What matters in funding: The value of research coherence and alignment in evaluators' decisions

Maastricht University

UNITED NATIONS UNIVERSITY

UNU-MERIT

Working Paper Series

#2020-010

What matters in funding: The value of research coherence and alignment in evaluators' decisions

Charles Ayoubi, Sandra Barbosu, Michele Pezzoni and Fabiana Visentin

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: **info@merit.unu.edu** | website: **http://www.merit.unu.edu**

Boschstraat 24, 6211 AX Maastricht, The Netherlands
Tel: (31) (43) 388 44 00

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology
UNU-MERIT

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT to stimulate discussion on the issues raised.

What matters in funding:

The value of research coherence and alignment in evaluators' decisions

Charles Ayoubi*
Chair in Economics and Management of Innovation - École Polytechnique Fédérale de Lausanne
charles.ayoubi@epfl.ch

Sandra Barbosu
Alfred P. Sloan Foundation
barbosu@sloan.org

Michele Pezzoni
Université Côte d'Azur, CNRS, GREDEG, France;
Observatoire des Sciences et Techniques, HCERES, Paris, France;
ICRIOS, Bocconi University, Milan, Italy;
michele.pezzoni@unice.fr

Fabiana Visentin**
UNU-MERIT, Maastricht University, the Netherlands
visentin@merit.unu.edu

Abstract: Entrepreneurs, managers, and scientists participate in competitive selection processes to obtain resources. The project they propose is a crucial aspect of their success. In this paper, we focus on the selection of scientists applying for academic funding by submitting a research proposal. We argue that two core dimensions of the research proposal affect the probability of funding success: its *coherence* with the applicant's previous work, and its *alignment* with subjects of general interest for the scientific community. Employing a neural network algorithm, we analyse the text of 2,494 research proposals for a prestigious fellowship awarded to promising early-stage North American researchers. We find field-specific heterogeneity in the committees' evaluations. In life sciences and chemistry, evaluators value the research proposal's coherence positively with the scientist's recent work and the proposals' alignment with the current subject of general interest for the scientific community. Conversely, in physics, evaluators give more weight to bibliometric indicators and less to the proposal coherence and alignment. Our results can be extended beyond the academic context to managerial implications in cases such as entrepreneurs and managers submitting project proposals to investors.

Keywords: Research trajectories, research funding, coherence, alignment

JEL codes: I23, O38

**Authors are in alphabetical order. All contributed equally. ** Corresponding author*

Introduction

Competitive selection processes are prevalent in many arenas. Entrepreneurs having to persuade investors to fund their start-ups (Astebro and Elhedhli, 2006; Scott et al., 2015), job candidates going through hiring processes and interviews (Burton and Beckman, 2007; Dahl & Klepper, 2015; Noe et al. 2017), and scientists drafting proposals to sponsor their research (Jacob & Lefgren, 2011) all face fierce competition. A core concern for any such candidate is to identify the factors affecting the probability of being selected. The impact of several salient factors, such as gender, ethnicity, and skills on success, have been extensively studied in various contexts (Bohnet et al. 2015; Ginther et al., 2011; Scott et al., 2015). However, the effect of the detailed content of the project proposed on the probability of being selected remains rather unexplored. We address this gap in the context of scientific research funding using novel data on applications from the Alfred P. Sloan Foundation's Sloan Research Fellowship (SRF) program. We explore two core dimensions affecting an applicant's probability of success: the *coherence* of the proposal submitted with their previous research, and the *alignment* of the candidate's proposal with research trends in the scientific community. We consider these two dimensions as representing the research trajectory chosen by the scientist. Using a neural network algorithm, we compare a scientist's research statement, included in the proposal, with both her past publications (coherence) and with publications in top generalist scientific journals (alignment) and estimate these two measures' impact on the probability of receiving funding.

The coherence of a proposal aims to assess the degree of similarity between the future research directions of an applicant and her previously published work. Nowadays, with the increasing difficulty in accessing funds, researchers seeking to finance their labs behave similarly to entrepreneurs aiming to attract investments for their start-ups (Etzkovitz, 2003). In the context of venture capital investment, several studies have investigated whether it is the project or the entrepreneur characteristics that make a winning start up (Kaplan et al. 2009; Zhang, 2011; Mitteness et al. 2012). In most literature on the subject, entrepreneurs are classified according to salient macro classifications such as age, network, and previous career positions, neglecting the detailed content of their previous work. In the context of scientific research funding, the richness of our data allows us to go beyond this limitation and to identify the fine-grained content of an individual's previous experience. Precisely, we can follow the candidates' previous work history codified in their publication paper trail (Gläser & Laudel, 2009; Franzoni et al., 2009). Doing so,

we can evaluate the actual content of earlier work and infer the expertise of an individual. From the publication text, we capture the subjects on which an applicant has previously worked and compare those subjects with the ones described in her proposal. Furthermore, we add a temporal dimension to take into account the depreciation of knowledge capital accumulation over time (Boone et al. 2008). We integrate this dimension by estimating the time elapsed since the moment an applicant has explored the subject of the research proposal in a previous scientific publication.

Access to the detailed content of a research statement also allows us to assess the impact of the alignment of the proposal with research trends in the scientific community. Previous studies have shown that the researcher's subject choice tends to conform to the scientific orthodoxy (Foster, et al. 2015; Corsi et al. 2019). Scientists are incentivised to embrace traditional and mainstream subjects that are more rewarded and to discharge novel subjects (Boudreau et al. 2016). However, by inferring the subjects studied by scientists solely from their published work, previous studies face two main limitations. First, published papers represent only one part of a scientist's work, the observable and successful part. Second, the subjects of published papers might be the result of filtering activities by mentors, co-authors, and reviewers. Our empirical setting allows us to overcome this limitation as we capture the subjects that scientists intend to explore and not only the ones eventually leading to publications. Moreover, scientists in our sample are autonomous young scholars applying for a grant aiming to support their career choices. They express in the research proposal submitted their unconstrained choice of the subjects in which they want to invest time and effort. To evaluate the alignment of a scientist's research statement with well-accepted subjects, we estimate the research statement similarity with all the articles published in Nature and Science over the last two decades. We assume Nature and Science, being two multidisciplinary journals, publish articles on issues relevant for the entire scientific community. The research statement can either dig deeper into questions in line with previously highly published topics as confirmed by a top generalist journal publication or explore new strands of research. Furthermore, to take into account the obsolescence of the subject (Sorensen and Stewart, 2000) with which the proposal is aligned, we also add a temporal dimension. Specifically, we include in our analysis the time elapsed since the subject was published in Nature or Science.

Coherence and alignment can affect the selection committee's decision through several mechanisms. On the one hand, evaluators may appreciate coherence if they perceive exploiting the extant expertise as a low-risk investment (Levinthal and March 1993) and a signal of the commitment in creating a focused identity (Zuckerman et al. 2003). The broadness of a research agenda is often perceived as riskier, less attractive, and less impactful by reviewers, compared to a more coherent agenda (Bateman 2015). We can expect that the coherence of a scientist's research agenda could be considered a positive signal by evaluators. On the other hand, funding institutions also intended to finance novel interdisciplinary research with high levels of uncertainty that would otherwise remain under-provisioned (Nelson, 1959; Arrow, 1972; Stephan, 2012) and often express a desire to do so[1]. Therefore, researchers with less coherent profiles might be perceived as competent to run such ambitious projects. Coherence might be seen as a signal of lack of flexibility (Pontikes 2012), being the researcher stuck in her 'comfort zone' (Evans 2019) and failing in adapting to future environmental changes (March 2003).

Regarding the alignment of a scientist's research path with articles published in top generalist journals, applicants who study trendy subjects with a broad audience may be considered more relevant and therefore be more likely to receive funding. Non-alignment with issues considered as highly relevant for the scientific community might be penalised by the selection committee. In fact, in economics, Corsi et al. (2019) argue that not conforming to mainstream subjects is detrimental to the likelihood of obtaining a top-tier position. Also, as recently raised by Oswald and Stern (2019), new subjects take time to emerge and be accepted in the field and published. Therefore, scientists might prefer to stick to research lines with an established scientific interest. On the other hand, if evaluators perceive the alignment as a lack of originality and replication of existing studies, they might penalise the choice (Foster et al. 2015; Stephan 1996).

For our analysis, we use a novel dataset of 2,011 young scientists who apply for the Sloan Research Fellowship (SRF) program, one of the most prestigious programs supporting early-career researchers in North America. For the period 2015-2019, we collected 2,494 complete application packages, including the applicants' CVs and research proposals. A unique,

[1] https://erc.europa.eu/funding/advanced-grants
https://www.nih.gov/news-events/news-releases/2019-nih-directors-awards-high-risk-high-reward-research-program-announced
https://www.nsf.gov/about/transformative_research/submit.jsp

fundamental feature of our data is the availability of the full-text research statement where scientists outline their 2-year future research plans. We complement the application package data with the applicants' publication data. Specifically, we gather the abstracts of all the papers published by each scientist until the application date. Then, we construct *Coherence* and *Alignment* measures. To identify the unbiased effect of coherence and alignment, we add detailed information regarding the applicant's background – age, gender, Ph.D. completion date, and institution, as well as current affiliation – and the scientist's publication record – number of publications, citations received, and number of co-authors. Finally, we construct a measure of career specialisation based on the scientist's past publications. Controlling for scientist specialisation is crucial in our analysis since, as recently shown by Nagle and Teodoridis (2019), as long as a scientist has a solid prior set of skills, her ability to diversify and integrate various types of knowledge leads to more impactful discoveries and could, therefore, be appreciated by the funding agency.

We find evidence of heterogeneity across scientific fields. In Life Sciences & Chemistry, the coherence between an applicant's proposal and her current research increases by 6.6 percentage points the probability of obtaining the fellowship, although the positive effect erodes over time. Similarly, alignment with current subjects of general interest is rewarded with a ten percentage point higher probability of obtaining the fellowship. This latter advantage decreases over time according to the obsolescence of the subject to which the proposal is aligned. In physics, coherence and alignment do not affect the chances of obtaining the fellowship. In this field, bibliometric indicators weight the most in evaluators' decisions.

Understanding the effect of scientists' research subject selection on the reward provided by the scientific community remains a widely unexplored subject, although crucial for both individual decision-making and policy considerations, with Tirole (2017) recently calling for more empirical research on the topic. We contribute to this with our analysis by evaluating the incentives that funding schemes give in terms of subject selection for young researchers. Our findings have important policy implications, suggesting to scientists the most rewarding choices when developing their future research plans. Evaluator committees appear to be rewarding coherent research trajectories, i.e., research agendas through which scientists build upon previous research, with a preference for recent research. This finding suggests a preference in funding

research trajectories where future knowledge incrementally builds upon existing knowledge, penalising "radical jumps." Moreover, scientists seem to be rewarded when pursuing research in subjects aligned with the current general scientific interest. Interestingly, for those scientists working in fields dominated by large labs, like physics, where it is challenging to attribute individual contribution, and the choice of research subject tends to be a collegial decision, the bibliographic profile – the number of publications and citations received – seem to remain a key aspect in funding decisions.

The remainder of the paper proceeds as follows. Section 2 describes our data and empirical setting. Section 3 presents our empirical strategy and the main results, including several robustness checks. Section 4 discusses and interprets the results, and concludes.

