*before enrollment* publication-*thesis*" pair, i.e., $cos(\beta)$ in Figure 2, then we calculate the average distance value of all the pairs[7], i.e., $\overline{1 - cos(\beta)}$. For instance, for Student 1 in Figure 1, we calculate the content distance of each pair formed by her thesis and the supervisor's articles published between *t-6* (= *d-8*) and *t-2* (= *d-4*). After following the same procedure for Students 2 and 3, we define the *Student exploration* variable as the average distance of all the "*before enrollment* publication–*thesis*" pairs.

Finally, we define the variable *Alignment* between the supervisor's current research and the student's research as the average cosine similarity of the supervisor's "*current* publication-*thesis*" pairs[8], i.e., $\overline{cos(\alpha)}$. The value of *Alignment* ranges from 1, representing perfect alignment of the supervisor and student research activities, to a theoretical value of -1, representing perfect misalignment of the activities. For instance, for Student 1 in Figure 1, we calculate the content similarity of each pair formed by her thesis and the supervisor's articles

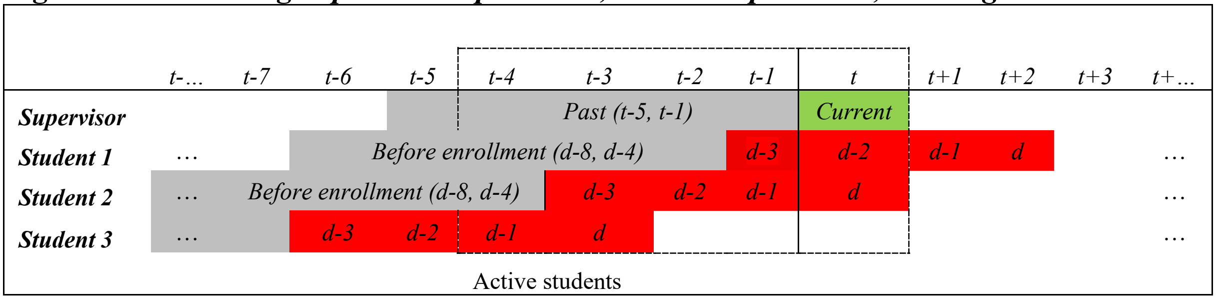
[6] See Appendix A for the details on how we computed similarity between documents.

[7] The average distance value is weighted by the number of years the student appears as active in the time window between *t* and *t-4*.

[8] As in the case of *Student exploration*, the average distance value is weighted by the number of years the student appears as active in the time window between *t* and *t-4*.

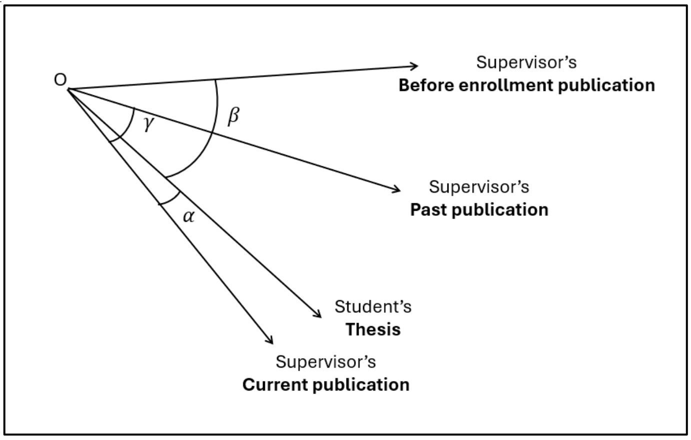
published in *t*. After following the same procedure for Students 2 and 3, we define the *Alignment* variable as the average similarity of all the "*current* publication-*thesis*" pairs.

Figure 1: Calculating *Supervisor exploration*, *Student exploration*, and *Alignment* in a team



NOTE: The figure is an illustrative example of a research team configuration in year *t* . The team consists of a supervisor and 3 students. The letter *d* represents the student's defence year, *d-3* the student's enrollment year, and *t* the focal year. *Current* represents the supervisor's publications at time *t*, and *Past* represents the supervisor's publications in the time window from *t-1* to *t-5*. *Current* and *Past* publications are used to calculate the variable *Supervisor exploration*. *Before enrollment* represents the supervisor's publications in the time window before the student's enrollment, from *d-4* to *d-8*. The dashed line represents the time window, i.e., from *t* to *t-4*, used to identify the active students in the supervisor's team during *t*. Active students' theses and *Before enrollment* supervisor's publications are used to calculate the variable *Student exploration*. Supervisor's *Current* publications and active students' theses are used to calculate the variable *Alignment.* In the figure, the supervisor has three students active for 1 year (Student 1), 3 years (Student 2), and 2 years (Student 3), respectively.

Figure 2: Cosine similarities between a student's thesis, a supervisor's current, past, and before enrollment publications



NOTE: The figure is an illustrative example of four bidimensional vectors representing one document for each of the four types used to calculate *Supervisor exploration*, *Student exploration*, and *Alignment*. Specifically, the figure shows a vector representing a student's *thesis*, a supervisor's *past* publication, a supervisor's publication before the student's enrollment (*before enrollment*), and a supervisor's *current* publication with respect to the origin of the axes (O). The bidimensional space is used to provide a simple graphical representation of the similarity measures; however, the variables used in the regression are calculated in a 100-dimensional space (see Appendix A for details). The similarity between each pair of vectors is calculated as the cosine of the angle between them.

2.2.2 Supervisor's performance measures

The dependent variables of our regression models are proxies for three different dimensions of the supervisor's research performance: quantity, impact, and novelty.

We proxy the quantity of the supervisor's research outcomes by counting the number of publications on scientific journals authored by the supervisor *i* in year *t*. We define the variable *Supervisor publications* accordingly.

We calculate two proxies to measure the impact of the supervisor's research in year *t*. The first proxy is the variable *Citations*, calculated as the average number of citations per article received within five years after publication, i.e., from *t* to *t+4*, by the supervisor's articles published in year *t*. The second proxy for the supervisor's research impact is the variable *Top 10*, which is calculated in two steps. First, for each supervisor publishing in year *t* we calculate the citations received within the following four years. Then, the variable Top 10 takes the value 1 when supervisor *i* ranks in the top 10^{th} percentile of citations received by publications published in year *t* in at least one of her supervision disciplines; otherwise, it is coded as 0.

As for the supervisor's research novelty, we consider two dimensions: global novelty and individual novelty. *Global novelty* is a dummy variable that equals one if the supervisor *i* publishes at least one article in year *t* that contains at least one novel word as defined in Arts et al. (2021). Novel words are words that appear in year *t* for the first time in the history of published articles. Different from global novelty, *Individual novelty* aims to measure the extent to which the supervisor's publications in year *t* are novel compared to the supervisor's previous research history. To do so, we consider journals as knowledge blocks and the supervisor's entry into a novel research field as publishing in a journal in which she has never published before. We calculate the variable *Individual novelty* as equal to one if supervisor *i* publishes in year *t* at least one article in a journal in which she has never published before.

2.2.3 Other control variables

To estimate unbiased coefficients in our regression models, we include control variables. Specifically, we calculate the variable *N. of PhD students* in year *t*, as the number of supervisor *i*'s PhD students who were active in the last 5 years, between *t-4* and *t*. We define a PhD student to be active for 4 years, from the enrollment in the PhD program, 3 years before the defence (*d-3*), to her defence year (*d*) (see Figure 1). In our econometric estimates, we create a set of 4 dummy variables based on the discrete variable *N. of PhD students*: *1 PhD*, *2 PhDs*, *3 PhDs*, and *4 PhDs or more*. We calculate the dummy variable *Co-supervisor*, if at least one of the active students in *t-4* and *t* is supervised by at least another researcher in addition to supervisor *i*. We define *N. supervisor pub.* as the number of the supervisor's articles published between *t-5* and *t*-1. We also control for the discipline by calculating a set of eight *Supervision discipline* dummy variables. We define the supervisor's discipline according to the disciplines of the theses[9] supervised by *i* between *t-5* and *t-1*. These dummy variables are not mutually exclusive since a supervisor i might have supervised students in more than one discipline during the period considered. Finally, we define a set of thirteen *Calendar year* dummy variables, from 2009 to 2022. Table 1 reports a brief description of each variable.

