The Impact of Project-Based Funding in Science: Lessons from the ANR Experience*

Nicolas Carayol[a,b,c] and Marianne Lanoë[a,b]

[a] *GREThA, UMR CNRS 5113, France*
[b] *Université de Bordeaux, France*
[c] *Observatoire des Sciences et Techniques, HCERES, France*

January 27, 2017

Abstract

Competitive allocation of funds to research proposals is a mechanism widely used by government agencies to sustain the projects of researchers in universities and other research institutions. However, little is known about how efficient this mechanism is in practice, how it affects the recipients' behaviors and how it would be possible to improve the precise design of such funding allocation mechanisms. This article provides new answers to those questions, relying on empirical evidence stemming from the creation of a French generalist and nationwide research funding agency in 2005. The impacts of receiving a grant on the research outputs as well as on the collaborations of the grantees is precisely quantified. Moreover, the impact on citations turns out to be more than double when funds are distributed in the more competitive non-thematic programs and to be significantly larger when allocated to younger recipients.

Keywords: project-based funding, competitive grants, scientific productivity, conditional difference-in-differences.

JEL codes: D04, O3, C31.

*Nicolas Carayol is Professor of Economics at the Université de Bordeaux, Research Fellow of the CNRS laboratory GREThA and Scientific Delegate at the Observatoire des Sciences et Techniques of the HCERES in Paris. Email: nicolas.carayol@u-bordeaux.fr. Marianne Lanoë is a PhD student at the Université de Bordeaux and Research Fellow at GREThA (Email: marianne.lanoe@u-bordeaux.fr). We would like to thank the research teams at OST-HCERES and GREThA and, in particular, C. Mouradian, E. de Turkheim and G. Filliatreau. We acknowledge the financial support of the ANR, deliberately very limited to avoid any form of conflict of interest. Special thanks go to Philippe Freyssinet, Michael Matlosz and Anne Portier from the ANR, for their advice, fruitful exchanges, and their complete respect of the independence of the authors, who are the only ones responsible for any remaining errors.

1 Introduction

Governments financially support research carried out in universities and research organizations via various mechanisms. In Europe, governments mostly used to fund (public) institutions via hard money. By contrast, the US has developed a tradition of competitive project-based federal funding since World War II. Some sub-periods have witnessed significant budget increases by the main federal agencies such as the NIH or the NSF (see Stephan, 2013 for a detailed overview). According to the National Science Board (2016), yearly extramural federal funding of US universities and colleges has exceeded 40 billion dollars since 2010. The presumed advantages of that mechanism has led European countries and institutions as well as many other countries around the globe to develop similar policies. Despite the huge amount of public money at stake worldwide and although the way money reaches research presumably affects efficiency, there is still little large-scale systematic evidence about the impact of such fund allocation schemes. Further, competitive allocation of funds to research proposals actually reveals a significant variability. The precise rules and goals of the programs are also likely to affect the outcomes. It is thus important to understand how and why the returns may vary with respect to the specific designs of the funding programs.

This article provides new clues on these issues, relying on the recent French experience. In 2005, the French government created a dedicated agency, the Agence Nationale de la Recherche (ANR) to implement project-based research funding in the country. Our study focuses on the first five years of the ANR's existence (2005-2009) since sufficient post-funding time has now elapsed for some of the first consequences of this policy to be observed. Over this period, this institution allocated nearly two and a half billion euros to research projects, the total cost of which amounts to approximately ten billion euros. The ANR was set up as a nationwide generalist player welcoming applications from all disciplines. Behind that single desk, a series of distinct programs has been processed. Though all of them do share the common goal of supporting the research and excellence of the recipients, they vary according to some specific rules and in their very focuses.