2. Data and Empirical Setting

2.1 Institutional context

In this paper, we use novel data from the Alfred P. Sloan Foundation's Sloan Research Fellowship (SRF) program. The program, founded in 1955, sponsors promising early-career scientists. Eligible candidates are tenure–track assistant professors employed at a university in the United States or Canada, who obtained their PhDs within six years of the date of application. The fellowship is offered in eight fields: chemistry, computer science, economics, mathematics, molecular biology, neuroscience, ocean sciences, and physics. The two-year fellowships "are awarded yearly to 126 researchers in recognition of distinguished performance and a unique potential to make substantial contributions to their field" (Alfred P. Sloan Foundation website). The fellowship consists of a financial award of roughly $70,000, meant to support the future recipients' career development, which "may be used by the fellow for any expense judged supportive of the fellow's research including staffing, professional travel, lab experiences, equipment, or summer salary support." "Fellows are selected on the basis of their independent research accomplishments, creativity, and potential to become leaders in the scientific community through their contributions to their field" (Alfred P. Sloan Foundation's website).

To apply for the fellowship, candidates submit an application package containing CV, selected publications, and a research statement with a detailed description of a 2-year research plan.

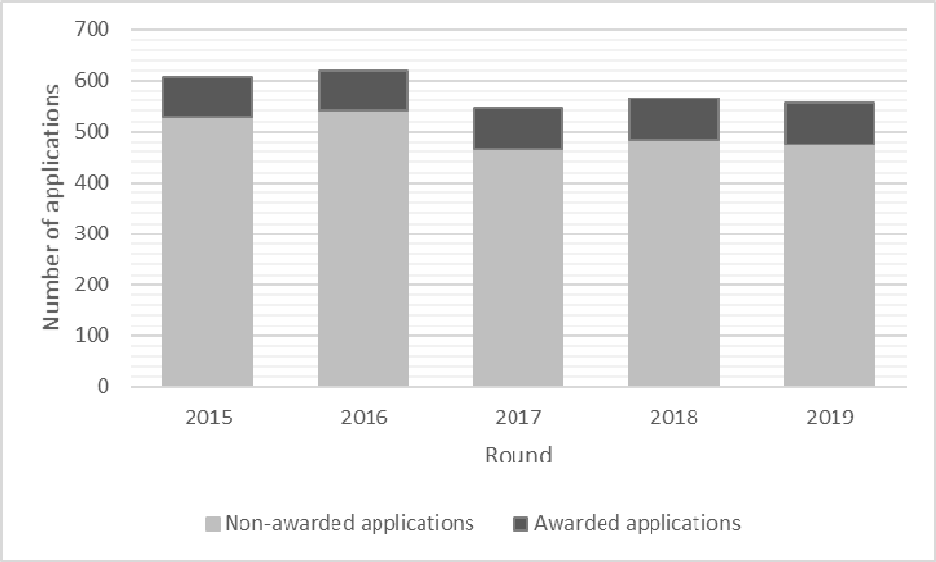
Applications are then reviewed, and winners selected by independent selection committees of three to four distinguished scientists in each field.

2.2 The study sample

Our dataset includes all the applications to the SRF in the period 2015-2019[2]. We collected information for 2,494 applications in the fields of computational & evolutionary molecular biology (CEMB), chemistry, neuroscience, ocean sciences, and physics[3]. Our primary sources of information are the complete application packages. We complement this with publication data retrieved from Elsevier's Scopus database.

As shown in Figure 1, the number of applications is roughly constant over the years. The highest number of applications is in physics (35%), followed by chemistry (24%), neuroscience (18%), CEMB (15%), and ocean science (9%). Across the years, the average fellowship application success rate is 16%.

Figure 1: Distribution of the number of applications per year



[2] Starting from 2015 data about applications have been systematically collected and available in electronic format.
[3] Those applications refer to 2,011 distinct scientists, since some of them applied multiple time. We excluded applications in the fields of Computer Science, Economics and Mathematics because in these three fields it is difficult to reconstruct reliable publication records. Conference proceedings as well as books that are relevant outcomes for scientists in those disciplines are not well covered in bibliometric dataset like the Elsevier's Scopus database.

2.3 The research trajectory: Coherence and alignment

To evaluate the coherence of a scientist's research trajectory, we exploit the information contained in the research statement and the scientist's previous publications. Since each scientist expresses her research plans in the research statement, we interpret the proposal's content as the scientist's future research agenda. Using advanced neural network text analysis techniques[4], we compare the content of all the scientist's previous publications (i.e. past research), with the content of her research statement (i.e. planned future research). 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 research statement and each publication preceding the application. Specifically, we extract from those publications the abstract texts and pair each abstract with the research statement text. Then, we calculate the similarity between the research statement and each publication that is a continuous measure varying on the interval [-1, 1] with 1 denoting a perfectly similar content. Overall, we use the text of 2,494 research statements and 52,499 publication abstracts. At the time of the application, scientists have, on average, 27.6 published papers. After computing the similarity scores of all the research statement-abstract pairs for each scientist, we consider that a scientist has a research statement coherent with her previous research if at least one of the research statement-abstract similarity scores is above a fixed threshold[5]. We construct the dummy variable *RS coherent* accordingly. In 66% of the applications, the scientist presents a coherent profile, i.e., her past and future research are similar to each other. Interestingly, it appears that evaluators tend to reward scientists with a coherent research trajectory: 73% of scientists awarded have a coherent profile versus 65% of non-awarded scientists.

We consider the content of scientists' previous work as well as how it has evolved over time. To add a temporal dimension, we identified in the publication list of each scientist the publication with the highest similarity score with her research statement. The variable *Years elapsed max coherence* equals the number of years between the SRF application year and the most similar publication to the research statement. In our sample, a scientist published the most similar article

[4] See Appendix A and B for the technical details about the implementation of text analysis techniques.
[5] We fixed the threshold at a similarity level of 0.85. Appendix C provides the technical details on the threshold selection.

to the research statement two years and nine months before the application (2.73 years) with no significant differences between awarded and non-awarded applicants.

To evaluate the alignment of the scientist's research with subjects of general interest in the field, we compare the content of the scientist's research statement with all the articles that appeared in Nature and Science in recent years, i.e., from 2000 to the application date. We consider whether Nature or Science articles treat topics similar to the ones described in the research statement, and the date of publication of those articles. Knowing that Nature and Science are two leading generalist journals publishing at the frontier of research in STEM scientific fields, we expect that if the research statement arguments have been treated by those journals, the topics are of general interest to the scientific community. We compare the scientists' research statement content with all the abstracts of the articles published in Nature and Science before the application date. We mark as aligned with a subject of general interest those scientists' research statements having a similarity score with one Nature or Science article above the fixed similarity threshold of 0.85, as identified in Appendix C. We define the dummy variable *RS aligned* accordingly. We find that 63% of applications exhibit a research statement aligned with subjects of general interest. The group of scientists with this characteristic appears more numerous in the subsample of awarded scientists, 71% versus 61% of the cases in the non-awarded subsample.

We consider that the more time that has passed between the subjects proposed in an applicant's research statement and the time they appeared in Nature or Science, the more the research statement focuses on obsolete topics. Hence, to add the time dimension, we include in our analysis the time elapsed from the application date to the most similar article published in the top two generalist journals. We then generate the variable *Years elapsed max alignment* accordingly. On average, a paper on Nature or Science similar to the research statement appears about 6.70 years before the application time, and there is a significant difference between the subsample of awarded and non-awarded applications: the value of the variable *Years elapsed max alignment* is significantly higher for the non-awarded applications (+0.68 years).

Table 1 reports the summary statistics of our measures of coherence of the research trajectory, and alignment with subjects of general interest, i.e., our main dependent variables.

Table 1: Summary statistics main dependent variables for the full sample, and the sub-samples of awarded and non-awarded applicants, respectively

Variable	All		Awarded	Non-Awarded	
	Average	Sd	Average	Average	t-test
Coherence of the research trajectory					
RS coherent (dummy)	0.66	0.47	0.73	0.65	0.00
Years elapsed max coherence*	2.73	2.20	2.59	2.76	0.24
Alignment with subjects of general interest					
RS aligned (dummy)	0.63	0.48	0.71	0.61	0.00
Years elapsed max alignment*	6.70	4.74	6.14	6.82	0.03

*the variable average is calculated conditional on having a positive value of the associated dummy

2.4 Other researcher characteristics

In our study sample, the average applicant is a promising junior scientist who has been appointed as tenure-track assistant professor. The average applicant age is 34.78 years, with a negligible difference between awarded and non-awarded: 34.41 years in the case of awarded scientists, and 34.85 years for non-awarded. On average, scientists apply 5.62 years after obtaining their PhD. To fulfil the application requirements, the Alfred P. Sloan Foundation asks the candidates to apply within six years of the date they are granted their doctoral degree. Some exceptions, such as a period of parental leave or a change in the research trajectory, are allowed. About one-third of our sample (32% of the cases) claim such exceptions.

One-third of our applicants are female scientists. Interestingly, it seems that female scientists have slightly higher chances of being awarded than their male colleagues: 39% of scientists in the sub-sample of awardees are females compared to 32% in the non-awarded sample. Half of our applicants obtained their PhD in a top-20 university, and 30% of them are based at a top 20 university at the time of the application[6]. The average applicant has a notable publication record both in terms of quantity and quality: 27.6 publications that receive 8.07 citations per year. On average, each publication lists 8.2 authors. As expected, the selection committee seems to rely on the publication record as selection criteria. Awarded applicants have a higher number of publications: 31.71 compared to 26.81 for the non-awarded applicants. Looking at the number of

[6] To retrieve the list of the top-20 universities we relied on QS World University Rankings. We considered the following universities within the list: Massachusetts Institute, Berkeley University, Harvard University, Stanford University, Northwestern University, the California Institute of Technology, University of California –Los Angeles, Yale University, Austin University, Princeton University, Georgia Institute of Technology, Michigan University, Urbana University, Columbia University, Chapel Hill University, Madison University, University of California – San Diego, and University of Pennsylvania.

citations, awarded applicants received 10.81 yearly citations per paper, while non-awarded received 7.55. In addition to controlling for standard scientific productivity quantity and quality measures, we introduce a measure of specialisation of the applicant using the content of her publications and control for it in the regression. Precisely, we compute *Career specialisation* as the average cosine similarity between all the applicant's publications at the time of the application. The measure varies on a scale [-1, +1] where +1 denotes the highest level of specialisation. Our applicants have an average Career specialisation value of 0.66, with no significant differences between awarded and non-awarded applicants.

Table 2 reports the summary statistics for the full sample and the sub-samples of awarded and non-awarded applications, respectively, while Table 3 summarises the description of all the variables included in our analysis.

Table 2: Summary statistics for the full sample, and the sub-samples of awarded and non-awarded applications

Variable	All (2,494)		Awarded (399)	Non-Awarded (2,095)	
	Average	Sd	Average	Average	t-test
Awarded	0.16	0.37	1	0	.
Applicant's biography					
Age	34.78	2.86	34.41	34.85	0.01
Years since Ph.D. degree	5.62	1.86	5.58	5.63	0.61
Female	0.33	0.47	0.39	0.32	0
Top 20 current university	0.3	0.46	0.49	0.26	0
Top 20 Ph.D. university	0.5	0.5	0.62	0.48	0
Applicant's bibliographic characteristics					
Average yearly citations received per publication	8.07	8	10.81	7.55	0
Average number of co-authors per publication	8.2	9.97	8.08	8.23	0.78
Number of publications	27.6	30.09	31.71	26.81	0
Career specialisation	0.66	0.10	0.66	0.66	0.17
Other application characteristics					
RS length	44.56	20.56	44.45	44.59	0.9
Eligibility exception	0.32	0.47	0.32	0.32	0.92
Field					
Computational & Evolutionary Molecular Biology (CEMB)	0.15	0.36	0.15	0.15	0.94
Chemistry	0.24	0.43	0.28	0.23	0.04
Neuroscience	0.18	0.38	0.2	0.17	0.27
Ocean science	0.09	0.28	0.1	0.08	0.34
Physics	0.35	0.48	0.28	0.36	0
Grant year					
2015	0.21	0.41	0.2	0.21	0.46
2016	0.22	0.41	0.2	0.22	0.26
2017	0.19	0.39	0.2	0.18	0.43
2018	0.19	0.4	0.21	0.19	0.53
2019	0.19	0.39	0.2	0.19	0.59

Table 3: Variables' content description.