Table 1: Variable description

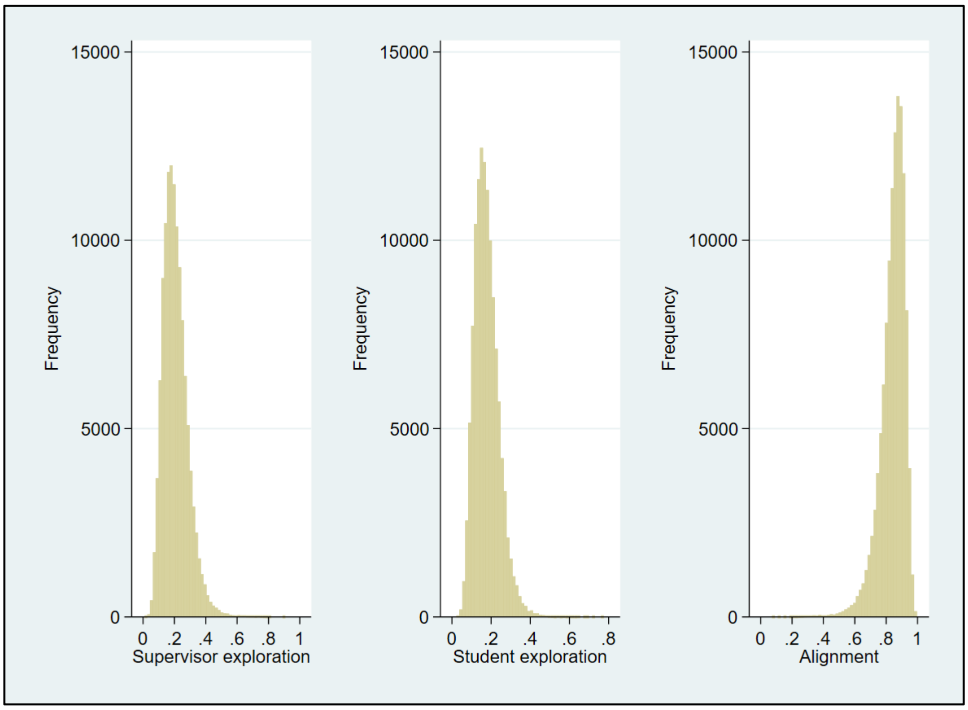
Variable name	Description
Dependent variables	
Quantity	
Supervisor publications	The number of publications authored by the supervisor *i* in year *t*.
Impact	
Citations	The average number of citations per article received by supervisor *i*'s articles published in year *t* within 5 years after publication.
Top 10 percentile	A dummy variable that equals one if the supervisor *i* is in the top 10th percentile in terms of citations received in a given year among the supervisors in the disciplines in which she supervises students.
Novelty	
Global Novelty	A dummy variable that equals one if the supervisor *i* published at least one article in year *t* that contains a novel word as defined in Arts et al (2021).
Individual Novelty	A dummy variable that equals one if the supervisor *i* published at least one article in year *t* in a journal in which she had never published before.
Variables of interest	
Supervisor exploration	Average distance between the content of supervisor's publications in year *t* and the content of her publications in the five years before *t*, i.e., $\overline{1 - cos(\gamma)}$
Student exploration	Average distance between the content of the theses of the students active in year *t* and the content of the supervisor's publications in the five years before each student's enrollment, i.e., $\overline{1 - cos(\beta)}$
Alignment	Average distance between the content of the theses of the students active in year *t* and the content of the supervisor's publications in year *t*, i.e., $\overline{cos(\alpha)}$.
Controls	
N. of PhD students	Is the number of supervisor *i*'s PhD students who are active in the 5 years between *t-4* and *t*. We define a PhD student as active in the period between her enrollment in the PhD program (*d-3*) to the defence year (*d*). In the econometric models, we create a set of 4 dummy variables based on this variable: *1 PhD*, *2 PhDs*, *3 PhDs*, and *4 PhDs or more*.
N. of past supervisor pub.	The number of supervisor *i*'s publications in the five years from *t-5* to *t-1*.
Co-supervisor	A dummy variable that equals one if at least one of the active students in the 5 years between *t-4* and *t* has a co-supervisor.
Supervision discipline	Eight dummy variables defined according to the thesis discipline of the students supervised by the supervisor *i* between *t-5* and *t-1*.
Calendar year	Thirteen dummy variables defined according to the year *t* in which we observe the supervisor *i*.

[9] The discipline is officially attributed to each thesis according to the *Dewey Decimal Classification* system.

2.3 Descriptive statistics

Figure 3 shows the marginal distributions of *Supervisor exploration*, *Student exploration*, and *Alignment*.

Figure 3: Histograms of Supervisor exploration, Student exploration, and Alignment



NOTE: Sample distribution of *Supervisor exploration*, *Student exploration*, and *Alignment*.

Columns 1, 2, and 3 of Table 2 present the correlations between *Supervisor exploration*, *Student exploration,* and *Alignment* of their research. Given the panel structure of the data, it is important to assess the within-group correlation of the three exploration variables, that is, the correlation between the deviations of each variable from its mean calculated at the supervisor level. For instance, the deviation of the variable *Supervisor exploration* form the supervisors' means is defined as $Supervisor\ exploration_{i,t} - \overline{Supervisor\ exploration_i}$. Columns 4, 5, and 6 of Table 2 show that the correlations calculated between the deviations from the means are smaller (in absolute value) than those calculated between the levels of the variables (Columns 1, 2, and 3), particularly for the correlation between *Supervisor exploration* and *Student exploration*, and between *Student exploration* and *Alignment*. This suggests that, in the fixed-effects estimators used in the econometric analysis, our main variables of interest exhibit a limited degree of collinearity.

Table 2: Correlation between Supervisor exploration, student exploration, and alignment

	Levels			Deviations from individual means		
	(1)	(2)	(3)	(4)	(5)	(6)
	Supervisor exploration	Student exploration	Alignment	Supervisor exploration	Student exploration	Alignment
Supervisor exploration	1.00			1.00		
Student exploration	0.48	1.00		-0.016	1.00	
Alignment	-0.79	-0.44	1.00	-0.77	-0.019	1.00

NOTE: The table shows the correlation matrix calculated between *Supervisor exploration*, *Student exploration*, and *Alignment* (Columns 1-3). It also shows the correlations calculated as deviation from the supervisors' means (Columns 2-6).

Table 3 presents the descriptive statistics for our dependent and independent variables. On average, the supervisors in our dataset publish 4.25 papers per year (*Number of publications*), which receive 21.18 citations over the following 5 years (*Citations*). By construction, 10 percent of the supervisors are highly cited every year (*Top 10*). 7% of them publish at least one article that contains some novelty with respect to the entire scientific production (*Global novelty*), and 79% publish in journals in which they have not previously published (*Individual novelty*).

On average, *Supervisor exploration*, i.e., the distance between the content supervisor's publications in year *t* and those in the previous 5 years, is equal to 0.21. Supervisors have an average of 2.2 PhD students active in year *t* who tend to explore less than their supervisors with their theses, i.e., *Student exploration* = 0.18 < *Supervisor exploration* = 0.21. Finally, the average alignment of the students with the content of their supervisor's current publications equals 0.84.