This experience therefore offers an excellent opportunity for investigating the impact of fund allocation on a large scale. It also provides interesting forms of variation as regards recipients, programs, disciplines, etc. The previous literature has been limited either to specific fields (bio-medical sciences in Carter et al., 1987, Arora et al., 2000, Jacob and Lefgren, 2011 and Azoulay et al, 2011; economics for Arora and Gambardella, 2005), to given funding programs (NIH Research Career Development Award in Carter et al., 1987, NIH R01 grants in Jacob and Lefgren, 2011, CNR biotechnology and bio-instrumentation programs in Arora et al., 2000, New Zealand Marsden Fund in Gush et al. 2015), or to a single emerging country (Argentina for Chudnovsky et al., 2008; Chile for Benavente et al., 2012). These studies are performed on samples of limited size, with the exception of Jacob

and Lefgren (2011) who study more than fifty thousand applications for NIH R01 grants.

The main methodological issue all these observational studies have to deal with is the selection bias. Why university professors and researchers apply to the funding agencies, and why evaluators and committees select them, are also often the self-same reasons why they are likely to be more productive. Confounding factors are thus likely to affect both funding and the scientific outcomes, which would skew estimates in a naive approach. In order to deal with this concern, a first stream of the literature has used the scores obtained in the evaluation process to control for the selection effect in a regression discontinuity design, which compares the barely funded with the nearly funded applicants (Carter et al., 1987; Jacob and Lefgren, 2011; Benavente et al., 2012). In a parametric version of this approach, scores are included in the regressions as controls of applicant quality, or as instruments of the funding decision (Arora and Gambardella, 2005; Gush et al. 2015). These scores are built according to a weighting scheme which is specific to the funding institution and, sometimes, to the funding program considered. Scores may reflect the referees' evaluation of the applicant, of the project, and/or of the hosting institution quality. Depending on what information those scores are based on exactly, the estimated impact of the funding agency will or will not incorporate the effects of the applicant, the project and/or the hosting institution quality differentials. Surprisingly, certain studies find a negative relation between evaluation scores and scientific productivity (e.g. Gush et al. 2015).[1]

The second stream of the literature has used the propensity scores approach, combined with differences-in-differences, to control for both the time-invariant individual fixed effects and the selection on observables (Chudnovsky et al. 2008; Azoulay et al, 2011). The present study adopts this approach. Here, the identification of the impact of funding relies significantly on the quality of the observables prior to treatment, on which the fund allocation process is modeled (estimation of the propensity scores). Fortunately, we were able to assemble detailed information that covered, in particular, scientists' age, their institutions, fine grained research fields, and multidimensional publication profiles, which we can use to model fund allocation. Moreover, information is available for almost the whole of the reference population (not just for applicants). More precisely, we match the list of applicants with the list of all professors and researchers associated with a laboratory accredited by the Ministry of Research and Higher Education in France. That represents more than thirty thousand tenured scientists while we restrict ourselves to those fields which are sufficiently covered by the publication database we use (Thomson-Reuters ISI Web of Science).[2] We can, therefore,

[1]Restricting outcome measurements to the publications likely to have been generated by the projects the PIs initially applied for, can help obtain positive correlation with the scores given to the applications. However, recent studies exhibit contradictory results on scores being good predictors of projects success (Li and Agha, 2015; Fang et al., 2016), and, more importantly, it not clear whether the unsuccessful applicants eventually work on the projects they submitted.

[2]Mainly hard and bio-medical sciences but not exclusively as it also includes some social sciences. The

estimate the impact of receiving an ANR grant using control groups picked either among unsuccessful applicants to the same ANR program and year, or from the whole reference population. There are good reasons to select controls in each way. On the one hand, all applicants self-select and are thus more "similar". On the other hand, picking individuals in the much larger reference population increases the chance of finding controls that are more similar to the treated scientist in terms of the observables (especially as regards publication profiles). We do not postulate that one control group is preferable to the other but rely on a placebo parallel path tests. As our approch also includes a difference-in-difference component which allows us to control for all residual time-invariant confounding factors, a proper identification requires that the controls and the treated have parallel production paths. These tests show that the former type of controls performs better and they are thus used to run the reference estimations. Moreover, balance diagnosis tests show that those (properly weighted) controls have very similar observables than the treated when they have similar propensity scores. That does not formally prove the chosen controls are fully similar to the treated, but it is strongly consistent with such statement. Further, we find that the productivity divergence between controls and treated only starts three years after funding. These remarks converge convincing us that the chosen controls differentiate from the treated for some reasons that are unrelated to scientific productivity.