Variable	Description
Awarded	Dummy equals one if the applicant is awarded the SRF.
Coherence of the research trajectory	
RS coherent (dummy)	Dummy that equals one if the cosine similarity distance between the research statement text and at least one applicant's article published before the application date overcomes the threshold of 0.85, zero otherwise.
Years elapsed max coherence	Years elapsed between the application time and the year of publication of the closest article to the RS, conditional on having at least one coherent publication.
Alignment with subjects of general interest	
RS aligned (dummy)	Dummy that equals one if the cosine similarity between the research statement text and the closest article published in Nature or Science publications after 1999 is above a threshold of 0.85, zero otherwise.
Years elapsed max alignment	Years elapsed between the application time and the year of publication of the closest article appeared in Nature or Science, conditional on having at least one aligned publication.
Applicant's biography	
Age	Applicant's age.
Years from Ph.D. degree	Years elapsed since the applicant's Ph.D. degree.
Female	Dummy that equals one if the applicant is a female scientist, zero otherwise.
Top 20 current university (dummy)	Dummy that equals one if the applicant's current university of affiliation is a top-20 university, zero otherwise.
Top 20 Ph.D. university (dummy)	Dummy that equals one if the applicant's Ph.D. university is a top-20 university, zero otherwise.
Field dummy variables: Computational & Evolutionary Molecular Biology, Chemistry, Neuroscience, Ocean science, Physics	Five dummy variables that equal one according to the application field of application.
Applicant's bibliographic characteristics	
Average yearly citations received per publication	Average yearly citations received by the applicant's stock of publications until the application year.
Average number of authors per publication	Average number of authors calculated for the applicant's stock of publications until the application year.
Number of publications	Applicant's stock of publications until the application year.
Career specialisation	
Average publication similarity	Average cosine similarity between the applicant's publications before the application
Other application characteristics	
RS length (number of pages)	Number of pages of the applicant's research statement.
Eligibility exception (dummy)	The applicant raised an eligibility exception when applying to avoid the eligibility constraint of the 6 years after the Ph.D.
Funding rounds: Round 2015-2019	Five dummy variables indicating the year of the funding round. If the funding round is in year *t*, it means that the scientist crafted her application in *t-1*.

3. Empirical Strategy and Main Results

3.1 Empirical approach

To analyse the impact of the coherence of the research trajectory and of the alignment with subjects of general interest on the probability of being awarded a SRF's Research Fellowship, we estimate Equation 1 with a Logit model.

*Pr(Being awarded a SRF Research Fellowship)=f(**RS coherent, RS coherent*Years elapsed max coherence, RS aligned, RS aligned*Years elapsed max alignment**, Applicant's biography, Applicant's bibliographic characteristics, Career specialisation, Other application characteristics),*

(Equation 1)

The vector *Applicant's biography* in Equation 1 includes information on age, gender, research field, ranking of the university where the candidate obtained her PhD degree, year of graduation, and ranking of the current affiliation. *Applicant's bibliographic characteristics* consider information about the applicant's publication record (publication quantity and quality and number of co-authors). Finally, the vector *Other application characteristics* includes the page-length of the application package and the candidate's eligibility exception (if any)[7].

3.2 Baseline Results

Table 4 reports the results of estimating Equation 1. Column 1 reports the baseline model including the main independent variables for the *RS coherent with at least one previous publication, Years passed since the most coherent publication* and *RS aligned with at least one N&S publication, Years passed since the most aligned N&S publication,* also controlling for *Career specialisation*, *Grant year* fixed effects and *Field* fixed effects. Column 2 introduces extensive controls about the applicant's biographic and bibliographic characteristics and application characteristics.

[7] To be eligible candidates need to have received their PhD degree at most 6 years before the application. Candidates who received their PhD degree earlier might declare an eligibility exception in case of family duties, change of research trajectories, or sickness.

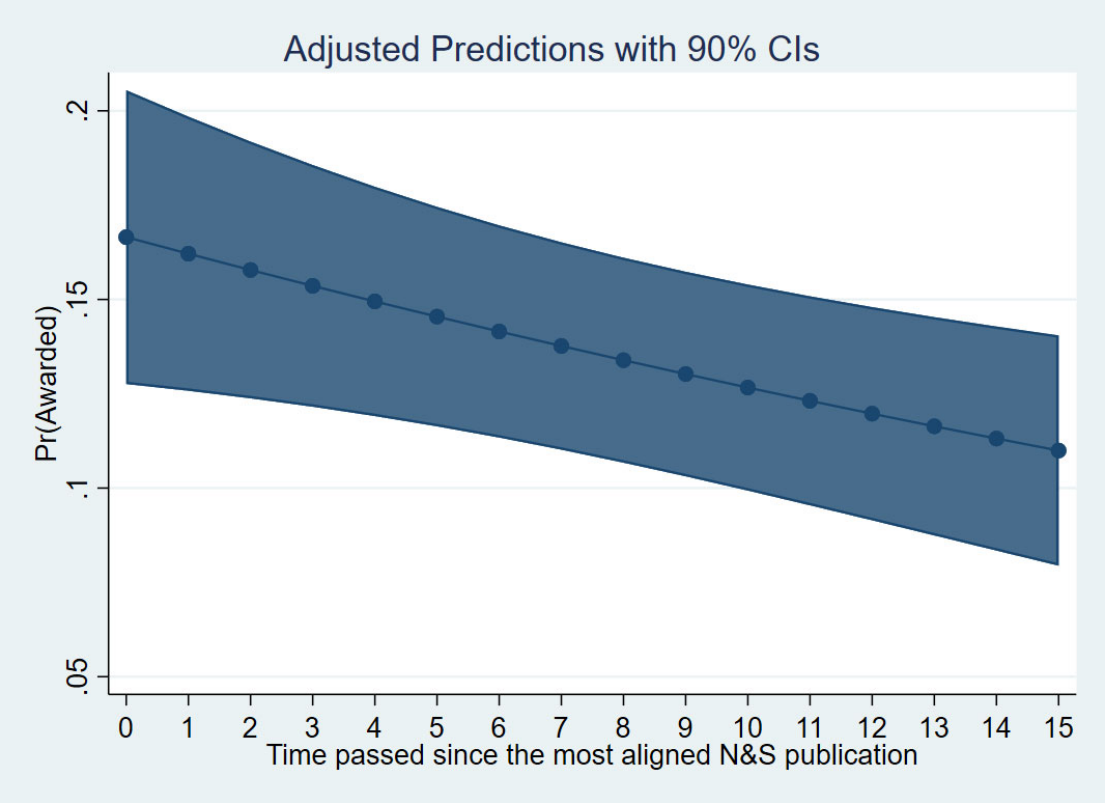
Table 4: Probability of being awarded a SRF Research Fellowship. Logit estimations. Marginal effects reported in the table. Full sample.

	(1) All disc. Awarded	(2) All disc. Awarded
RS coherent	0.053**	0.034
	(0.021)	(0.020)
RS coherent * Years elapsed max coherence	-0.0029	-0.0033
	(0.0040)	(0.0040)
RS aligned	0.095***	0.078***
	(0.022)	(0.021)
RS aligned * Years elapsed max alignment	-0.0047**	-0.0039**
	(0.0019)	(0.0018)
Career specialisation	-0.19**	-0.11
	(0.093)	(0.093)
Age		-0.0091***
		(0.0034)
Years from PhD degree		0.0054
		(0.0056)
Female		0.062***
		(0.015)
Top 20 current university		0.10***
		(0.014)
Top 20 PhD university		0.054***
		(0.015)
Average yearly citations received per publication		0.0043***
		(0.00086)
Average number of authors per publication		-0.00098
		(0.00095)
Number of publications		0.00079***
		(0.00029)
RS length (number of pages)		-0.00018
		(0.00036)
Eligibility exception (dummy)		0.0056
		(0.019)
Observations	2,494	2,494
Dummy grant year	Yes	Yes
Dummy field	Yes	Yes
Pseudo R2	0.0232	0.0899

While the impact of research coherence becomes insignificant when controls are added, having a research statement aligned with at least one article that appeared in Nature or Science increases the probability of being awarded the fellowship. All other things being equal, applicants having a research statement aligned with at least one Nature or Science publication have a 7.8 percentage points higher probability of funding success. The results also show that the temporal dimension counts. For each year passing from the publication of the most aligned Nature or Science article to the year of the application, there is a loss of 0.39 percentage points on the probability of being

awarded. Figure 2 illustrates how the probability of being awarded declines considering a period of 15 years.

Figure 2: Predicted probability of being awarded varying the time passed since the most aligned publication to the research statement. Predictions based on the model estimations reported in Column 2 of Table 4.



Looking at the controls, older applicants are slightly penalised. We observe that women have a 6.2% percentage point higher probability of being awarded, which partly compensates for the initial mismatch in applications between men and women (women represent 33% of all applicants, but the share of women goes up to 39% in the awarded group). As expected, being affiliated with a top-20 university or having obtained a PhD degree from one of those universities increases the probability of being awarded by 10 and 5.4 percentage points, respectively. A strong publication record is well perceived by the evaluation committee. A greater number of publications, as well as receiving more citations, increases the probability of being awarded. Considering the other characteristics of the application, i.e., the length of the proposal or having claimed an eligibility exception, do not significantly affect the probability of being awarded. As one would expect, we observe positive and significant effects of standard

bibliometric measures such as the number of publications and citations on the probability of being awarded. Nevertheless, more surprisingly, we note that the effect of coherence is much larger in scale. More specifically, we see that having a coherent profile increases an applicant's chance as much as 20 more citations per paper per year, or 100 publications, in total, all else being equal. The magnitude of these differences in effect size reflects the importance of the subject choice proposed in the research statement in comparison with the evaluation of the candidate's previous endeavours as reflected by her past scientific publications.

3.3 Exploring research field heterogeneity

The results of Table 4, obtained by pooling together all the applications, show that what matters in the fellowship selection is the alignment of the research statement with subjects of general interest in the discipline. However, one possible concern is that coherence and alignment of the research statement play different roles across disciplines.

Table 5, panel A (*Research trajectory*), shows the average values of the variables measuring the applicants' research statement coherence and alignment with topics of general interest by discipline. Table 5, panel B (*Bibliometric indicators*), reports the average values of two standard bibliometric indicators, i.e., number of publications and number of co-authors.

Table 5: Research trajectory measures and bibliometric indicators by discipline

Panel A - *Research trajectory* Discipline (Number of applications)	CEMB (372)	Chemistry (599)	Physics (871)	Neuroscience (439)	Ocean Sciences (213)
Average					
RS coherent	0.56	0.6	0.81	0.53	0.66
Years elapsed max coherence*	2.58	2.2	2.99	2.8	2.9
RS aligned	0.37	0.61	0.81	0.51	0.64
Years elapsed max alignment*	6.6	6.7	6.32	6.43	9.16

*the variable average is calculated conditional on a positive value of the associated dummy

Panel B – *Bibliometric indicators* Discipline (Number of applications)	CEMB (372)	Chemistry (599)	Physics (871)	Neuroscience (439)	Ocean Sciences (213)
Average					
Average number of authors per publication	8.74	5.72	11.40	5.77	6.19

Number of publications	19.53	27.38	37.18	18.35	22.16

We observe that the research statement coherence and alignment vary across disciplines, as well as the value of bibliometric indicators. Remarkably, physicists emerge as having the highest level of coherence and alignment, and the highest number of authors per paper and publications. Physicists also seem to organise their research activities differently, working in larger teams and producing a greater number of publications. Moreover, physics is the largest discipline in our sample, accounting for 34.92% of our sample.