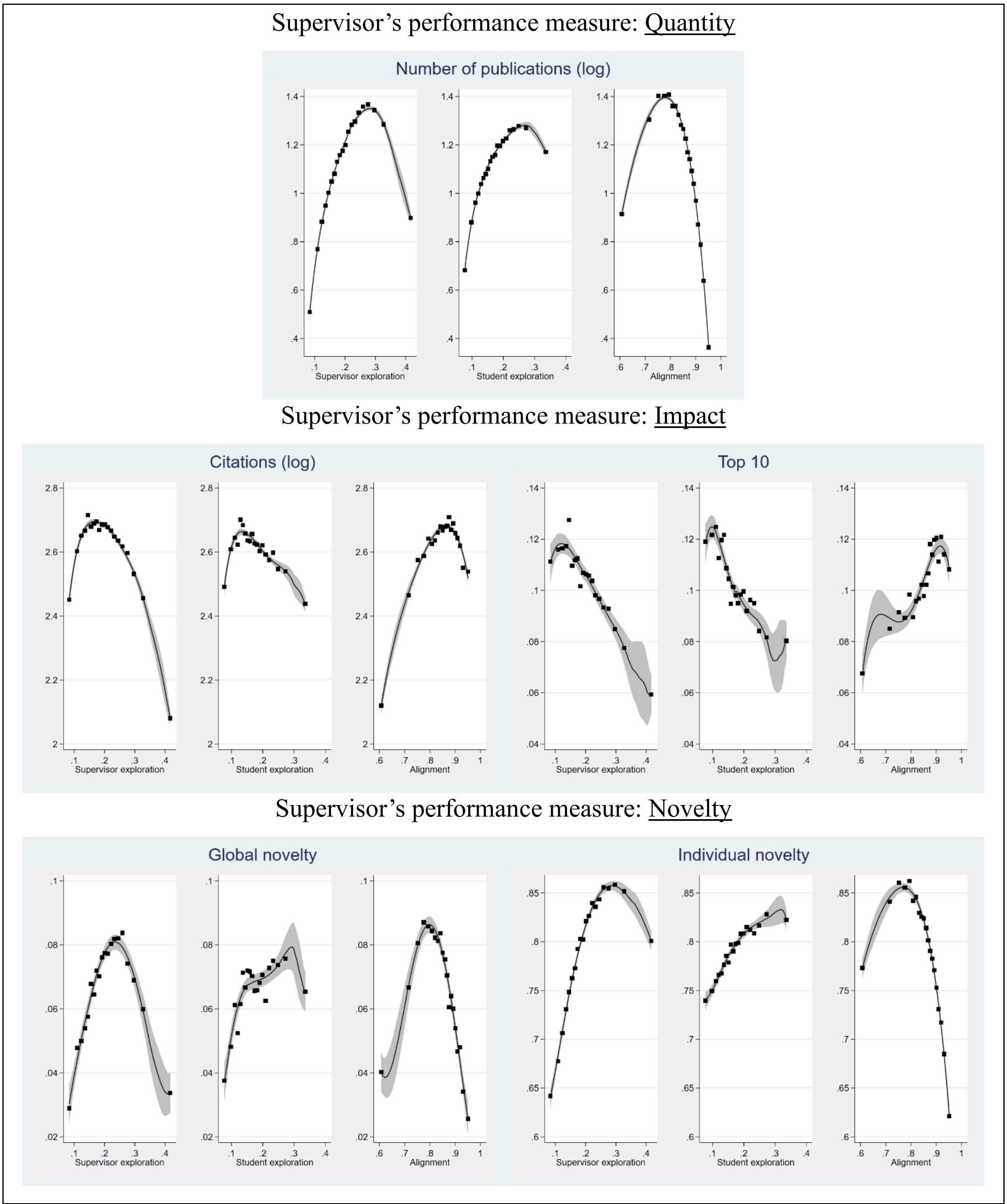
Table 3: Descriptive statistics

Variable	Mean	Std. Dev.	10th Percentile	90th Percentile
(120,935 observations, 14,978 supervisors)				
Dependent variables				
Quantity				
Number of Publications	4.25	4.00	1.00	9.00
Impact				
Citations	21.18	63.24	3.00	40.00
Top 10	0.10	0.30	0.00	1.00
Novelty				
Global novelty	0.07	0.25	0.00	0.00
Individual novelty	0.79	0.41	0.00	1.00
Independent variables				
Variables of interest				
Supervisor exploration	0.21	0.08	0.12	0.31
Student exploration	0.18	0.06	0.10	0.26
Alignment	0.84	0.08	0.74	0.92
Controls				
Co-supervisor	0.38	0.48	0.00	1.00
Number of PhD students	2.23	1.80	1.00	4.00
N. of past supervisor pub.	19.50	17.19	5.00	39.00

NOTE: for space reasons, the table does not report descriptive statistics of the *Supervision discipline* and *Calendar year* dummy variables.

Figure 4 presents preliminary evidence on the relationship between the dependent variables, which measures the supervisor's performance, and the variables that measure exploration and alignment. Specifically, Figure 4 includes 15 panels, one for each combination of the 5 dependent variables and the 3 independent variables. For instance, the top panel shows the relationship between *Supervisor exploration* and the *Number of Publications* expressed in logarithms. To better visualize the relationship between the two variables, we split the *Supervisor exploration* values into 20 consecutive bins, each containing 5% of the observations in our sample. For the observations included in each bin, we calculate the average of *Number of Publications*. With a similar approach, we visualize the relationships between dependent and independent variables in the other 14 panels. We find that supervisor and student exploration show a similar positive relationship with the supervisor's number of publications: the higher the exploration, the higher the number of publications. They also show a similar negative relationship with the probability of being a top-cited supervisor (*Top 10*). This latter result, particularly when considering the supervisor exploration, aligns with previous literature indicating that researchers' exploration tends to lower the impact of their research outcomes (Hill et al., 2025). Another result consistent with the previous literature is the inverted-U relationship observed between supervisor exploration and global novelty (Hill et al., 2025). Overall, a visual inspection of the 15 panels in Figure 4 suggests that many of the relationships between exploration and performance variables are not linear, prompting us to adopt a nonlinear specification for our econometric models by using a second-degree polynomial specification.

Figure 4: Relationships between dependent and independent variables.



NOTE: In each of the 15 panels, the values of the variables on the x-axis are split into 20 consecutive bins, each containing 5% of the sample observations. For each bin, the average value of the variable on the y-axis is calculated and represented by a dot. The gray area represents 10 percent confidence intervals.

2.4 Model

Our empirical strategy relies on five models that take, in turn, one of the supervisor's performance measures as the dependent variable, i.e., *Number of Publications*, *Citations*, *Top 10*, *Global novelty,* and *Individual novelty*. Each supervisor's performance measure is explained by three main explanatory variables: *Supervisor exploration*, *Student exploration,* and *Alignment*. The three variables are fully interacted in the models, which also include supervisor (δ) fixed effects and a full set of interactions between the discipline of the supervised theses (θ) and year fixed effects (τ). Supervisor fixed effects account for time-invariant personal characteristics that may affect the supervisor's performance and be related to our regressors, such as the supervisor's gender, cohort, and university effects, as well as other unobserved personal characteristics, including risk aversion and leadership skills. Equation 1 shows the estimated model, where **X** is a vector including the time-variant control variables described in Section 3.2. The term ε is the idiosyncratic error term.

$$
\begin{gathered}
Supervisor's\ performance_{i,t} = \beta_0 + \\
\beta_1 Supervisor\ exploration_{i,t} + \beta_2 Student\ exploration_{i,t} + \beta_3 Alignment_{i,t} + \\
\beta_4 Supervisor\ exploration_{i,t}^2 + \beta_5 Student\ exploration_{i,t}^2 + \beta_6 Alignment_{i,t}^2 + \\
\beta_7 Supervisor\ exploration_{i,t} \times Student\ exploration_{i,t} + \\
\beta_8 Student\ exploration_{i,t} \times Alignment_{i,t} + \\
\beta_9 Supervisor\ exploration_{i,t} \times Alignment_{i,t} + \\
\boldsymbol{X}_{i,t}\boldsymbol{\beta} + \delta_i + \theta_{i,t}\tau_t + \varepsilon_{i,t}
\end{gathered}
$$

Equation 1

We estimate the parameters using within-group OLS, with standard errors clustered at the supervisor level. As *Supervisor exploration*, *Student exploration,* and *Alignment* enter Equation 1 in a polynomial form, for interpretability reasons, in Table 4, we report the average marginal effects on the performance measure of these variables. The average marginal effects are calculated as the average of the linear combination of the estimated parameters ($\hat{\beta}_{1,} \ldots, \hat{\beta}_9$).

3. Results

The regression estimates reported in Table 4 explain three dimensions of the supervisor's performance: quantity, impact, and novelty.

Regarding the quantity of published articles, we rely on Column 1 of Table 4 to assess the relationship between supervisor exploration and the number of publications by the supervisor in year *t*. Specifically, we consider an increase from the first (Q1) to the third quartile (Q3) of the variable *Supervisor exploration,* and we calculate the percentage variation of the number of publications as $\widehat{log(y)}|_{x\,Q3} - \widehat{log(y)}|_{x\,Q1}$, where y is the *Number of Publications* and x is the *Supervisor exploration*. All the other variables in the regression model are considered at their average values. According to this approach, we find that an increase from Q1 to Q3 of *Supervisor exploration* is associated with a 10.4% increase in publications. This positive association can be explained by the beneficial effects of exploration, as suggested by the literature on research creativity (Shibayama et al., 2015). However, it may also be the result of the production of easy-to-publish conventional research that is expected to characterize the supervisor's activity when exploring new fields. Indeed, the lack of experience in new fields might lead the supervisor to produce outcomes that are original to her research history but are considered conventional for researchers with experience in the field being explored. This latter explanation appears to be supported by the results observed in Columns 2 and 5, which show that the supervisor's exploration is associated with lower research impact, but higher individual novelty. Unlike supervisor exploration, student exploration shows a negative relationship with the supervisor's number of publications. A Q1-Q3 increase in *Student exploration* is associated with a 1.9% publication decrease in the supervisor's publications[10]. Interestingly, we also find that alignment between supervisor and student explorations is negatively associated with the number of papers published by the supervisor: a Q1-Q3 increase in alignment is associated with a 13% decrease in the supervisor's publications. These two negative associations may be explained by the higher effort dedicated by the supervisor to students' training activities when she and her students work on aligned research subjects that are new for the supervisor. Indeed, having a research subject that aligns with her students might increase the supervisor's commitment to their training, subtracting time from her research activities and, ultimately, decreasing the quantity of her scientific output. This is especially true if students are working in fields unfamiliar to their supervisor.