The remainder of the article is organized as follows. The data are presented first. Methodology comes next. Then, in three separate sections, come the results concerning the quantification of the average impact of funding, the results on the design of funding programs, and the remaining available results. The last section wraps up and discusses the main results.[3]

2 Data and variables

2.1 The data set

Data collection starts with a list of all researchers and professors associated with one laboratory accredited by the French Ministry of Higher Education and Research around the year 2010, which contains information on 49,225 persons.[4] All of these persons are tenured, whenever as full or assistant professors, assistant researchers or research directors. Once all individuals for which we do not have full and consistent information (status, institutional

selection of treated and controls to be included is done at the level of the detailed research field.

[3] A Supplementary Material is available online. It provides, in seven appendices, numerous details on the raw data, the variables, the methodology, the tests, the results and the publication data disambiguation.

[4] Are thus excluded all the tenured researchers and professors who are not associated to a lab, and those associated to laboratories in schools funded only by other ministries (such as the ministries of industry, agriculture or defense), or to laboratories solely associated to national research institutes (such as CNRS or CEA internal labs).

employer, laboratory, age, etc.) have been excluded, we are left with 48,328 persons. This list has then been matched to the names of the authors of scientific articles, letters and reviews (on the basis of their surname and first name initials) in the Thomson-Reuters ISI Web of Science, a well-known database which gathers all the documents published in the main scientific journals. The publication period covered in this study goes up to and including year 2012. Thus, the last publication year considered (2012) stands three years after the last funding year (2009) and seven years after the first funding year (2005). We collected more than 9 million documents which received more than 40 million citations. As these large publication records show, we are faced with a huge homonymy problem due to the absence of any reliably unique identifier of researchers in publication databases. A disambiguation algorithm has thus been developed based on a "seed + expand" methodology (Reijnhoudt et al, 2014). Basically, this algorithm works as follows: in a first step (seed), the algorithm validates articles by imposing strong conditions, particularly on the field and institutions, which should be consistent with what we know for each person. At this stage, the goal is to minimize false positives. In the second step (expand), the algorithm uses the information on the articles already validated in the seed step, to accept other articles which did not fully meet the conditions of the seed step. The program then iterates until no further document is accepted. Typically the information used concerns the co-authors, the references and the keywords. New papers are validated either because, on the top of some of the first-step conditions which are maintained, they have the same co-authors or cite the same references as already validated articles.[5] By the end of this process, a little over one million documents have been validated, approximately 11% of the initial set.

The affiliation of professors and researchers to scientific fields of investigation is based on a fine grained organization of science in France into peer groups called "sections". Such sections are specific to the institutional employer, either a national research institute (such as CNRS or INSERM) or the Ministry of Higher Education and Research for all professors employed in universities and schools. Each section members elect a national committee which usually accredits PhDs for recruitment (or sometimes even recruits directly), evaluates individuals, allocates promotions, etc. Most of the time, sections tend to be organized around specific disciplinary orientations.[6] We computed, for each section, the percentage of professors and researchers for whom no article was found in the database. On the basis of this information, we excluded a long list of sections, mostly in the fields of humanities and social sciences. We

[5]In order to evaluate the quality of this disambiguation process, we have constituted a benchmark of nearly 300 scientists who have created an ORCID number and are thus likely to have disambiguated their own publications. Detailed information on the algorithm and on the quality of the disambiguation are presented in the Supplementary Material, Appendix E.

[6]For a few specialized research institutes, the specialty of the sections is not straightforward, and we had to develop specific strategies. For instance, for INRA (the national research body dedicated to agricultural research), the allocation to disciplines has been performed on an individual basis.

suspect that these disciplines are not well covered by the database, either because scientific journal articles are not the main outcomes of their research, or because the principal journals of these disciplines are not well covered by the database. This leaves 31,081 persons.