To explore the effect of these field specificities, we isolate physics and run a separate set of regressions where we distinguish Physics from the other disciplines, i.e., Life Sciences & Chemistry.

Table 6 reports the estimation results. We find that, in Life Sciences & Chemistry, the coherence of the research trajectory, as well as the alignment with subjects of general interest, affect the probability of being awarded. Looking at the temporal dimension, we find that both the time passed since the most coherent article and the time passed since the most aligned article decrease the probability of being awarded. For each year passed, the probability decreases by 1.7 and 0.7 percentage points, respectively. Figure 3 illustrates these trends.

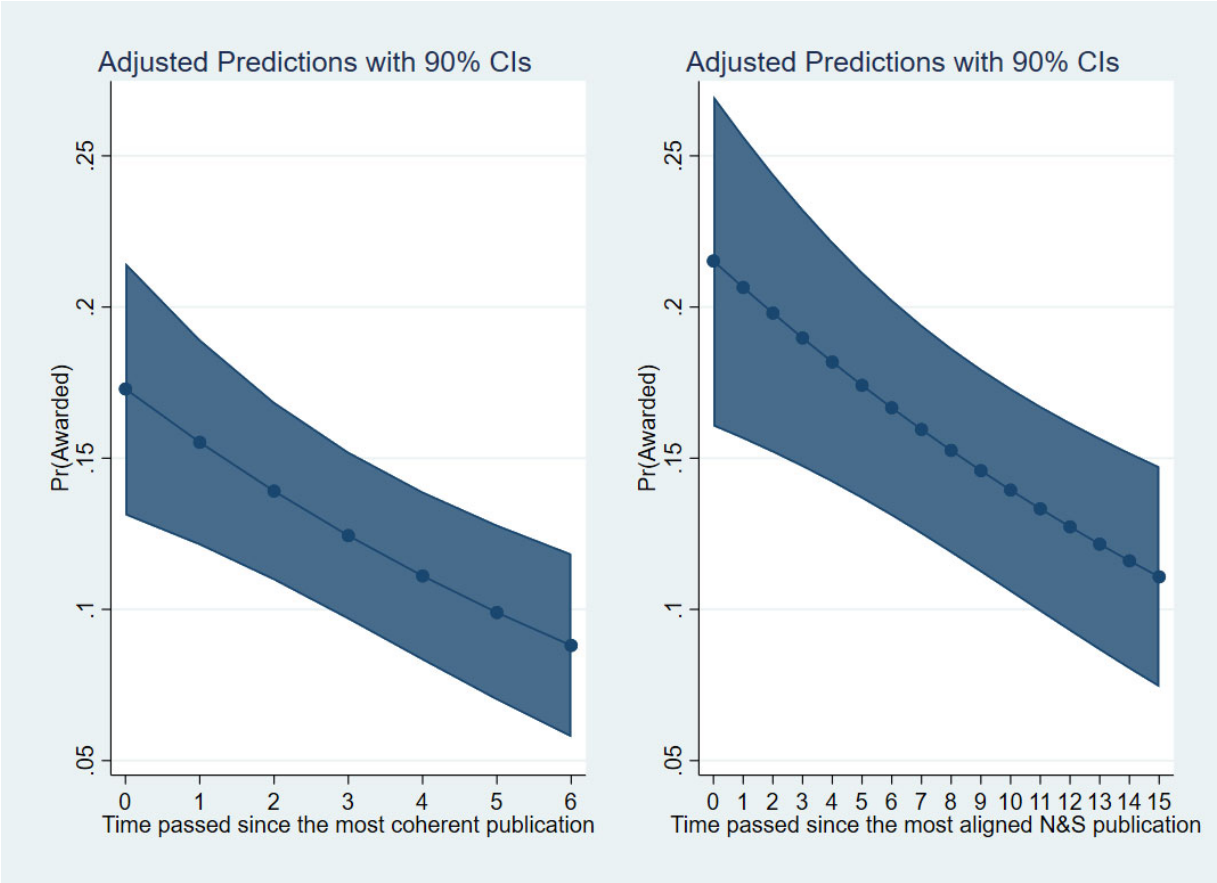
Interestingly, in Physics, the evaluation committee considers only the years passed since the most coherent article with the research statement. Our results show that in Physics, evaluators rely more on bibliometric indicators, i.e., number of publications and citations, than on the coherence or alignment of the applicant's work in the funding decision. Alternative explanations might explain this result. On the one hand, evaluators might pay less attention to the coherence of the content of physicists' publications and research statements, since physicists work in large teams where scientists are highly specialised and where the choice of the research subject is a collegial decision. In this scenario, the research trajectory is not an individual decision, but a team level one. On the other hand, evaluators might pay less attention to the content of physicists' publications since a high number of publications and co-authors make it challenging to assess the applicant's individual contribution to each publication. Moreover, it is also possible that in many areas of physics (e.g. particle physics) researchers do not publish in Science or

Nature, but in more specialised journals, so these journals, while broadly relevant for other disciplines, may not capture publication behaviour in physics.

Table 6: Probability of being awarded a SRF's Research Fellowship in Life Sciences & Chemistry and Physics. Logit estimations. Marginal effects reported in the table.

	(1) Life sciences & Chemistry Awarded	(2) Life sciences & Chemistry Awarded	(3) Physics Awarded	(4) Physics Awarded
RS coherent	0.085***	0.066***	0.022	-0.0073
	(0.025)	(0.025)	(0.043)	(0.043)
RS coherent * Years elapsed max coherence	-0.017***	-0.017***	0.011**	0.011**
	(0.0060)	(0.0058)	(0.0051)	(0.0051)
RS aligned	0.12***	0.10***	0.054	0.024
	(0.026)	(0.026)	(0.043)	(0.044)
RS aligned * Years elapsed max alignment	-0.0079***	-0.0069***	0.00088	0.0015
	(0.0025)	(0.0024)	(0.0027)	(0.0027)
Career specialisation	-0.33***	-0.25**	0.14	0.21
	(0.11)	(0.11)	(0.18)	(0.18)
Age		-0.010**		-0.0047
		(0.0042)		(0.0059)
Years from PhD degree		0.0053		0.0051
		(0.0070)		(0.0093)
Female		0.067***		0.043*
		(0.019)		(0.024)
Top 20 current university		0.12***		0.082***
		(0.018)		(0.023)
Top 20 PhD university		0.071***		0.014
		(0.019)		(0.023)
Average yearly citations received per publication		0.0038***		0.0060***
		(0.0011)		(0.0017)
Average number of co-authors per publication		0.00059		-0.0017*
		(0.0024)		(0.0010)
Number of publications		0.00056		0.00080**
		(0.00038)		(0.00033)
RS length (number of pages)		-0.00092		0.00039
		(0.00057)		(0.00040)
Eligibility exception (dummy)		-0.00072		0.00021
		(0.024)		(0.030)
Observations	1,623	1,623	871	871
Dummy grant year	Yes	Yes	Yes	Yes
Dummy field	Yes	No	Yes	No
Pseudo R2	0.031	0.106	0.0268	0.0896

Figure 3: Predicted probability of being awarded varying the time passed since the most coherent (aligned) publication to the research statement. Based on the model estimations for the subsample of Life Sciences & Chemistry (Column 3 of Table 6).



3.4 Robustness Checks

In this section, we further test the validity of our results by performing three robustness checks. First, in order to account for evaluators' characteristics as a determinant of the evaluation, we control for the how 'intellectually' close evaluators are to the research statement content (Boudreau et al. 2016). In a second robustness check, we test the validity of our main explanatory variables by replacing the binary variables identifying coherent and aligned research statements with two corresponding continuous variables measuring the degree of coherence and alignment. Finally, we check the sensitivity of our results to variations in the similarity threshold values that denote a research statement as coherent or aligned.

Evaluators' intellectual closeness

To calculate the intellectual closeness between the evaluator committee members and the research statement content, we proceed in three steps. First, we gather all the evaluators' publications before the research statement date. Second, we calculate the similarity between each evaluator's publication and the research statement. Finally, if at least one evaluator's publication shows a similarity level above the threshold of 0.85, we define the binary variable *RS evaluators* equals to one, zero otherwise. A positive value of *RS evaluators* means that evaluators are intellectually close to the content of the research statement. For those research statements having the *RS evaluators* equal to one, we calculate the years elapsed since the most similar evaluator's publication to the research statement (*Years elapsed max similarity evaluator*).

We find that facing evaluators who are intellectually close to the content of the research statement, increases the applicant's chances of being awarded – holding constant all the other factors – only in Life Sciences & Chemistry (Table 7, Column 1). When we control for the evaluators' intellectual closeness, our results on the impact of coherence and alignment remain unchanged.

Table 7: Probability of being awarded a SRF Research Fellowship in Life Sciences & Chemistry and Physics, including as controls the similarity of the research proposal to evaluators' publications and the years elapsed since the evaluators' article with the maximum similarity. Logit estimations. Marginal effects reported in the table.

	(1) Life Sciences & Chemistry Awarded	(2) Physics Awarded
RS coherent	0.065***	-0.0074
	(0.024)	(0.043)
RS coherent * Years elapsed max coherence	-0.018***	0.011**
	(0.0057)	(0.0051)
RS aligned	0.075***	0.023
	(0.026)	(0.045)
RS aligned * Years elapsed max alignment	-0.0072***	0.0015
	(0.0024)	(0.0027)
RS evaluators	0.064**	-0.0042
	(0.025)	(0.033)
RS evaluators * Years elapsed max similarity evaluator	0.0028	0.00041
	(0.0017)	(0.0016)
Specialisation	-0.32***	0.21
	(0.11)	(0.19)
Age	-0.011**	-0.0048
	(0.0042)	(0.0060)
Years from Ph.D. degree	0.0076	0.0051
	(0.0070)	(0.0093)
Female	0.072***	0.043*
	(0.019)	(0.024)
Top 20 current university	0.12***	0.082***
	(0.018)	(0.023)
Top 20 Ph.D. university	0.069***	0.014
	(0.018)	(0.023)
Average yearly citations received per publication	0.0037***	0.0059***
	(0.0011)	(0.0017)
Average number of authors per publication	0.00075	-0.0017*
	(0.0024)	(0.0010)
Number of publications	0.00054	0.00080**
	(0.00038)	(0.00033)
RS length (number of pages)	-0.00096*	0.00038
	(0.00057)	(0.00041)
Eligibility exception (dummy)	0.0013	0.000058
	(0.023)	(0.030)
Observations	1,623	871
Dummy grant year	Yes	Yes
Dummy field	Yes	Yes
Pseudo R2	0.119	0.0897

One possible concern is that having a dummy measuring Coherence and Alignment might limit the validity of our results to an assigned threshold. To respond to this concern, first we replace the dummies with the corresponding continuous variables, then we implement a sensitivity analysis considering alternative thresholds.

Coherence and Alignment as continuous variables

We replace the binary variables *RS coherent* and *RS aligned* with the corresponding continuous variables *Max RS coherence* and *Max RS alignment*. *Max RS coherence* is calculated as the maximum similarity score of all the possible scientist's "research statement-previous publication" pairs. Similarly, we define *Max RS alignment* as the maximum similarity score of all the possible scientist's "research statement-Nature & Science publication" pairs. Table 8 reports the descriptive statistics of the two variables.