As for the impact of the supervisor's research in Column 2, similar to Hill et al. (2025), we observe that supervisor exploration is negatively associated with the number of citations received. A Q1-Q3 increase in *Supervisor exploration* is associated with a 1.6% decrease in the citations received by the supervisor's publications in *t*. This negative relationship can be explained by the fact that exploration leads the supervisor to publish in unfamiliar fields (as

[10] We calculate the percentage change variation of the continuous dependent variables according to a Q1-Q3 variation of the independent variable, similarly to the case of the percentage variation of the *Number of publications* according to a Q1-Q3 variation in the *Supervisor exploration*.

shown in Column 5), which hinders her visibility and ultimately decreases the impact of her research as measured by the citations received. In other words, supervisors' lack of reputation in the new fields explored may harm the citations received for their work (Hill et al., 2025). Another explanation can be that exploring new fields in which the supervisor is inexperienced leads her to publish research that is less original and, therefore, less cited by colleagues. Different from supervisor exploration, student exploration is beneficial to the impact of their supervisors' work. A Q1-Q3 increase in *Student exploration* is associated with a 3.3% increase in the citations received by the supervisor. The positive relationship between students' exploration and the citations received by their supervisors' work can be explained by the fact that students acquire new knowledge through exploration and transfer this knowledge to their supervisors, thereby enhancing the quality of their supervisors' work. A second possible explanation is that students disseminate information about supervisors' work, increasing supervisors' visibility and leading them to receive more citations. Both these explanations are coherent with the positive relationship observed between *Alignment* and the citations to the supervisor's work shown in Column 2: the more aligned the students' and supervisors' research is, the higher the probability that students bring useful knowledge from their exploration activities to their supervisor, and the higher the probability that they disseminate information related to their supervisor's research. Interestingly, the probability of producing top-cited research (*Top 10*) is associated with supervisor exploration and alignment but not with student exploration[11]. A Q1-Q3 increase in *Supervisor exploitation* is associated with 0.6 percentage points in the probability of publishing top-cited work (6% of the unconditional probability of producing top-cited research reported in Table 3), while a Q1-Q3 increase in *Alignment* is associated with a 1.5 percentage points increase in the probability of publishing top-cited research (15% of the unconditional probability of producing top-cited research).

As for the novelty of the supervisor's research, the probability of producing global novel research (*Global novelty*) is negatively associated with the supervisor's alignment with students' research (Column 4). Specifically, a Q1-Q3 increase in *Alignment* is associated with a decrease of 0.4 percentage points in the probability of publishing global novel research, accounting for 5.71% of the unconditional mean.

[11]We calculate the percentage point variation of the binary dependent variable *Top 10* according to a Q1-Q3 variation of the independent variable as $\widehat{Pr(y)}|_{x_{Q3}} - \widehat{Pr(y)}|_{x_{Q1}}$, where y is the *Top 10* and x is the *Supervisor exploration*. Similarly, we calculate the percentage point variation of the dependent variable across all the other regression models explaining a binary variable, i.e., *Global novelty* and *Individual novelty*.

Contrary to global novelty, the supervisor's probability of publishing individual novel research (*Individual novelty*) is positively associated with the supervisor's exploration and alignment (Column 5). Specifically, a Q1-Q3 increase in *Supervisor exploration* is associated with an 11.4 percentage point increase in the probability of publishing in a new journal, i.e., producing individual novel research, while a Q1-Q3 increase in *Alignment* is associated with a 2.6 percentage point decrease in the probability of publishing in a new journal. The positive association between the supervisor's probability of publishing individual novel research and her exploration activity is expected, given that learning about new research fields is the primary goal of the supervisor's exploration activity.

Overall, our results indicate that student exploration activities and their alignment with the supervisor's research are crucial determinants of the supervisor's performance.

Table 4: OLS estimates of the fully interacted model explaining supervisor's performances (Equation 1). Average marginal effects reported.

	(1)	(2)	(3)	(4)	(5)
	Quantity	Impact		Novelty	
	OLS	OLS	OLS	OLS	OLS
Variables	Number of Publications (log)	Citations (log)	Top 10	Global novelty	Individual novelty
Supervisor exploration°	0.87***	-0.20**	0.065**	-0.0043	1.07***
	(0.064)	(0.093)	(0.028)	(0.020)	(0.043)
Student exploration°	-0.23**	0.40***	0.065	0.045	0.052
	(0.10)	(0.14)	(0.047)	(0.036)	(0.063)
Alignment°	-1.07***	0.86***	0.15***	-0.032*	-0.17***
	(0.054)	(0.082)	(0.024)	(0.017)	(0.038)
1 PhD	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*
2 PhDs	0.027***	-0.0019	-0.0022	0.0065***	0.0012
	(0.0060)	(0.0081)	(0.0029)	(0.0025)	(0.0041)
3 PhDs	0.066***	0.011	-0.0017	0.0055*	0.0080
	(0.0081)	(0.011)	(0.0039)	(0.0032)	(0.0052)
4 PhDs or more	0.13***	0.021	-0.0051	0.015***	0.025***
	(0.010)	(0.013)	(0.0046)	(0.0039)	(0.0063)
N. of past supervisor pub. (log)	-0.010	0.000028	0.0025	-0.0030	-0.067***
	(0.0066)	(0.0084)	(0.0027)	(0.0021)	(0.0037)
Co-supervisor	YES	YES	YES	YES	YES
Supervision discipline FE	YES	YES	YES	YES	YES
Calendar year FE	YES	YES	YES	YES	YES
Supervisor FE	YES	YES	YES	YES	YES
Observations	120,935	120,935	120,935	120,935	120,935
R-squared	0.613	0.458	0.281	0.202	0.235
Number of supervisors	14,978	14,978	14,978	14,978	14,978

NOTE: °The average marginal effects reported in the table refer to the fully interacted model in Equation 1, with squared terms of the variables *Supervisor exploration*, *Student exploration*, and *Alignment*. The estimates of all the coefficients are reported in Table B1 in Appendix B.

3.1 Further analysis

In this section, we conduct two additional analyses. The first analysis reported in Section 3.1.1 aims to assess the bias introduced in the estimates of the *Supervisor exploration* coefficient when only the supervisor’s exploration activity is considered to explain the supervisor's performance. The second analysis, reported in Section 3.1.2, examines the discipline heterogeneity associated with the estimates of the relationships between the supervisor’s performance measures and the variables *Supervisor exploration*, *Student exploration,* and *Alignment*.

3.1.1 Analysis considering *Supervisor exploration* only

In previous studies, exploration has been considered a researcher’s individual choice (Hill et al., 2025). In this section, we follow the same approach, assuming that exploration is limited to the supervisor, neglecting the other team members. In other words, we replicate the analysis reported in Table 4, but remove *Student exploration* and *Alignment* from the model specification. The difference in the marginal effects estimated for *Supervisor exploration* in Table 4 and those estimated in Table 5 quantifies the bias introduced by omitting *Student exploration* and *Alignment.* As for the marginal effects of *Supervisor exploration* calculated for the *Number of publications* (Column 1) and *Citations* (Column 2) in Table 5, we find that, although they maintain the same sign and statistical significance as in Table 4, they are substantially larger in absolute values. Moreover, the marginal effect of *Supervisor exploration* is negatively associated with the probability of being a top-cited supervisor (Column 3), inverting the result obtained in Table 4. The marginal effects calculated for *Supervisor exploration* in Columns 5 and 6 maintain the same sign and statistical significance as those calculated in the same columns of Table 4. Overall, comparing Tables 4 and 5 reveals that omitting *Student exploration* and *Alignment* from the model specification substantially biases the estimates of the coefficient of *Supervisor exploration*, especially for variables measuring the quantity and impact of the supervisor’s performance.

Table 5: OLS estimates of the model explaining supervisor's performances in case we include only *Supervisor exploration* and controls as explanatory variables. Average marginal effects reported.

	(1)	(2)	(3)	(4)	(5)
	Quantity	Impact		Novelty	
	OLS	OLS	OLS	OLS	OLS
Variables	Number of Publications (log)	Citations (log)	Top 10	Global novelty	Individual novelty
Supervisor exploration°	1.43***	-0.92***	-0.046**	0.0100	1.11***
	(0.044)	(0.066)	(0.020)	(0.014)	(0.030)
Other controls°°	YES	YES	YES	YES	YES
Co-supervisor	YES	YES	YES	YES	YES
Supervision discipline FE	YES	YES	YES	YES	YES
Calendar year FE	YES	YES	YES	YES	YES
Supervisor FE	YES	YES	YES	YES	YES
Observations	120,935	120,935	120,935	120,935	120,935
R-squared	0.613	0.458	0.281	0.202	0.235
Number of supervisors	14,978	14,978	14,978	14,978	14,978

NOTE: °The average marginal effects reported in the table refer to the model reported in Equation 1, from which we removed the variables *Student exploration*, *Alignment*, and the interaction terms involving one of these two variables. °°The models estimated in this table include all time-variant control variables that are also included in the models estimated in Table 4 (Other controls).