The ANR sent us the list of all applications from 2005 to 2009, comprising 67,812 partners × applications. A project "partner" is defined as an institution which will directly receive the planned funds from the ANR if the application is successful. Each partner has its own scientific coordinator. For multipartner projects, there is only one project coordinating partner, whose scientific coordinator is the project PI. In multipartner projects, each partner recieves its funds directly from the ANR. Each partner coordinator is fully responsible for the engagement of the funds recieved by its institution and thus enjoys significant autonomy. Keeping only the partners × applications emanating from academia and for which the variables of interest are correctly documented (scientific coordinator's surname and first name, the partner, funding decision, amount, and duration), leaves 54,852 partners × applications. The success rate is 30%. The total amount allocated is 2.4 billion euros, but the expected total cost of the funded projects is 9.5 billion euros because the ANR funds only the marginal cost of the projects it supports for public partners.[7] The median fund per partner is 136,000 euros, while the mean is 138,000 euros. The mean total cost per partner is 545,000 euros.

We next basically matched the list of scientific coordinators of all ANR applications with the personnel list obtained previously. Two types of matching were performed subsequently: an exact matching and a fuzzy one.[8] In the event of homonymy in the full initial list of scientists, a manual check was made, based on the consistency between the discipline of the scientist and the project description, and between the employer of the scientist and the project partner. This matching allowed us to find, in the list of the 31,081 professors and researchers, the scientific coordinators of 46.2% of all applications and of 45,5% of the funds, and 46.9% of the total amount of money allocated.

It turns out that more than one third (10,722) of all these persons applied as scientific coordinators of the partners involved in the projects submitted between 2005 and 2009, and that 18.6% (5,786 persons) obtained at least one grant. The age distribution of the three populations (reference population, applicants and funded) is similar, though the 35-50 year-olds (in year 2010) are proportionately more numerous among the applicants and the funded. Researchers and full professors are more likely to have applied at least once. Researchers from CNRS and INSERM apply more often and their applications are more

[7] The grants cover the wages of the non-tenured personel hired for the purpose of the project and overheads limited to 4% of the grant. The total costs typically include the grant and all the resources included in the project, in particular the salaries of the tenured researchers and professors paid by the research institutes and universities.

[8] Fuzzy matching authorizes small variations in the surnames and first names and then requires manual verification and cleaning, basically comparing individual information and project information before validation.

likely to be successful. The applicants identified have applied on average 2.4 times over the period (25,364 applications). The distribution of applications is asymmetric, with most professors and researchers not applying or applying only once, while some apply many times. On average, the applicants obtained 1.2 grants over the five years considered (12,757 funds allocated). Like the applications, the funds are also unevenly allocated across the population: More than 75% of the applicants received only one funding, while a few got many. In this study, we will consider only the first funding for those who got multiple grants. There are two types of programs: thematic programs that have a specific thematic orientation, and non-thematic ones which are fully open to any application. While half of the applications go to thematic programs and the other half to non-thematic programs, thematic programs account for 65% of the grants allocated, because these programs have significantly higher rates of success.

When we break down applications by discipline, we observe that the highest rate of application is found for physics (with more than one application per scientist), followed by fundamental biology (.94), chemistry (.91) and applied biology and ecology (.90). The lowest rate is found for mathematicians who applied only one-third time on average. The highest average rate of funding can be observed for applied biology and ecology (with .34 funds per scientist). Physics follows immediately (with .32 funds per capita). These two fields differ strongly, however, in terms of supporting programs: physics is most often funded by non-thematic programs, whereas nearly two-thirds of the funds allocated to applied biology and ecology come through thematic programs. Similarly, fundamental biology, medicine and engineering sciences are mostly funded by thematic programs, while the sciences of the universe and mathematics are most often funded by non-thematic programs.

2.2 Variables

Three main basic output measurements will be used to assess the impact of funding.[9] The first sums the number of articles published, each being adjusted by the number of co-authors (fractional counts). The second measurement weights these articles by the number of citations they received (in a three-year time window), while the third weights them by the average number of citations which papers published in the journal that year received on average (again in a three-year window).[10] We label these three measurements respectively as Volume, Citations and Impact Factor. Though they are not independent, these indicators are distinct.