Table 8: Descriptive statistics for the variables Max RS coherence and Max RS alignment

Discipline (Number of applications)	Life Sciences & Chemistry (1,623)			Physics (871)		
	Mean	Min	Max	Mean	Min	Max
Max RS coherence	0.85	0.13	0.96	0.88	0.08	0.96
Max RS alignment	0.85	0.51	0.96	0.88	0.73	0.95

Table 9 shows the result of the regression exercise using the same model specification as in Table 6 but replacing the binary variables *RS coherent* and *RS aligned* with the continuous variables *Max RS coherence* and *Max RS alignment*.[8]

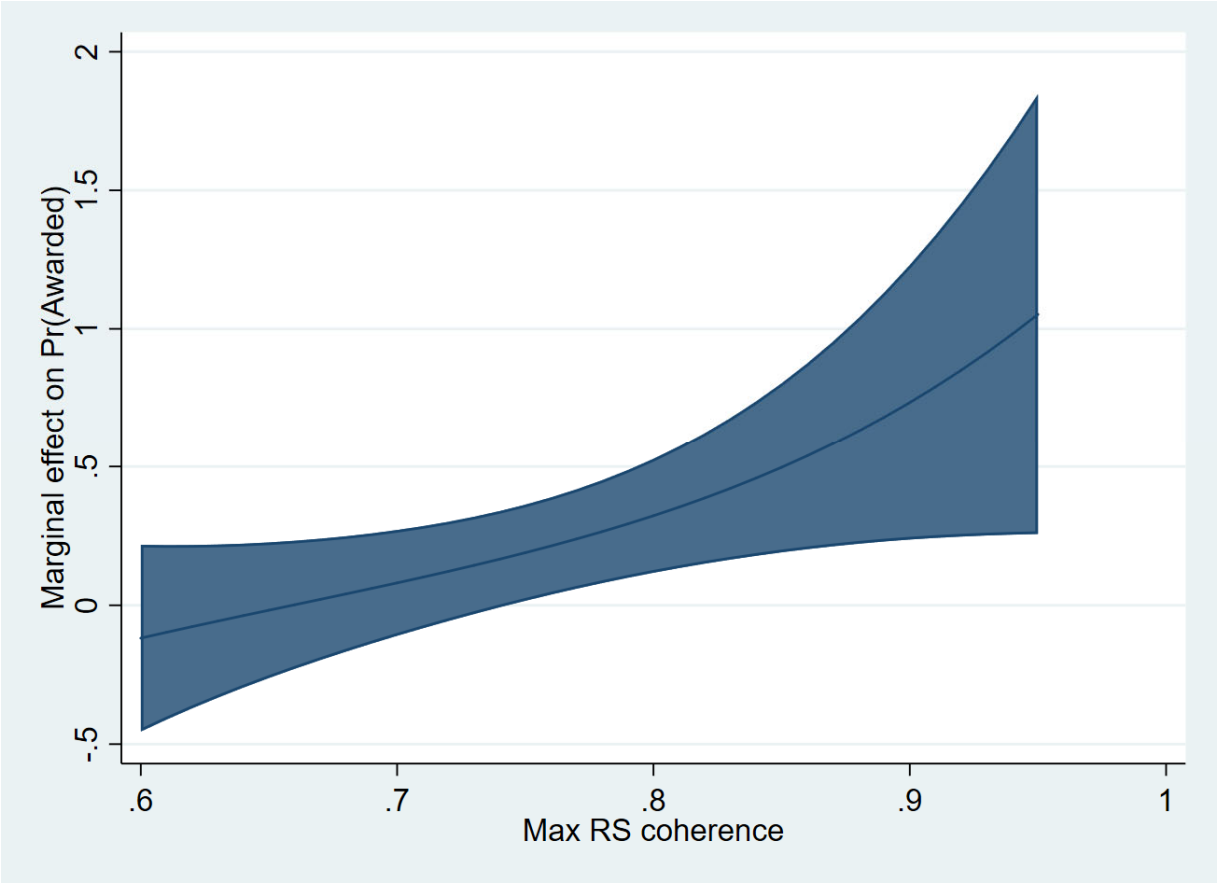
Columns 1 and 2 in Table 9 report the marginal effects of the estimated coefficients of *Max RS coherence* and *Max RS alignment*, while Columns 4 and 5 report the logit coefficients, including the quadratic term of *Max RS coherence* to allow for non-linear effects. According to the results in Columns 1 and 2 of Table 9, the signs of *Max RS alignment* are in line with those reported in Table 6 for the binary version of the variable. Differently from Table 6, the coefficient of *Max RS coherence* is no longer significant for Life Science & Chemistry. The lack of significance of *Max RS coherence* can be explained by the non-linear nature of its impact on the probability of being awarded. Relying on the estimates reported in Column 3, including the quadratic term of *Max RS coherence*, we find a U-shaped effect of *Max RS coherence* that is statistically different from zero for values larger than 0.75 (see Figure 4). For the sake of simplicity, in the main analysis in Table 6, we capture this non-linear effect by defining the binary variable *RS coherence.*

[8] Since the meaning of the variables Years elapsed max coherence and Years elapsed max alignment are meaningless when the values of Max RS coherence and Max RS alignment are low, we excluded these two variables from the regression model.

Table 9: Probability of being awarded a SRF's Research Fellowship in Life Sciences & Chemistry and Physics, including RS coherence and alignment measured as continuous variables. Columns 1 and 2 report marginal effects, while columns 3 and 4 the logit coefficients.

	(1) Life Sciences & Chemistry Awarded	(2) Physics Awarded	(3) Life Sciences & Chemistry Awarded	(4) Physics Awarded
Max RS coherence	0.27	0.19	-14.0*	80.7
	(0.18)	(0.32)	(7.30)	(82.6)
Max RS coherence^2			10.6**	-45.7
			(4.77)	(47.2)
Max RS alignment	0.65**	0.61	4.65**	6.68*
	(0.26)	(0.40)	(1.95)	(4.01)
Specialisation	-0.30***	0.12	-2.65***	1.40
	(0.12)	(0.20)	(0.89)	(1.93)
Age	-0.011***	-0.0045	-0.085***	-0.042
	(0.0042)	(0.0060)	(0.032)	(0.058)
Years from Ph.D. degree	0.0048	0.0060	0.036	0.052
	(0.0070)	(0.0094)	(0.053)	(0.092)
Female	0.065***	0.045*	0.50***	0.45*
	(0.019)	(0.024)	(0.14)	(0.23)
Top 20 current university	0.12***	0.077***	0.90***	0.77***
	(0.018)	(0.023)	(0.14)	(0.22)
Top 20 Ph.D. university	0.075***	0.017	0.56***	0.17
	(0.019)	(0.023)	(0.14)	(0.22)
Average yearly citations received per publication	0.0038***	0.0056***	0.026***	0.054***
	(0.0011)	(0.0017)	(0.0081)	(0.017)
Average number of authors per publication	0.00015	-0.0016	0.0047	-0.015
	(0.0024)	(0.0010)	(0.018)	(0.0099)
Number of publications	0.00045	0.00080**	0.0027	0.0080**
	(0.00038)	(0.00033)	(0.0027)	(0.0032)
RS length (number of pages)	-0.00091	0.00050	-0.0069	0.0049
	(0.00058)	(0.00040)	(0.0044)	(0.0039)
Eligibility exception (dummy)	0.0015	0.0054	-0.0038	0.063
	(0.024)	(0.029)	(0.18)	(0.29)
Constant			2.05	-44.8
			(3.34)	(36.8)
Observations	1,623	871	1,623	871
Dummy round	Yes	Yes	Yes	Yes
Dummy field	Yes	Yes	Yes	Yes
Pseudo R2	0.106	0.0938	0.109	0.0957

Figure 4: Marginal effect of the variable Max RS coherence on Pr(Awarded) for Life Science & Chemistry. The marginal effect is calculated according to the estimates reported in Table 9, Column 3.



Sensitivity to the threshold chosen to define coherence and alignment

We test the sensitivity of our results for different values of the threshold used to define coherent and aligned research statements. Specifically, we consider a high threshold equal to 0.88 and a low threshold equal to 0.82. These two values are obtained by adding and subtracting 0.03 to the threshold of 0.85. The threshold is calculated in Appendix B as the average similarity value of 100 randomly drawn highly-similar publication pairs. The value 0.03 corresponds to half of the standard deviation of the similarity scores of the 100 highly-similar publication pairs. In case of a high threshold, 47.2% of the research statements are defined as coherent (37.4% in Life Sciences & Chemistry and 65.3% in Physics), while 38.1% are defined as aligned (27.7% in Life Sciences & Chemistry and 57.4% in Physics). In case of a low threshold, 86% of the research statements

are defined as coherent (84.8% in Life Sciences & Chemistry and 88.3% in Physics) while 80.5% are defined as aligned (74.5% in Life Sciences & Chemistry and 91.7% in Physics).

Table 11 reports the results of our analysis for a high and low threshold. Columns 1 shows for Life Sciences & Chemistry, when we adopt a looser definition of coherence and alignment setting a low threshold, the coefficients of the variable *RS aligned* and of the interaction *RS aligned * Years elapsed max alignment* are less significant. This result is expected since the high share of research statements classified as aligned (74.5%) reduces the discriminating power of the dummy to identify research statements that are actually similar to Nature and Science articles. On the contrary, *RS coherent* and *RS coherent * Years elapsed max alignment* maintain the same sign and significance as the results in Table 6. When we adopt a stricter definition of coherence and alignment in Column 3, i.e., a high threshold, coherence and alignment maintain their significance as in Table 6. Physics, Column 2 and 4, shows the positive and significant effect of the time elapsed since the most coherent article for coherent research statements as in Table 6.

Table 11: Probability of being awarded a SRF's Research Fellowship in Life Sciences & Chemistry and Physics changing the threshold used to define coherent and aligned research statements.

	Low threshold (0.82)		High threshold (0.88)	
	(1)	(2)	(3)	(4)
	Life Sciences & Chemistry Awarded	Physics Awarded	Life Sciences & Chemistry Awarded	Physics Awarded
RS coherent	0.10***	-0.022	0.074***	0.0063
	(0.028)	(0.056)	(0.026)	(0.035)
RS coherent * Years elapsed max coherence	-0.014***	0.011**	-0.018**	0.013**
	(0.0051)	(0.0048)	(0.0077)	(0.0055)
RS aligned	0.051*	0.063	0.11***	0.024
	(0.028)	(0.063)	(0.029)	(0.034)
RS aligned * Years elapsed max alignment	-0.0035	0.0010	-0.0090***	0.00037
	(0.0021)	(0.0026)	(0.0033)	(0.0031)
Specialisation	-0.23**	0.25	-0.23**	0.12
	(0.11)	(0.18)	(0.11)	(0.19)
Age	-0.0099**	-0.0051	-0.011***	-0.0047
	(0.0042)	(0.0059)	(0.0043)	(0.0059)
Years from Ph.D. degree	0.0052	0.0045	0.0055	0.0051
	(0.0070)	(0.0093)	(0.0071)	(0.0093)
Female	0.066***	0.043*	0.068***	0.044*
	(0.019)	(0.024)	(0.019)	(0.024)
Top 20 current university	0.12***	0.082***	0.12***	0.077***
	(0.018)	(0.023)	(0.018)	(0.023)
Top 20 Ph.D. university	0.073***	0.015	0.074***	0.011
	(0.019)	(0.023)	(0.018)	(0.023)
Average yearly citations received per publication	0.0039***	0.0063***	0.0037***	0.0059***
	(0.0011)	(0.0016)	(0.0010)	(0.0017)
Average number of co-authors per publication	0.00030	-0.0018*	0.0010	-0.0016
	(0.0024)	(0.00099)	(0.0024)	(0.00100)
Number of publications	0.00053	0.00086***	0.00057	0.00075**
	(0.00036)	(0.00032)	(0.00038)	(0.00032)
RS length (number of pages)	-0.0010*	0.00040	-0.0010*	0.00040
	(0.00057)	(0.00040)	(0.00057)	(0.00041)
Eligibility exception (dummy)	-0.00067	-0.00021	0.0010	0.0032
	(0.024)	(0.029)	(0.024)	(0.029)
Observations	1,623	871	1,623	871
Dummy grant year	Yes	Yes	Yes	Yes
Dummy field	Yes	No	Yes	No
Pseudo R2	0.102	0.090	0.106	0.094

4. Discussion and conclusion

This paper examines the role of an individual scientist's research trajectory on the probability of being awarded a prestigious fellowship. We conducted our analysis in the context of the Sloan Research Fellowship program, which awards promising young researchers to support their early careers. The setting provides us the unique opportunity to access detailed information on the candidate's profile as well as her full research statement.