3.1.2 Discipline heterogeneity

In this section, we analyze how the association between the exploration variables and the supervisor's performance varies across disciplines. The estimates of the model specification described in Equation 1 by discipline are reported in Table 6. Column 1 shows a positive association between the *Supervisor exploration* and *Number of publications*, and a negative association between the *Supervisor exploration* and *Number of publications*, across all disciplines. On the contrary, *Student exploration* is negatively associated with the supervisor's number of publications in Computer Science, Physics, and Biology, while it is positively associated with the supervisor's publications in Mathematics.

The positive association between *Alignment* and the two impact measures, i.e., *Citations* and *Top 10*, appears to be common across all disciplines. If we compare the results reported in Table 6 with those aggregated in Table 4, the negative association between *Citations* and *Supervisor exploration* is driven by the supervisors in Medicine, while the positive association between *Student exploration* and *Citations* is driven by supervisors in Chemistry, Medicine, Physics, and Biology. Similarly, the positive association between *Supervisor exploration* and *Top 10* is driven by supervisors in Chemistry and Engineering.

As for the novelty of the supervisor's outcomes, Computer Science, Biology, and Medicine seem to be the only fields in which our exploration variables are associated with the probability of producing *Global novelty*. As expected, *individual novelty* is positively associated with *Supervisor exploration* in all disciplines.

Table 6: OLS estimates of the fully interacted model explaining supervisor's performances (Equation 1), by discipline. Average marginal effects reported.

	(1)	(2)	(3)	(4)	(5)
	Quantity	Impact		Novelty	
	OLS	OLS	OLS	OLS	OLS
Variables	Number of Publications (log)	Citations (log)	Top 10	Global novelty	Individual novelty
Chemistry (n=16,546)	**1.33 [mean]**	**2.63 [mean]**	**0.10 [mean]**	**0.07 [mean]**	**0.80 [mean]**
Supervisor exploration°	1.30***	0.40*	0.17**	0.030	1.29***
	(0.17)	(0.21)	(0.073)	(0.053)	(0.11)
Student exploration°	0.099	0.75**	0.037	0.092	0.012
	(0.31)	(0.33)	(0.14)	(0.11)	(0.17)
Alignment°	-1.13***	1.22***	0.28***	0.011	-0.051
	(0.14)	(0.18)	(0.064)	(0.046)	(0.098)
Engineering (n=27,001)	**1.09**	**2.37**	**0.10**	**0.04**	**0.79**
Supervisor exploration	1.24***	-0.14	0.12*	0.035	1.11***
	(0.14)	(0.21)	(0.063)	(0.039)	(0.096)
Student exploration	-0.060	0.23	0.029	-0.0043	0.0088
	(0.24)	(0.32)	(0.097)	(0.069)	(0.14)
Alignment	-1.08***	0.67***	0.27***	-0.014	-0.27***
	(0.12)	(0.18)	(0.055)	(0.035)	(0.088)
Geology (n=8,703)	**1.14**	**2.92**	**0.11**	**0.07**	**0.75**
Supervisor exploration	1.56***	-0.21	0.065	-0.087	1.55***
	(0.24)	(0.33)	(0.12)	(0.086)	(0.19)
Student exploration	-0.24	-1.09*	-0.23	-0.19	-0.15
	(0.46)	(0.59)	(0.23)	(0.16)	(0.28)
Alignment	-1.42***	0.20	0.020	-0.10	-0.50***
	(0.21)	(0.29)	(0.10)	(0.080)	(0.17)
Mathematics (n=8,829)	**0.81**	**2.15**	**0.09**	**0.03**	**0.77**
Supervisor exploration	0.48**	-0.16	0.065	-0.017	0.80***
	(0.22)	(0.30)	(0.079)	(0.045)	(0.12)
Student exploration	0.51*	-0.13	0.21	-0.10	0.50**
	(0.31)	(0.45)	(0.13)	(0.085)	(0.22)
Alignment	-1.15***	0.68**	0.13*	-0.044	-0.16
	(0.18)	(0.27)	(0.072)	(0.042)	(0.12)
Computer science (n=10,575)	**0.78**	**2.24**	**0.11**	**0.04**	**0.77**
Supervisor exploration	0.38**	0.21	0.092	0.18***	1.06***
	(0.18)	(0.35)	(0.098)	(0.055)	(0.15)
Student exploration	-0.62*	0.61	0.079	0.15	0.067
	(0.34)	(0.54)	(0.16)	(0.11)	(0.22)
Alignment	-1.29***	1.40***	0.27***	0.10**	0.016
	(0.16)	(0.34)	(0.085)	(0.051)	(0.14)
Physics (n=26,088)	**1.22**	**2.51**	**0.11**	**0.05**	**0.73**
Supervisor exploration	1.11***	-0.33	-0.015	-0.051	1.42***
	(0.15)	(0.20)	(0.063)	(0.042)	(0.10)
Student exploration	-0.40*	0.53*	0.16	0.0049	-0.020
	(0.24)	(0.30)	(0.10)	(0.069)	(0.15)
Alignment	-1.49***	0.27	0.11**	-0.025	-0.18**
	(0.12)	(0.19)	(0.056)	(0.036)	(0.091)
Biology (n=36,019)	**1.10**	**2.92**	**0.11**	**0.09**	**0.83**
Supervisor exploration	1.34***	-0.089	0.081	0.024	1.17***
	(0.11)	(0.17)	(0.057)	(0.045)	(0.081)
Student exploration	-0.37*	0.52*	0.14	0.029	0.17
	(0.21)	(0.29)	(0.11)	(0.088)	(0.12)
Alignment	-1.14***	1.07***	0.11**	-0.12***	-0.30***
	(0.096)	(0.15)	(0.049)	(0.039)	(0.071)
Medicine (n=19,696)	**1.32**	**2.75**	**0.09**	**0.11**	**0.86**
Supervisor exploration	1.16***	-0.52*	0.026	-0.033	1.04***

	(0.16)	(0.26)	(0.077)	(0.068)	(0.10)
Student exploration	-0.39	0.88**	0.14	0.088	-0.30*
	(0.29)	(0.40)	(0.13)	(0.14)	(0.16)
Alignment	-1.17***	0.63***	0.040	-0.12**	-0.16*
	(0.13)	(0.23)	(0.066)	(0.055)	(0.087)
Other controls°°	YES	YES	YES	YES	YES
Supervision discipline FE°°	YES	YES	YES	YES	YES
Calendar year FE°°	YES	YES	YES	YES	YES
Supervisor FE°°	YES	YES	YES	YES	YES

NOTE: °The average marginal effects reported in the table refer to the fully interacted model in Equation 1, with squared terms of the three variables: *Supervisor exploration*, *Student exploration*, and *Alignment*. The mean values of the dependent variables in each discipline are reported in bold. °°The models estimated in this table include all time-variant control variables that are also included in the models estimated in Table 4 (Other controls).

4. Discussion

Exploration in scientific enterprises is a collective effort, and our analysis shows that the exploration attitude of every member of the research team plays a role in determining the researchers' performance. In our context, the team's exploration attitude is described by three variables: *Supervisor exploration*, *Student exploration*, and *Alignment*. In this section, we use an empirical approach based on our estimates of Equation 1 to identify the combinations of *Supervisor exploration*, *Student exploration*, and *Alignment* values that maximize each of the supervisor's performance measures.

Relying on the coefficients reported in Table B1 in Appendix B, we identify the combination of *Supervisor exploration*, *Student exploration*, and *Alignment* values that predict the maximum supervisor's performance. To do so, we proceed in four steps.

First, we group the observations in our study sample according to the decile values of *Supervisor exploration*, *Student exploration*, and *Alignment*, obtaining 10 groups of equal numerosity for each variable.

Second, we combine the 10 groups associated with each variable, resulting in a grid of 1,000 cells (= 10 deciles for *Supervisor exploration* x 10 deciles for *Student exploration* x 10 deciles for *Alignment*). Although each of these 1,000 cells theoretically contains observations, only 988 cells contain at least a combination of *Supervisor exploration*, *Student exploration*, and *Alignment* that exists in our study sample. For instance, none of the observations in our sample fall within the cell defined by the 1st decile of *Supervisor exploration*, 1st decile of Student exploration, and 1st decile of *Alignment*. On the contrary, the most populated cell, with 2.46% of the observations in our sample, is the 1st decile of *Supervisor exploration*, the 1st decile of *Student exploration*, and the 10th decile of *Alignment*. To ensure that the cells considered in our grid are populated by a minimum number of observations, and therefore represent combinations of *Supervisor exploration*, *Student exploration*, and *Alignment* that are relevant in our sample, we select the cells that contain at least 0.01% of the observations in the study sample.