[9] Details on the calculation of all outcome variables are presented in Appendix B of the online Supplementary Material.

[10] This weighting scheme is very close to, but distinct from, the traditional Journal Impact Factor which divides the number of citations received in a given year (thus to articles published that year but also to those published previously) by the number of articles published that year. Therefore our approch is less sensitive to the yearly variations in the average quality or in the number of articles published.

The first relates more to the volume/quantity of scientific production, the second captures its direct impact on the scientific literature, while the third captures the capacity to publish in well-established, large-readership journals.

Publication data also prove to be very helpful in investigating the collaboration behaviors of professors and researchers (Wuchty et al. 2007). We use the number of authors of the article to evidence the size of the research teams, information which is averaged for each given period and person to obtain the Average Team Size. Collecting all collaborators' names and initials over given time periods and dropping double counts, we also compute the total number of distinct coauthors, labeled Coauthors. This number proxies the size of the collaboration network. We also compare the sets of collaborators between two consecutive time periods to assess the number of new coauthors they are working with (New Coauthors). The addresses of the authors' institutions can be used to assess the capacity of professors and researchers to extend their collaboration networks at the international level. The variable International Collaborations equals the number of articles that have at least one foreign address.

We define an indicator of Novelty for any set of scientific articles based on the atypicality of their keywords. All keywords appearing in research articles indexed in Thomson Reuters Web of Science (WoS) from 2000 to 2012 are considered. Let N_{ct} denote the number of (non-distinct) keywords to be found in the articles published in field c[11] and year t and let N_{kct} be the number of times the token k is actually used as a keyword by an article published that year in that field. The novelty of keyword k, in year t and field c, is defined as $-log\frac{N_{kct}}{N_{ct}}$. Note that as journals may be attached to multiple fields, so are articles whose keyword novelty is to be appreciated in different contexts. For instance, an article with three keywords published in a journal associated with two subject categories leads to six keywords×fields. To characterize the novelty of a set of articles, we use the value of the 90th percentile of the distribution of novelty of its associated keywords×fields. The underlying idea is that the most novel keywords×fields are likely to characterize the novelty of the whole set.[12]

Project variables are also available. In particular we have information on the role each person plays in the project: is she/he scientific coordinator of the whole project (the PI of the project), or only scientific coordinator of one institutional partner in a multipartner project. As the design of the ANR grant system provides each partner's scientific coordinator with a significant level of autonomy (in particular financial), we have chosen the partner level of analysis rather than the project level. However, the project PI role is specific, often not a desirable one to play and one that keeps busy with administration and coordination tasks. We thus keep track of the status of each partner's scientific coordinator in the project with

[11] A field is defined as one of the 251 WoS subject categories.

[12] We could have taken the maximum but the 90th percentile value has the advantage to cut some extreme values. Very similar results are obtained with alternative novelty measures such as taking the maximum.

a dummy labeled PI, which will allow us to check whether PIs are compensated for their efforts by increased scientific productivity and/or collaborations.

The literature has long emphasized that age plays a significant role in scientific outcomes.[13] In particular an inversed-U shape of scientific productivity has been observed in most fields of science. It will thus be used in some regressions as a control variable. As age may affect the odds of being granted, it will also be used to model treatment allocation (together with age squared). Moreover, we will investigate the differential impact of funding on younger and older persons. Therefore, a variable reporting the age at treatment/application time is built, which is dichotomized breaking the population in two sets: the ones below versus the ones above the median age at the treatment/application year (43).

3 Identifying the impact of funds

Controlling selection on observables: propensity scores

In this paper, as we focus on the effect of receiving an ANR award on successful applicants, we are interested in the so-called average treatment effect on the treated individuals, which is defined as follows:

$$ATT = E\left(Y(1) - Y(0) | T = 1\right), \tag{1}$$

where $Y(1)$ denotes the production when the applicant is funded, while $Y(0)$ refers to the counterfactual, i.e. the production if the applicant had not been funded. The event noted $T = 1$ means treatment occurs. The problem is that the counterfactual outcome is non-observable: either he/she is funded, or is not, but not both.