4.1 Results

Our results suggest that the determinants of selection vary substantially across disciplines. In this respect, we consider two sets of disciplines: Life Sciences & Chemistry on the one hand and Physics on the other. In Life Sciences & Chemistry, the coherence of the research trajectory and the alignment with articles published in major generalist scientific journals are the main factors of evaluation. More specifically, we observed that having a coherent research trajectory (i.e. a research statement highly similar to at least one past publication) and being aligned with a Nature or Science publication increases the candidate's chances of being awarded the grant by 6.6 and 10 percentage points respectively, all else being equal. Interestingly, the positive effect of coherence and alignment in Life Sciences & Chemistry is not driven by a preference for more specialised profiles as career specialisation (average similarity among past publications of the applicant) is discounted by the selection committee. Furthermore, in Life Sciences & Chemistry, bibliometric measures such as the number of publications and the number of citations received have a smaller effect. On the other hand, in Physics, the coherence of the research trajectory does not significantly affect the funding chances of applicants. In fact, a specificity of the field of Physics is the fact that the resume of the applicant (i.e. past publications, citations, and quality of the institution) is the main factor driving the evaluation committee's decision.

4.2 Interpretation

Our findings might be driven by several possible mechanisms. Regarding the results on the coherence of the research trajectory, the similarity of a candidate's research statement with her previous publications denotes prior knowledge of the subject submitted in the proposal and might suggest higher chances of successfully implementing the proposed project. The reduced uncertainty in the realisation of the project would then explain the positive relationship between

coherence and the probability of being awarded in Life Sciences & Chemistry. Interestingly, for those scientists working in fields dominated by large labs like Physics where it is challenging to attribute individual contribution, and the choice of research subject tends to be a collegial decision, the detailed content of the research statement and its alignment with previous endeavours hold less importance in the selection decision. In fact, in Physics, it is rather the bibliographic profile of the applicant – number of publications and citations received – that functions as the key aspect in the funding decision.

Concerning the appreciation of research statements highly similar to an article published in Nature or Science can reflect two different phenomena. A first interpretation is that articles that make it into one of these two top journals deal with a subject considered as very relevant for the entire scientific community with strong implications for the advancement of science[9]. It is then logical for the evaluation committee to appreciate proposals aiming to work on subjects with high relevance for the scientific community, with obsolescence of this relevance as time passes. Beyond the mere relevance of the topic, an article published in a top generalist journal also embeds the fashion and trends in the scientific community. Hence, a second explanation of the positive effect of alignment on funding could be the fact that it reflects the "hotness" of a topic (Wei et al., 2013) and is therefore financially encouraged as so.

Furthermore, we observe across all fields that holding constant of all other characteristics (applicant profile and research trajectory), the prestige of the institutions are strong determinants of the selection decision. This last result can be driven by mere prestige being interpreted as a signal of quality (McGuiness, 2003), or by applicants from top institutions having more influential networks (Clauset et al., 2015; Chevalier and Conlon, 2003).

4.3 Contribution and relationship to literature

This paper seeks to contribute to the field of the science of science, an emerging, multidisciplinary field focused on identifying the drivers of science, its rate and direction, and developing policies to accelerate scientific progress (Fortunato et al., 2018). The emergence of the field is driven by data availability (such as Scopus, PubMed, Google Scholar, Microsoft

[9] Both journals underline the relevance of the subject for the scientific community as a factor of publication in the journal: https://www.nature.com/nature/about
https://www.sciencemag.org/about/mission-and-scope

Academic) about scientists and their outputs, and new computational capabilities driven by collaborations between natural, computational, and social scientists (Fortunato et al., 2018). While the large majority of the existing studies explore the effect of funding on science (Jacob and Lefgren, 2011; Ganguli, 2017; Azoulay et al., 2018; Ayoubi et al., 2019), we investigate the factors that lead to funding success, in order to understand the antecedents of funding. We do that looking at young researchers since early successes starkly increase future success chances in securing research funding (Bol, de Vaan and van de Rijt 2018). With the rising concern on the growing importance of bibliometric measures in evaluating scientific impact (Stephan et al. 2017), we bring evidence on the key place still being taken by the content of applicant research proposal and the effect of research subjects choices on the probability of being awarded. Our paper is not the first to explore research trajectories as a core feature of scientists' careers, but most works on the matter have thus far been mainly descriptive (Franzoni et al., 2009; Gläser and Laudel, 2009) and use citation patterns to identify "research trails" (Gläser, 2012).

In addition to the science of science literature, we contribute to other streams of research exploring the determinants of success in competitive selection processes such as venture capital investment and recruitment procedures. The process of selection when choosing among several potential candidate firms to fund is similar to the funding procedure in science as candidates seeking funds (i.e. entrepreneurs) submit a detailed description of their future lines of work (i.e. a business plan) (Boudreau et al; 2016). Venture capitalists are asked to select the most promising project to put money in (Baum and Silverman, 2004). Scholars have identified two main factors affecting the selection: the characteristics of the project presented, on the one hand, and the leading proponent and her past experiences, on the other (MacMillan et al., 1985). However, the empirical findings of this literature have not exhibited convergent results, with some putting forward the importance of the proponent and her previous experience (MacMillan et al., 1987; Clarysse et al., 2005) while others finding that the project presented is the key factor to make the cut (Tyebjee and Bruno, 1984; Sudek et al. 2008). Being based mostly on survey answers given by venture capital investors, these findings can be affected by the subjectivity in the answers of the survey participants and are limited by the binarity of the answering options. Our approach allows us to assess the key factors of success in being funded with objective measures. First, while we find that the characteristics of proponent and project alone matter, our results on coherence bring empirical evidence to the hypothesis of MacMillan et al. (1985) suggesting that

the most important is probably whether the "jockey is fit to ride", i.e., if the project is coherent with the past experience of the proponent. Second, the diversity of the fields in our data suggests that one should expect some heterogeneity in the selection process among sectors. In other words, as the difference in results we find between Physics and other fields suggests, it is very likely that the process of selection for venture capital investors would be different depending on the inherent characteristics of the business sector. Finally, the importance of alignment that we observe infers that the accordance of the business plan with global business trends might also be a key factor of selection.

In the context of firms seeking new employees, the hiring process of firms is often based on the evaluation of the previous career achievements of the candidate and her profile match with the firm's current and future projects (Acharya and Wee, 2019). Extant literature on recruitment determinants has questioned the relevance of previous job experiences on the probability of being hired. The works of Zuckerman (1999) and Leung (2014) have shown that building a coherent identity in past experiences increases the chances of being selected. Our findings bring more accurate insights showing that coherence and alignment with current trends matter and that one can expect a high variability across sectors. Furthermore, with respect to the hiring literature that uses a broad job classification, we contribute by highlighting the impact of the actual content of work (i.e. scientists' research agendas) on funding success in the labour market.

4.4 Policy implications and future research

Our findings have important policy implications, suggesting to scientists the most rewarding choices when developing their future research plans. The positive impact of having a coherent research trajectory suggests that the trajectories rewarded are those in which future knowledge incrementally builds upon existing knowledge while "radical jumps" are penalised.

Moreover, the preference of the evaluation committee for topics of general interest for the scientific community can be seen as the propensity of the funding agency to direct funds towards matters relevant to the scientific community with previous proof of success and avoiding niche projects with excessive uncertainty. However, one might also see it as a confirmation of the claim of Nicholson and Ioannidis (2012) that funding in science follows the rule of "Conform and be funded" and is probably missing out on potentially more impactful projects. As

researchers often pursue their research projects regardless of whether they received the funds for it (Ayoubi et al. 2019), future research could investigate whether non-funded projects work on more impactful ideas.

Our focus on the Sloan Research Fellowships is partly motivated by the fact that it targets promising early-career scientists[10], who are still in the process of developing a scientific identity. Our motivation in studying these scholars is that we are interested in understanding the incentives given to these future top researchers in terms of subject selection in the funding process. Specifically, does the funding process encourage them to stick to a set of research subjects in which they have already shown some productivity, or to explore topics in which they have little to no expertise? Does it stimulate them to study topics that are aligned with already popular subjects in the field, or to delve into unexplored research questions? We bring first empirical evidence on how the funding process can be favouring certain types of scientific issues and specific research trajectories. However, basing our analysis on planned projects, it remains rather unsure whether the reception of funds does effectively stir the direction of scientific research and if so, to what extent. These are interesting questions to be explored in future research.

[10] The outstanding quality of awarded fellows can be seen in the recognition they receive later in their career with 43 fellows winning a Nobel Prize (https://web.archive.org/web/20160127182945/http://www.sloan.org/sloan-research-fellowships/nobel-laureates/) and 16 winning the Fields Medal in mathematics (https://web.archive.org/web/20120908235152/http://www.sloan.org/sloan-research-fellowships/fields-medalists/).

5. References

Acharya, S., & Wee, S. L. (2019). Rational inattention in hiring decisions. *FRB of New York Staff Report*, (878).

Astebro T, Elhedhli S (2006) The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*. 52(3):395–409.

Ayoubi, C., Pezzoni, M., & Visentin, F. (2019). The important thing is not to win, it is to take part: What if scientists benefit from participating in research grant competitions?. *Research Policy*, *48*(1), 84-97.

Azoulay, P., Graff Zivin, J. S., Li, D., & Sampat, B. N. (2018). Public R&D investments and private-sector patenting: evidence from NIH funding rules. The Review of Economic Studies, 86(1), 117-152.

Arora, Ashish, Gambardella, Alfonso, 2005. The impact of NSF support for basic research in economics. Annales d'Economie et de Statistique 79 (80), 91–117.

Baron, J. N., & Hannan, M. T. (2002). Organizational blueprints for success in high-tech start-ups: Lessons from the Stanford project on emerging companies. *California Management Review*, *44*(3), 8-36.

Baum, J. A., & Silverman, B. S. (2004). Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology start-ups. *Journal of business venturing*, *19*(3), 411-436.

Berlin, I. (2013). *The hedgehog and the fox: An essay on Tolstoy's view of history*. Princeton University Press.

Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. "Are ideas getting harder to find?" (No. w23782). *National Bureau of Economic Research*. (2017)

Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225-1234.

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, *115*(19), 4887-4890.

Bornmann, L., Mutz, R., & Daniel, H. D. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226-238.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, *62*(10), 2765-2783.

Burton, M. D., & Beckman, C. M. (2007). Leaving a legacy: Position imprints and successor turnover in young firms. *American Sociological Review*, *72*(2), 239-266.

Chan, Y. S. (1983). On the positive role of financial intermediation in allocation of venture capital in a market with imperfect information. *The Journal of Finance*, *38*(5), 1543-1568.

Chevalier, A., & Conlon, G. (2003). Does it pay to attend a prestigious university?.

Clarysse, B., Knockaert, M., & Lockett, A. (2005). How do early stage high technology investors select their investments. *Venture Capital*.