Third, considering only the observations within each cell, we compute the median values of *Supervisor exploration*, *Student exploration*, and *Alignment*. Then, we input the median values of the three variables calculated for each decile into the models estimated in Table B1 to predict, for each cell in the grid, the supervisor's performance, i.e., $\widehat{Number\ of\ Publications}\ (log)$, $\widehat{Citations}\ (log)$, $\widehat{Top\ 10}$, $\widehat{Global\ Novelty}$, and $\widehat{Individual\ Novelty}$.

Finally, we identify the *Optimal strategy* that maximizes each supervisor's performance measure by searching within the grid for the cell that corresponds to the combination of *Supervisor exploration*, *Student exploration*, and *Alignment* that predicts the maximum performance value.

Table 7 illustrates the *Optimal strategy* characteristics, i.e., *Supervisor exploration*, *Student exploration*, *Alignment* values, and the corresponding predicted supervisor's performance. To assess the benefits of adopting the *Optimal strategy* on the supervisor's performance, we compare it with the *Average strategy*, which corresponds to setting *Supervisor exploration*, *Student exploration*, and *Alignment* at their average values[12]. Columns 1 and 2 in Table 7 report the predicted performance of adopting the *Optimal strategy* and the *Average strategy*, respectively. We observe that the predicted increase in the supervisor's performance is positive and statistically significant for all the indicators considered (Column 3 reports the P-value of a test for the statistical equivalence of the predicted performance measures for the two strategies). For instance, the predicted number of supervisor's publications grows by approximately 17% (=1.357-1.187) when switching from the *Average strategy* to the *Optimal strategy*.

Concerning the characteristics of the *Optimal strategy*, Columns 4, 5, and 6 report the values of the three exploration variables when the *Optimal strategy* is adopted, along with the percentage variation of these values from the ones characterizing the *Average strategy*. For instance, the optimal strategy that maximizes the supervisor's number of publications requires the supervisor to substantially increase her level of exploration by 67% with respect to the *Average strategy*. Interestingly, the *Optimal strategy* always requires the supervisor to increase her level of exploration, except when she aims to maximize citations. Similarly, when adopting optimal strategies, students are often required to set their exploration level 60% higher than the *Average strategy*, except when the supervisor aims to maximize the probability of being among the *Top 10* cited. Finally, a level of alignment between students' and supervisor's exploration

---

[12] The average values of *Supervisor exploration*, *Student exploration*, and *Alignment* are those reported in the descriptive statistics in Table 3

higher than *Average strategy* is needed to maximize the impact of supervisor's publications, while a lower level of alignment is needed to maximize publication quantity and novelty.

Taken together, these results show that the exploration strategy adopted by team members can significantly change the supervisor's predicted performance, providing supervisors with a precise exploration strategy to follow according to the performance measure they aim to maximize.

Table 7: Supervisor's predicted performance and exploration *Optimal strategy*.

	Supervisor's performance			Optimal strategy					
	(1)	(2)	(3)	(4)		(5)		(6)	
	Optimal strategy	Average strategy	P-value H0: (1)=(2)	Supervisor exploration	Δ%	Student exploration	Δ%	Alignment	Δ%
$\widehat{Number\ of\ Publications}\ (log)$	1.357	1.187	< 0.01	0.35	+67%	0.29	+60%	0.69	-18%
$\widehat{Citations}\ (log)$	2.721	2.624	< 0.01	0.19	-11%	0.29	+60%	0.94	+12%
$\widehat{Top\ 10}$	0.139	0.104	< 0.01	0.29	+36%	0.16	-10%	0.94	+12%
$\widehat{Global\ novelty}$	0.075	0.069	0.085	0.25	+20%	0.29	+60%	0.80	-5%
$\widehat{Individual\ novelty}$	0.947	0.821	< 0.01	0.35	+67%	0.29	+60%	0.76	-9%

NOTE: Columns 1 and 2 report the predicted supervisor's performance based on the estimations of Table B1 and obtained by applying two different exploration strategies: the *Optimal strategy* and the *Average strategy*. The optimal exploration strategy is defined by inputting the values reported in Columns 4, 5, and 6 to the variables *Supervisor exploration*, *Student exploration*, and *Alignment*, respectively, while the *Average strategy* is defined by inputting the values 0.21, 0.18, and 0.84 (reported in Table 3) to the three exploration variables. The Δ% in Columns 4, 5, and 6 reports the % variation of the three exploration variables from their values in the *Average strategy*. Column 3 reports the P-values of a test of the equivalence between the means of Columns 1 and 2.

5. Conclusion

Exploration is a crucial activity that drives creativity and innovation. In the context of scientific teams led by a supervisor and populated by PhD students, our analysis examines the relationship between the supervisor's performance and three exploration dimensions at the team level: *Supervisor exploration*, *Student exploration*, and *Alignment*. Overall, our results show that supervisors aiming to maximize their performance should carefully design their teams' exploration strategies.

The contribution of our study to the literature is threefold. First, it shows that efficiently conducting exploration in modern science does not depend only on the choices of the single researcher but requires coordination at the team level. Second, our study contributes to team management literature by proposing specific guidelines for team leaders to redesign their teams' exploratory activities. Finally, our study makes a methodological contribution to exploration studies by proposing a new approach that traces exploration through a semantic analysis of the document content. This methodology departs from the commonly used exploration measures based on document metadata. e.g., reference lists.

Our results are expected to raise policy-relevant issues. Indeed, exploration strategies, highly influenced by the supervisor's decisions, might have negative effects on the career trajectories of her PhD students, who have limited influence on defining these strategies. For instance, according to our results, a supervisor who aims to maximize citations to her publications may allocate a high level of exploration to her PhD students and reduce her own exploration. Being exploratory activities associated with a higher risk of failure than non-exploratory activities, the supervisor may be transferring the risk of exploration to her students. Therefore, guided by their supervisors toward exploration, students may end up undertaking too risky research avenues in the early phases of their careers, potentially harming their future career outcomes. One possible policy intervention to prevent students from taking excessive risks is to ask the *Comité de Suivi de Thèse*, a committee that evaluates the yearly progress of students, to monitor the exploration activity undertaken by the students.

Bibliography

Arts, S., Fleming, L., 2018. Paradise of Novelty—or Loss of Human Capital? Exploring New Fields and Inventive Output. Organization Science 29, 1074-1092.

Bourdieu, P., 1975. The Specificity of the Scientific Field and the Social Conditions for the Progress of Reason. Social Science Information 14, 19–47.

Delamont, S., Atkinson, P., 2001. Doctoring Uncertainty: Mastering Craft Knowledge. Social Studies of Science 31, 87-107.

Gupta, A.K., Smith, K.G., Shalley, C.E., 2006. The Interplay between Exploration and Exploitation. Academy of Management Journal 49, 693-706.

He, Z.-L., Wong, P.-K., 2004. Exploration Vs. Exploitation: An Empirical Test of the Ambidexterity Hypothesis. Organization Science 15, 481-494.

Hill, R., Yin, Y., Stein, C., Wang, X., Wang, D., Jones, B.F., 2025. The Pivot Penalty in Research. Nature 642, 999-1006.

Jansen, J.J.P., Van Den Bosch, F.a.J., Volberda, H.W., 2006. Exploratory Innovation, Exploitative Innovation, and Performance: Effects of Organizational Antecedents and Environmental Moderators. Management Science 52, 1661-1674.

Kuhn, T.S., 1977. The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago Press, Chicago.

March, J.G., 1991. Exploration and Exploitation in Organizational Learning. Organization Science 2, 71-87.

Mattsson, P., Shibayama, S., 2025. Shaping Scientific Careers: The Role of Early-Career Research Topic Dynamics in Fostering Novel Knowledge. Studies in Higher Education, 1-19.

Mcgrath, R.G., 2001. Exploratory Learning, Innovative Capacity, and Managerial Oversight. Academy of Management Journal 44, 118-131.

Merton, R.K., 1973. Sociology of Science. University of Chicago Press, Chicago.

Shibayama, S., Baba, Y., Walsh, J.P., 2015. Organizational Design of University Laboratories: Task Allocation and Lab Performance in Japanese Bioscience Laboratories. Research Policy 44, 610-622.

Shibayama, S., Mattsson, P., Broström, A., 2025. Risk in Science: The Socialization of Risk-Taking in Early-Career Training. Research Policy.