Propensity scores can help reduce the bias related to the selection on observable characteristics. Rosenbaum and Rubin (1983) show that, under the ignorability condition which states that adjusting for a set of covariates X is sufficient to remove all confounding factors, controlling for the propensity scores is sufficient. The propensity score $P(X)$ is defined as the probability of being "treated" (obtaining a grant in our case) given X: $P(X) = P(T = 1|X)$, with $0 < P(T = 1|X) < 1$. The propensity scores are reliably estimated when the conditional independence assumption (CIA) is verified.[14] It states that the potential outcome is independent of the treatment status, conditional on the propensity score. In other words, the treated individuals would have reached the same outcome levels as the controls having the same propensity score, if they had not been assigned to the treatment:

[13] To name a few: Lehman (1953), Zuckerman and Merton (1972), McDowell (1982), Levin and Stephan (1991).

[14] It is also known as Weak Unconfoundedness for the ATT.

$$E\left(Y(0)|T=1,p(X)\right)=E\left(Y(0)|T=0,p(X)\right)=E\left(Y(0)\right). \tag{2}$$

This equation can be rewritten as:

$$Y(0)\perp T|p(X). \tag{3}$$

Such an assumption, which cannot be tested directly, implies that there is no confounder influencing both the assignment of the treatment and the outcome that is not included in X. Heckman, Ichimura and Todd (1997) show that the non-inclusion of a relevant covariate causes the introduction of a bias in the estimated impact. In other words, the CIA assumption is valid only if all the covariates which influence both the treatment and the outcome variables are included in the set of explanatory covariates used for the estimation of the propensity scores.

Therefore, the covariates that are included in the vector X, which is used for estimating the propensity scores, need to be selected with caution. In this study, we use an "agnostic" approach whereby we investigate several specifications of the selection model that we test later. Further, the propensity scores are calculated in a standard way using logit regressions.

Matching and weighting using propensity scores

Different methods using the estimated propensity scores can be applied to remove the bias due to the differences between the observed characteristics of the treated and those of the untreated individuals. In this paper, we consider two matching procedures, nearest neighbors matching with replacement, and kernel matching, as well as inverse probability of treatment weighting (IPTW).

In the nearest neighbors matching, each treated individual is assigned its most similar controls (up to five) in terms of propensity score. To improve the quality of the matching, a caliper width is specified, which restricts the selection of the controls within a caliper around the propensity score of the treated individual (to avoid capturing controls that are too distant). The caliper value is calculated in line with Cochran and Rubin (1973), who tested the bias reduction when applying a caliper width $c=a\sqrt{(\sigma_1^2+\sigma_2^2)/2}$, along with σ_1 and σ_2, which are the standard deviations of the propensity scores among the treated individuals and the controls respectively, as well as with a as a positive parameter. Following Rosenbaum and Rubin (1985), we set $a=0.2$ which is expected to remove around 99% of the bias. Using a caliper condition however reduces the subset of available controls. Note that treated individuals will be excluded from the analysis if no control meets the imposed conditions (the caliper or the common support restriction).

Unlike the nearest neighbor approach which assigns the same weight to all controls of a given treated individual, the kernel matching approach assigns a different weight to each control, which is inversely proportional to the difference between its propensity score and that of the treated individuals. The kernel method provides an interesting solution when the nearest controls have very different propensity scores to those of the treated individuals. Frolich (2004) argues that kernel matching is always preferable to nearest neighbors matching. We exclude observations with extreme propensity score values. Following Imbens and Wooldridge (2008), we remove all individuals i, such that $p(x_i) > .9$ or $p(x_i) < .1$. We also apply the common support restriction, which implies that we do not consider controls with a lower propensity score than the lowest score among the treated individuals (Dehejia, 1999).