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science advances*, *1*(1), e1400005.

Dahl, M. S., & Klepper, S. (2015). Whom do new firms hire?. *Industrial and corporate change*, *24*(4), 819-836.

Etzkowitz, H. (2003). Research groups as 'quasi-firms': the invention of the entrepreneurial university. *Research policy*, *32*(1), 109-121.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Vespignani, A. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Franzoni, C., Simpkins, C., Li, B., & Ram, A. (2009). Using content analysis to investigate the research paths chosen by scientists over time. *Scientometrics*, *83*(1), 321-335.

Franzoni, C., & Rossi-Lamastra, C. (2017). Academic tenure, risk-taking and the diversification of scientific research. *Industry and Innovation*, *24*(7), 691-712.

Harrison, R. T., & Mason, C. M. (2017). Backing the horse or the jockey? Due diligence, agency costs, information and the evaluation of risk by business angel investors. *International Review of Entrepreneurship*, *15*(3), 269-290.

Gibbons, M. (Ed.). (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. Sage.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science*, *333*(6045), 1015-1019.

Gläser, J., & Laudel, G. (2009). Identifying individual research trails. In *Proceedings of ISSI* (pp. 14-17).

Gläser, J. (2012). How does Governance change research content? On the possibility of a sociological middle-range theory linking science policy studies to the sociology of scientific knowledge. Technical University Berlin. *Technology Studies Working Papers*

Gompers, P. A. (1995). Optimal investment, monitoring, and the staging of venture capital. *The journal of finance*, *50*(5), 1461-1489.

Groysberg, B., & Lee, L. E. (2009). Hiring stars and their colleagues: Exploration and exploitation in professional service firms. *Organization science*, *20*(4), 740-758.

Gush, J., Jaffe, A., Larsen, V., & Laws, A. (2018). The effect of public funding on research output: The New Zealand Marsden Fund. *New Zealand Economic Papers*, *52*(2), 227-248.

Hall, J., & Hofer, C. W. (1993). Venture capitalists' decision criteria in new venture evaluation. *Journal of business venturing*, *8*(1), 25-42.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?. *The Review of Economic Studies*, *76*(1), 283-317.

Kambourov, G., & Manovskii, I. (2008). Rising occupational and industry mobility in the United States: 1968–97. *International Economic Review*, *49*(1), 41-79.

Kaplan, S. N., Sensoy, B. A., & Strömberg, P. (2009). Should investors bet on the jockey or the horse? Evidence from the evolution of firms from early business plans to public companies. *The Journal of Finance*, *64*(1), 75-115.

Kaplan, S. N., Klebanov, M. M., & Sorensen, M. (2012). Which CEO characteristics and abilities matter?. *The Journal of Finance*, *67*(3), 973-1007.

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

Leahey, E. (2007). Not by productivity alone: How visibility and specialization contribute to academic earnings. *American sociological review*, *72*(4), 533-561.

Leung, M. D. (2014). Dilettante or renaissance person? How the order of job experiences affects hiring in an external labor market. *American Sociological Review*, *79*(1), 136-158.

Leung, M. (2016). Failed searches: Hiring as a cognitive decision making process and how applicant variety affects an employer's likelihood of making an offer. *Available at SSRN 2833689*.

Lungeanu, R., & Zajac, E. J. (2016). Venture capital ownership as a contingent resource: how owner–firm fit influences IPO outcomes. *Academy of Management Journal*, *59*(3), 930-955.

MacMillan, I. C., Siegel, R., & Narasimha, P. S. (1985). Criteria used by venture capitalists to evaluate new venture proposals. *Journal of Business venturing*, *1*(1), 119-128.

MacMillan, I. C., Zemann, L., & Subbanarasimha, P. N. (1987). Criteria distinguishing successful from unsuccessful ventures in the venture screening process. *Journal of business venturing*, *2*(2), 123-137.

Mairesse, J., & Pezzoni, M. (2015). Does gender affect scientific productivity?. *Revue économique*, *66*(1), 65-113.

Mairesse, J., Pezzoni, M., & Visentin, F. (2019). Impact of family characteristics on the gender publication gap: evidence for physicists in France. *Interdisciplinary Science Reviews*, *44*(2), 204-220.

McGuinness, S. (2003). University quality and labour market outcomes. *Applied Economics*, *35*(18), 1943-1955.

Merluzzi, J., & Phillips, D. J. (2016). The specialist discount: Negative returns for MBAs with focused profiles in investment banking. *Administrative Science Quarterly*, *61*(1), 87-124.

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

Mitteness, C. R., Baucus, M. S., & Sudek, R. (2012). Horse vs. jockey? How stage of funding process and industry experience affect the evaluations of angel investors. *Venture Capital*, *14*(4), 241-267.

Negro, G., Hannan, M. T., & Rao, H. (2011). Category reinterpretation and defection: Modernism and tradition in Italian winemaking. *Organization Science*, *22*(6), 1449-1463.

Noe, R. A., Hollenbeck, J. R., Gerhart, B., & Wright, P. M. (2017). *Human resource management: Gaining a competitive advantage*. New York, NY: McGraw-Hill Education.

Oswald, A. J., & Stern, N. (2019). Why does the economics of climate change matter so much, and why has the engagement of economists been so weak?. *Royal Economic Society Newsletter, October*.

Page, S. E. (2006). Path dependence. *Quarterly Journal of Political Science*, *1*(1), 87-115.

Pontikes, E. G. (2008). *Fitting in or starting new? An analysis of invention, constraint, and the emergence of new categories in the software industry*. Stanford University.

Rosenbaum, J. E. (1979). Tournament mobility: Career patterns in a corporation. *Administrative science quarterly*, 220-241.

Ruef M, Patterson K. 2009. Credit and classification: the impact of industry boundaries in 19th century America. *Admin. Sci. Q.* 54:486–520

Schmutte, I. M. (2014). Job referral networks and the determination of earnings in local labor markets. *Journal of Labor Economics*, *33*(1), 1-32.

Scott, E. L., Shu, P., & Lubynsky, R. M. (2015). *Are'better'Ideas More Likely to Succeed?: An Empirical Analysis of Startup Evaluation* (No. 16-013). Harvard Business School.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A. L. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312), aaf5239.

Spivey, C. (2005). Time off at what price? The effects of career interruptions on earnings. *ILR Review*, *59*(1), 119-140.

Stijepic, D. (2018). Trends and Cycles in US Job Mobility. *Available at SSRN 3260290*.

Sudek, R., Mitteness, C. R., & Baucus, M. S. (2008, August). Betting On The Horse Or The Jockey: The Impact Of Expertise On Angel Investing. In *Academy of Management Proceedings* (Vol. 2008, No. 1, pp. 1-6). Briarcliff Manor, NY 10510: Academy of Management.

Tassier, T., & Menczer, F. (2008). Social network structure, segregation, and equality in a labor market with referral hiring. *Journal of Economic Behavior & Organization*, *66*(3-4), 514-528.

Tetlock, P. E. (2017). *Expert Political Judgment: How Good Is It? How Can We Know?-New Edition*. Princeton University Press.

Tirole, J. (2017). *Economics for the common good*. Princeton University Press.

Topel, R. H., & Ward, M. P. (1992). Job mobility and the careers of young men. *The Quarterly Journal of Economics*, *107*(2), 439-479.

Tyebjee, T. T., & Bruno, A. V. (1984). A model of venture capitalist investment activity. *Management science*, *30*(9), 1051-1066.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G. and Jain, A., 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), p.95.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416-1436.

Wei, T., Li, M., Wu, C., Yan, X. Y., Fan, Y., Di, Z., & Wu, J. (2013). Do scientists trace hot topics?. *Scientific reports*, *3*, 2207.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036-1039.

Zhang, J. (2011). The advantage of experienced start-up founders in venture capital acquisition: evidence from serial entrepreneurs. *Small Business Economics*, *36*(2), 187-208.

Zuckerman, E. W. (1999). The categorical imperative: Securities analysts and the illegitimacy discount. *American journal of sociology*, *104*(5), 1398-1438.

Zuckerman, E. W., Kim, T. Y., Ukanwa, K., & Von Rittmann, J. (2003). Robust identities or nonentities? Typecasting in the feature-film labor market. *American Journal of Sociology*, *108*(5), 1018-1074.

Appendix

A. Representing documents with vectors

For evaluating the degree of similarity between two documents, we need to transform the documents into vectors so that we can compute the cosine similarity of the two resulting vectors. To produce the vector representation of documents, we proceed in two steps: First, we generate the vector representation of a vocabulary of words, then we use this global representation to represent each document by a unique vector.

For the first step, in order to produce the vector representation of a full vocabulary of words, we rely on the Word2vec algorithm for text analysis proposed by Mikolov et al. (2013). Word2vec is a neural network-based approach generating a vector representation of a word based on the word's context within a large corpus of documents. The logic behind Mikolov et al.'s algorithm is that words sharing common contexts end up close to one another in the vector space. Precisely, Word2vec works on predicting a word based on the words surrounding it (Continuous-Bag-Of-Words or *CBOW* method) or by predicting the missing words surrounding a certain word (Skip-gram method). For instance, if the sequence analysed is "New scientific discoveries are great" and the window is two words, the *Skip-gram* method works on predicting the four missing words in "__ __ discoveries __ __" (often called *negative sampling*) while the CBOW method tries to predict the missing word in "New scientific __ are great". Following the recent works on text analysis (Tshitoyan et al. 2019) we use the Skip-gram method in our analysis.
The algorithm performs the prediction by training its estimation on a large corpus of texts (often called the training dataset) and readjusting the predicted values based on the words' apparitions. Specifically, Word2vec produces its prediction by constructing a vector representation of words in a vector space of an arbitrary number of dimensions *N*. Adopting Mikolov et al.' s terminology, the vector space where words are represented is called the *hidden layer*. The *hidden layer* is unobservable, while the *input layer* and the *output layer* are used to estimate it (see Figure A1). According to the *Skip-Gram* model estimated using *negative sampling* (see Rong, 2014 for a detailed description), the *target word*, i.e., the word selected in the text, is represented in the *input layer* as a vector having only one unit that equals one (the one corresponding to the *target word*) and all the other units equal zero (the ones corresponding to all the other *V-1* words in the vocabulary). The *output layer* are the *C* vectors of size *V* representing the *C context words* appearing in a window of size *C* centred on the target word (see Figure A2 for a representation of the *Skip-Gram* model with a window of size *C*). We parametrised our algorithm setting the number of dimensions *N* equal to 100 and the window *C* used to identify the context words equal to 10. To train the algorithm and obtain a reliable estimation of the vector representation of the *V* words in the vocabulary we use all the words (and the corresponding context words) appearing in all the article abstracts published in two leading generalist journals, Nature and Science, from 2000 to 2017. We obtained a corpus of 28,872 abstracts, including a vocabulary of 35,993 words (*V*). We end up with a matrix of size *VxN* that corresponds to the vector representation of a vocabulary of words.

For the second step, the goal is to transform each document into a vector. We therefore extract from the text of the document the list of words and we drop the most common stop-words such as “the”, “a”, “an”, etc. We end up with a list of words of length *L* representing the words appearing in the document. Then, we assign to each word its vector representation derived using the Skip-Gram model described in the first step. After matching the vector representation of the words in the vocabulary with the list of words appearing in the document, each document is represented by a matrix of size *LxN* where *L* are the words appearing in the document and *N* is the size of the vector representing each word. To reduce the document to a unique vector of size *1xN*, we calculate the centroid of all the *L* words which represents the weighted average of all vectors in the *LxN* matrix.

Figure A1: The basic Word2vec model with the three layers neural network with a vocabulary of size V and a hidden layer of dimension N.