Stephan, P., Levin, S., 1992. Striking the Mother Lode in Science: The Importance of Age, Place, and Time. Oxford University Press, New York.

Tushman, M.L., O'reilly, C.a.I., 1996. Ambidextrous Organizations: Managing Evolutionary and Revolutionary Change. California Management Review 38, 8-30.

Wang, J., Lee, Y.-N., Walsh, J.P., 2018. Funding Model and Creativity in Science: Competitive Versus Block Funding and Status Contingency Effects. Research Policy 47, 1070-1083.

Wang, J., Shibayama, S., 2022. Mentorship and Creativity: Effects of Mentor Creativity and Mentoring Style. Research Policy 51, 104451.

Wang, J., Veugelers, R., Stephan, P., 2017. Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators. Research Policy 46, 1416-1436.

Wuchty, S., Jones, B.F., Uzzi, B., 2007. The Increasing Dominance of Teams in Production of Knowledge. Science 316, 1036-1039.

Zuckerman, H., 1988. The Sociology of Science, in: Smelser, N.J. (Ed.), Handbook of Sociology. Sage, Newbury Park, CA, pp. 511-574.

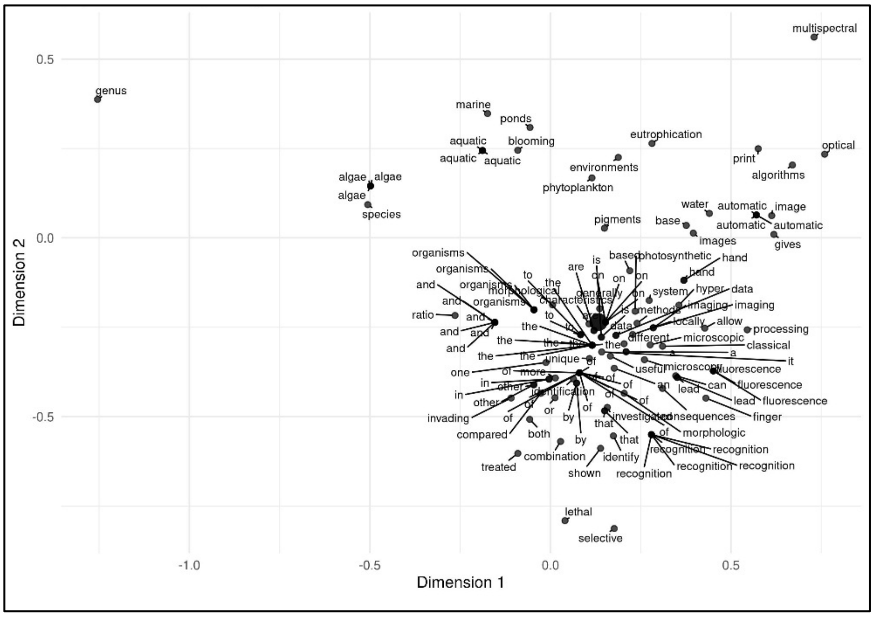
Appendix

Appendix A: Document content similarity

The calculation of the variables *Student exploration*, *Supervisor exploration*, and *Alignment* relies on assessing the similarity between text documents, such as scientific articles and PhD theses. To measure the similarity between documents, we proceed in three steps.

First, we produce a vectorial representation of each word included in the abstracts and titles of the documents considered in our analysis. To achieve this goal, we use the Word2Vec algorithm (Mikolov et al., 2013; Rong, 2014), which is trained on a corpus of 1,268,232 article abstracts in STEM published in English between 1990 and 2018 and having at least one author with a French affiliation. The Word2Vec algorithm's main inputs are the size of the vectors representing the semantic meaning of words and the corpus of documents used for training Word2Vec. In our case, we assume that the semantic meaning of each word can be well represented by a vector of real values representing the word in a 100-dimensional space. The corpus used for training Word2Vec includes 609,840 distinct words, enabling Word2Vec to create a vocabulary of equal size, where each word is associated with a 100-dimensional vector. The attribution of vectors to words is based on their co-occurrence in the text. Specifically, for each focal word (the "target" word), the algorithm attempts to predict the co-occurrence of words in its neighborhood (the "context" words). Once the words in the vocabulary are assigned a probability to appear as context words of the target word, Word2Vec compares the predicted context words with the actual context words observed in the training corpus to calculate a prediction error. Finally, Word2Vec updates the vectorial representation of the target word to minimize the prediction error. Each word in the corpus becomes, in turn, a target word, generating an iterative process, i.e., the algorithm's training. At the end of the training, two words with a similar semantic meaning are expected to be close in the 100-dimensional vector space, while words with different meanings are expected to be far. For instance, Figure 1 presents a simplified bidimensional representation of words, showing that words with similar semantic meaning, such as "marine" and "aquatic", are close in the vectorial space.

Figure A1: Vectorial representation of a set of words used in the abstract of a document included in our study sample.



NOTE: To obtain a convenient graphical representation of the vectorial space, we use Principal Component Analysis to reduce from 100 to 2 the dimensions of the vectorial space. All analyses are conducted in a 100-dimensional vector space.

Second, once we have obtained the vectorial representation of words, we calculate the vectorial representation of documents. To do so, we calculate the centroid of all vectors corresponding to the words appearing in the document abstract, obtaining a 100-dimensional vector for each document. In the bidimensional representation reported in Figure 1, the big dot at the center of the cloud of words corresponds to the vectorial representation of the document.

Third, we calculate the similarity between documents. Specifically, for each pair of documents, article-article or article-thesis, we calculate the cosine similarity between their vector representations. Cosine similarity is calculated as the ratio between the product of the two vectors representing the documents (numerator) and the product of their norms (denominator). The cosine similarity value ranges from -1 to 1, where 1 indicates content similarity between documents and -1 indicates content dissimilarity between documents. To obtain a measure of content dissimilarity between documents, as required by the *Supervisor exploration* and *Student exploration* variables, we calculated (1-cosine similarity), obtaining a measure of dissimilarity that ranges between 0 and 2.

References:

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (2013) (May 7, 2019).

X. Rong, word2vec Parameter Learning Explained. *arXiv:1411.2738 [cs]* (2014) (April 25, 2019).

Appendix B: Fully interacted regression model

Table B1: Regression analysis, regression coefficients

Variables	(1) OLS Number of Publications (log)	(2) OLS Citations (log)	(3) OLS Top 10	(4) OLS Global novelty	(5) OLS Individual novelty
Supervisor exploration	21.683***	0.5671	-0.6711*	1.1213***	8.8449***
	(1.0112)	(1.1940)	(0.3491)	(0.2291)	(0.5768)
Student exploration	4.0878***	1.8914*	0.8980***	0.006833	0.08996
	(0.7854)	(1.1337)	(0.3290)	(0.2068)	(0.4948)
Alignment	29.755***	1.9578	-1.0790***	1.4169***	10.372***
	(1.1467)	(1.2988)	(0.3579)	(0.2291)	(0.6318)
Supervisor exploration^2	-12.401***	-2.8001***	0.2194	-0.6526***	-5.1557***
	(0.5730)	(0.7681)	(0.2100)	(0.1382)	(0.3578)
Supervisor exploration x Student exploration	1.4603*	0.9180	-0.7143**	0.06293	0.5950
	(0.8077)	(1.1098)	(0.3040)	(0.2007)	(0.5094)
Supervisor exploration x Alignment	-18.963***	0.2792	0.9195***	-1.0313***	-6.8429***
	(0.9412)	(1.0854)	(0.3198)	(0.2065)	(0.5287)
Student exploration^2	-0.4940	-0.6601	-0.5660	0.04724	0.4067
	(0.7864)	(1.0269)	(0.3564)	(0.2288)	(0.4331)
Student exploration x Alignment	-5.2967***	-1.7281*	-0.5762**	0.01012	-0.3638
	(0.6975)	(0.9749)	(0.2659)	(0.1666)	(0.4396)
Alignment^2	-15.454***	-0.5051	0.6771***	-0.7362***	-5.3943***
	(0.5544)	(0.6293)	(0.1732)	(0.1108)	(0.3077)
1 active PhD	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*
2 active PhDs	0.02710***	-0.001934	-0.002161	0.006491***	0.001213
	(0.005983)	(0.008145)	(0.002880)	(0.002452)	(0.004093)
3 active PhDs	0.06601***	0.01141	-0.001704	0.005461*	0.007969
	(0.008053)	(0.01080)	(0.003874)	(0.003183)	(0.005247)
4 active PhDs or more	0.1322***	0.02075	-0.005071	0.01466***	0.02503***
	(0.01025)	(0.01299)	(0.004635)	(0.003905)	(0.006312)
N. supervisor pub. (log)	-0.01023	2.793e-05	0.002549	-0.002981	-0.06671***
	(0.006621)	(0.008372)	(0.002671)	(0.002103)	(0.003718)
Co-supervisor	-0.004642	-0.002696	-0.001557	0.003110	0.005730
	(0.005831)	(0.007658)	(0.002712)	(0.002344)	(0.003749)
*Calendar year * Supervision discipline*	*YES*	*YES*	*YES*	*YES*	*YES*
Observations	120,935	120,935	120,935	120,935	120,935
Number of supervisors	14,978	14,978	14,978	14,978	14,978
R2	0.613	0.458	0.281	0.202	0.235
R2 within	0.05574	0.07690	0.001714	0.005909	0.02675

Clustered standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

DOCUMENTS DE TRAVAIL GREDEG PARUS EN 2025
GREDEG Working Papers Released in 2025

2025-01 Bruno Deffains & Frédéric Marty
Generative Artificial Intelligence and Revolution of Market for Legal Services

2025-02 Annie L. Cot & Muriel Dal Pont Legrand
"Making war to war" or How to Train Elites about European Economic Ideas: Keynes's Articles Published in L'Europe Nouvelle during the Interwar Period

2025-03 Thierry Kirat & Frédéric Marty
Political Capitalism and Constitutional Doctrine. Originalism in the U.S. Federal Courts

2025-04 Laurent Bailly, Rania Belgaied, Thomas Jobert & Benjamin Montmartin
The Socioeconomic Determinants of Pandemics: A Spatial Methodological Approach with Evidence from COVID-19 in Nice, France

2025-05 Samuel De La Cruz Solal
Co-design of Behavioural Public Policies: Epistemic Promises and Challenges

2025-06 Jérôme Ballet, Damien Bazin, Frédéric Thomas & François-Régis Mahieu
Social Justice: The Missing Link in Sustainable Development

2025-07 Jérôme Ballet & Damien Bazin
Revoir notre manière d'habiter le monde. Pour un croisement de trois mouvements de pensée: capabilités, services écosystémiques, communs

2025-08 Frédéric Marty
Application des DMA et DSA en France : analyse de l'architecture institutionnelle et des dynamiques de régulation

2025-09 Éléonore Dodivers & Ismaël Rafaï
Uncovering the Fairness of AI: Exploring Focal Point, Inequality Aversion, and Altruism in ChatGPT's Dictator Game Decisions

2025-10 Hayk Sargsyan, Aleksandr Grigoryan & Olivier Bruno
Deposits Market Exclusion and the Emergence of Premium Banks

2025-11 Lubica Stiblarova & Anna Tykhonenko
Talent vs. Hard Work: On the Heterogeneous Role of Human Capital in FDI Across EU Member States

2025-12 Mathieu Chevrier
Social Reputation as one of the Key Driver of AI Over-Reliance: An Experimental Test with ChatGPT-3.5

2025-13 Rania Belgaied
L'accessibilité aux médecins généralistes libéraux pour la ville de Nice

2025-14 Lorenzo Corno, Giovanni Dosi & Luigi Marengo
Behaviours and Learning in Complex Evolving Economies

2025-15 Grégory Donnat
Real Exchange Rate and External Public Debt in Emerging and Developing Countries

2025-16 Michela Chessa
Politics as A (Very) Complex System: A New Methodological Approach to Studying Fragmentation within a Council

2025-17 Mathieu Lambotte & Benjamin Montmartin
Competition, Conformism and the Voluntary Adoption of Policies Designed to Freeze Prices

2025-18 LEONARDO CIAMBEZI & ALESSANDRO PIETROPAOLI
Relative Price Shocks and Inequality: Evidence from Italy

2025-19 MATTEO ORLANDINI, SEBASTIANO MICHELE ZEMA, MAURO NAPOLETANO & GIORGIO FAGIOLO
A Network Approach to Volatility Diffusion and Forecasting in Global Financial Markets

2025-20 DANIELE COLOMBO & FRANCESCO TONI
Understanding Gas Price Shocks: Elasticities, Volatility and Macroeconomic Transmission

2025-21 SIMONE VANNUCCINI
Move Fast and Integrate Things: The Making of a European Industrial Policy for Artificial Intelligence

2025-22 PATRICE BOUGETTE, OLIVER BUDZINSKI & FRÉDÉRIC MARTY
Revisiting Behavioral Merger Remedies in Turbulent Markets: A Framework for Dynamic Competition

2025-23 AMAL BEN KHALED, RAMI HAJ KACEM & NATHALIE LAZARIC
Assessing the Role of Governance and Environmental Taxes in Driving Energy Transitions: Evidence from High-income Countries

2025-24 SANDYE GLORIA
Emergence. Another Look at the Mengerian Theory of Money

2025-25 BARBARA BRIXOVÁ, ANNA TYKHONENKO, ĽUBICA ŠTIBLÁROVÁ & MARIANNA SINIČÁKOVÁ
Middle Income Convergence Trap Phenomenon in CEE Countries

2025-26 GRÉGORY DONNAT, MAXIME MENUET, ALEXANDRU MINEA & PATRICK VILLIEU
Does Public Debt Impair Total Factor Productivity?

2025-27 NATHALIE LAZARIC, LOUBNA ECHAJARI & DOROTA LESZCZYŃSKA
The Multiplicity of Paths to Sustainability, Grand Challenges and Routine Changes: The Long Road for Bordeaux Winemakers

2025-28 OLIVIER BRETTE & NATHALIE LAZARIC
Some Milestones for an Evolutionary-Institutional Approach to the Circular Economy Transition

2025-29 PIERRE BOUTROS, ELIANA DIODATI, MICHELE PEZZONI & FABIANA VISENTIN
Does Training in AI Affect PhD Students' Careers? Evidence from France

2025-30 LUCA BARGNA, DAVIDE LA TORRE, ROSARIO MAGGISTRO & BENJAMIN MONTMARTIN
Balancing Health and Sustainability: Optimizing Investments in Organic vs. Conventional Agriculture Through Pesticide Reduction

2025-31 JEAN-LUC GAFFARD
Equilibre général et évolution: un éclairage théorique

2025-32 PATRICE BOUGETTE, FRÉDÉRIC MARTY & SIMONE VANNUCCINI
Competition Law Enforcement in Dynamic Markets: Proposing a Flexible Trade-off between Fines and Behavioural Injunctions

2025-33 TAEKYUN KIM, GABRIELA FUENTES & SIMONE VANNUCCINI
When Robots Met Language (Models): An Exploration of Science and Strategy in the vision-language-action Models Space

2025-34 FRÉDÉRIC MARTY & THIERRY WARIN
Prévenir les risques de parallélisme de comportements par développeur commun : de la constatation des limites des règles de concurrence au retour du Nouvel Instrument Concurrentiel

2025-35 LUCA FONTANELLI, MAURO NAPOLETANO & ANGELO SECCHI
Rethinking Volatility Scaling in Firm Growth

2025-36 CHARLIE JOYEZ
Connectivity and Contagion: How Industry Networks Shape the Transmission of Shocks in Global Value Chains

2025-37 Frédéric Marty
L'expérience de l'Union Européenne en matière de contrôle des subventions étrangères - quels enseignements pour la Nouvelle Calédonie ?

2025-38 Frédéric Marty & Thierry Warin
Data and Computing Power: The New Frontiers of Competition in Generative AI

2025-39 Jean-Luc Gaffard
Finance et globalisation : les enjeux d'une mutation

2025-40 Jean-Luc Gaffard
Crise de la globalisation et mutations du capitalisme

2025-41 Sara Amoroso & Simone Vannuccini
Teaming up with Large R&D Investors: Good or Bad for Knowledge Production and Diffusion?

2025-42 Ilona Dielen & Jeanne Mouton
Impact Assessment of the "Damages Directive": Abuse of Dominance Cases

2025-43 Jean-Luc Gaffard & Mauro Napoletano
Vers une nouvelle industrialisation : une stratégie polycentrique

2025-44 Mauro Napoletano & Francesco Toni
The Tartar Steppe of Italian Growth: Strategies for Renewal in a Slowing Europe

2025-45 Laurence Saglietto & Sylvie Graziani-Invernon
Les relations auditeur/audité au sein de micro-espaces de qualité

2025-46 Charlie Joyez
Robustness with Adaptation. Ownership Networks of Multinationals through COVID-19

2025-47 Thierry Blayac, Patrice Bougette & Jules Duberga
Share the Ride: The Determinants of Long-Distance Carpooling Pricing Strategies in France

2025-48 Jeremy Srouji, Dominique Torre & Qing Xu
Should BRICS Members Adopt a Wholesale Digital Payment System?

2025-49 Raffaele Miniaci, Michele Pezzoni & Sotaro Shibayama
Exploration in Research Teams: Building on the Shoulders of PhD Students