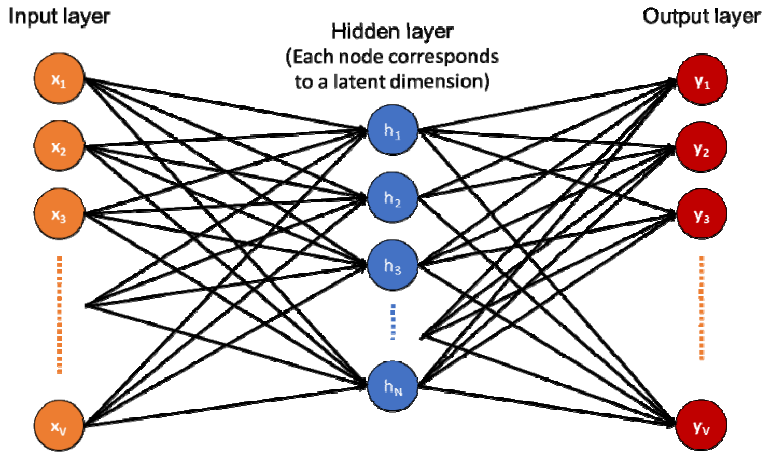
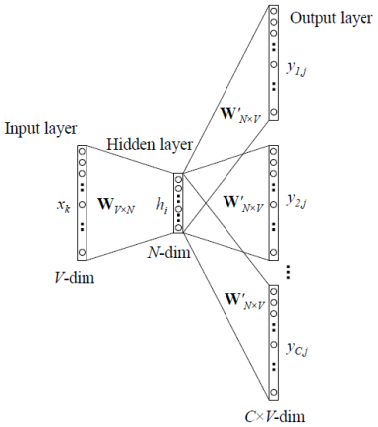


Figure A2: A Skip-gram model with N latent dimensions, a vocabulary of size V and a window of size C.



(Source: Rong et al. 2014)

B. An example of document similarity

To illustrate how we implemented the Word2vec algorithm, we calculate the similarity between three documents. Two documents, Bougher et al. (2015) and Jakosky et al. (2015), reported in the issue 6261 of Science have similar subjects. Specifically, they include a description of the analyses conducted by the Mars Atmosphere and Volatile Evolution (MAVEN) spacecraft being part of the same special issue of the journal on MAVEN. The third document, Soderquist (2015), also published in the same issue of Science (but not in the MAVEN special issue), treats a very different subject: the isolation of the Americium, a radioactive element.

For each article abstract, we calculate the document vector representation by using the Word2vec algorithm, as explained in Appendix A. Then, we calculate the cosine similarity between each pair of articles. The results are reported in Table B1.

Table B1: Similarity between the three selected documents.

	Bougher et al. 2015	Jakosky et al. 2015	Soderquist 2015
Bougher et al. 2015	1.00		
Jakosky et al. 2015	0.86	1.00	
Soderquist 2015	0.22	0.21	1.00

Table B1 shows, as expected, that the value of similarity between the Bougher's and Jakosky's article is high, while the similarity of both articles with the Soderquist is low.

To allow for a graphical representation of the similarity between the three documents in a two-dimensional space, we re-estimated the Word2vec algorithm reducing the size of the vector space from N=100 to N=2. Figure B1 shows the result. The angle α between the dashed lines connecting the origin of the vector space and the point representing the Bougher's and Jakosky's articles is close to 0, leading to a value of cos(α) close to 1. On the contrary, the angle β between the dashed line connecting the origin of the vector space with the Soderquist article and the dashed lines of the Bougher's article is large, leading to a value of cos(β) smaller than cos(α). The value of cos(α) higher than cos(β) shows that Bougher's and Jakosky's articles are more similar than the Soderquist's and Bougher's articles.

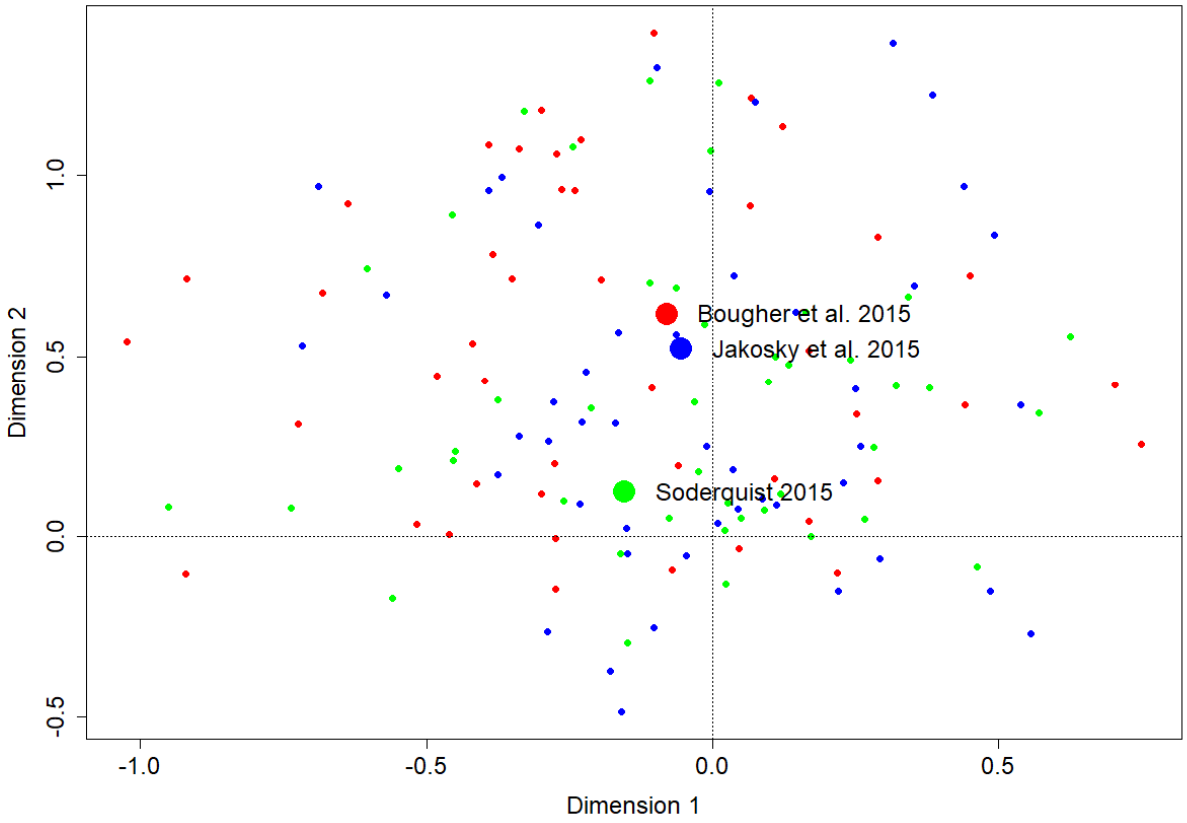
References:

Bougher, S., Jakosky, B., Halekas, J., Grebowsky, J., Luhmann, J., Mahaffy, P., ... & Mcfadden, J. (2015). Early MAVEN Deep Dip campaign reveals thermosphere and ionosphere variability. *Science*, *350*(6261), aad0459.

Jakosky, B. M., Grebowsky, J. M., Luhmann, J. G., Connerney, J., Eparvier, F., Ergun, R., ... & Mitchell, D. L. (2015). MAVEN observations of the response of Mars to an interplanetary coronal mass ejection. *Science*, *350*(6261), aad0210.

Soderquist, C. (2015). How to isolate americium. *Science*, *350*(6261), 635-636.

Figure B1: Representation of three articles in a 2-dimensional space obtained applying the Word2vec algorithm.



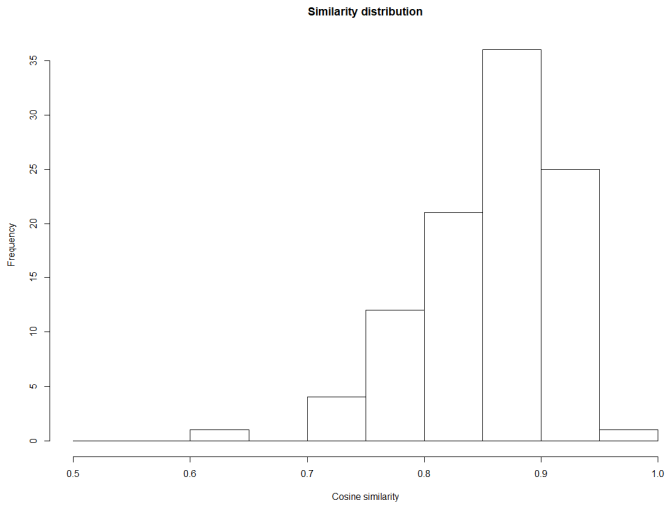
C. Fixing a threshold to define similar/aligned documents

To define a threshold above which we consider two documents as coherent/aligned, we adopt two different approaches that lead to consistent results.

According to the first approach, *Similarity threshold based on selected articles*, we deduct the similarity threshold by comparing two documents for which we have some *a priori* on their level of similarity. Specifically, we select two articles that are likely to be similar since they appeared in the same Science special issue on the analyses conducted by the Mars Atmosphere and Volatile Evolution (MAVEN) spacecraft. As shown in Appendix B, the similarity between two MAVEN articles equals 0.86. According to the first approach, we consider 0.86 as the threshold above which we two articles are similar.

According to the second approach, *Similarity threshold based on 100 randomly drawn articles*, we randomly draw 100 article abstracts, i.e., the core articles, from a large sample of 28,872 scientific articles published in Nature and Science. Then, we calculate the similarity between each core article and the remaining 28,872-1 articles, i.e. the comparison articles, retaining only the pair core-comparison article with the highest similarity score. We end up with 100 similarity score values distributed as shown by Figure B2. Finally, we calculate the average similarity of the 100 article pairs, and we considered it as the threshold above which two articles are similar. We find that the 100 articles' similarity average equals to 0.85 and the standard deviation to 0.06.

Figure C1: Similarity distribution for the 100 randomly drawn articles paired with their most similar article retrieved in Nature and Science publications.



The two approaches lead to similar results identifying a threshold of 0.86 and 0.85, respectively. We decided to adopt the threshold resulting from the statistical exercise conducted in this appendix, i.e., 0.85, in our analyses.

The UNU-MERIT WORKING Paper Series

2020-01 *Debating the assumptions of the Thirlwall Model: A VECM analysis of the Balance of Payments for Argentina, Brazil, Colombia, and Mexico* by Danilo Spinola

2020-02 *The La Marca Model revisited: Structuralist Goodwin cycles with evolutionary supply side and balance of payments constraints* by Danilo Spinola

2020-03 *Uneven development and the balance of payments constrained model: Terms of trade, economic cycles, and productivity catching-up* by Danilo Spinola

2020-04 Time-space dynamics of return and circular migration: Theories and evidence by Amelie F. Constant

2020-05 *Mapping industrial patterns and structural change in exports* by Charlotte Guillard

2020-06 *For real? Income and non-income effects of cash transfers on the demand for food* by Stephan Dietrich and Georg Schmerzeck

2020-07 *Robots and the origin of their labour-saving impact* by Fabio Montobbio, Jacopo Staccioli, Maria Enrica Virgillito and Marco Vivarelli

2020-08 *STI-DUI innovation modes and firm performance in the Indian capital goods industry: Do small firms differ from large ones?* By Nanditha Mathew and George Paily

2020-09 *The impact of automation on inequality across Europe* by Mary Kaltenberg and Neil Foster-McGregor

2020-10 *What matters in funding: The value of research coherence and alignment in evaluators' decisions* by Charles Ayoubi, Sandra Barbosu, Michele Pezzoni and Fabiana Visentin