Robins, Hernan and Brumback (2000) and Hirano and Imbens (2001) argue that the controls with higher probabilities of being treated are likely to be under-represented in the control population (because they are likely to have been treated), whereas the controls with lower propensity scores are likely to be over-represented. To correct for this bias, the authors suggest weighting the controls by the inverse of the probability of being treated. The weights allow under-represented controls (because they are likely to have been treated) to have a more important role in the analysis as compared to the controls who have a low probability of receiving the treatment (who are thus likely to be over-represented). Hirano, Imbens and Ridder (2003) argue that this approach is more efficient.

Conditional difference-in-differences

So far we have considered that observed heterogeneity was sufficient for explaining the selection into treatment. However, in the applicants' CVs or in their project proposals, the selection committees and the external solicited referees can find relevant information that cannot be observed in our data, but which reveals their ability to perform in science. If this occurs, and if it influences the selection, then propensity scores are not sufficient for identification. However, if these unobserved variables are time-invariant, such as personal fixed-effects, then time differentiation can be used to solve the problem. The relevant approach is the so-called difference-in-differences methodology, which basically compares the variation in the performances of the treated individuals and the controls, before and after treatment. The outcomes variables are calculated by pooling together the information on the three years before and the three years after the year of funding.[15] Therefore, the publication outcomes issued in the year of funding are not considered. The three-year window ensures that the post-funding publication period considered is complete, even for the last funding year considered (2009) because publication data are available until 2012.

[15] In principle, this analysis could be done on a yearly basis. However, we follow Bertrand et al. (2004) who show that using only two periods is preferable because it reduces serial auto-correlation.

We use the conditional difference-in-differences model (Abadie, 2005) that combines a treatment selection model based on the estimation of the propensity scores with the difference-in-differences method. The estimation of the impact can be calculated as follows:

$$\hat{\delta}^{psm} = \frac{1}{|N_T|} \sum_{i \in N_T} \omega_i \left(Y_{i,1} - Y_{i,0}\right) - \frac{1}{|N_{NT}|} \sum_{j \in N_{NT}} \omega_j \left(Y_{j,1} - Y_{j,0}\right), \tag{4}$$

where N_T denotes the set of treated individuals and N_{NT} the set of controls. $Y_{i,t}$ is the outcome variable observed in period t, with $t = 1$ in the period after the treatment assignment, and $t = 0$ in the period before treatment. The weights ω_j are defined according to the chosen matching method. When the nearest neighbors or the kernel methods are chosen, the treated individuals have a unitary weight ($\omega_i = 1$) when included and the controls have a total weight which is accumulated over the treated individuals to which they are associated: $\omega_j = \sum_{i \in N_T} \frac{1}{|M(i)|} \omega_{j,i}$, with $\omega_{j,i}$ the weight of control j vis-à-vis treated agent i, and with $M(i)$ the set of controls for treated agent i. With the IPTW approach, the weights are calculated following a slightly different logic as controls are no longer specifically associated with given treated individuals. They are calculated as follows (cf. Robins, Hernan and Brumback, 2000; Hirano and Imbens, 2001): $\omega_i = T_i + \frac{(1-T_i)p(x_i)}{1-p(x_i)}, \forall i \in N_T \cup N_{NT}$, with $T_i = 1_{\{i \in N_T\}}$ the treatment dummy and $p(x_i)$ the propensity score of agent i.

The impact is then estimated by interacting post-funding dummy with treatment dummy in a fixed effect regression using two time periods' panel data, where observations are weighted according to the chosen method.

Conditional difference-in-difference-in-differences model

The conditional difference-in-difference-in-differences approach allows us to differentiate the effect according to program types or researcher characteristics in order to investigate the heterogeneity of the impact. It builds upon the basic conditional difference-in-differences model by introducing a supplementary level of differentiation based on the characteristics of interest. For instance, if we investigate the differentiation of the effect between, say, two types of programs (type 1 vs. type 2), the ATT of being treated by program type 1 as compared to being treated by program type 2 is given by:

$$\begin{aligned}\hat{\delta}^{psm}_{1-2} = & \frac{1}{\left|N^1_T\right|} \sum_{i \in N^1_T} \omega_i \left(Y_{i,1} - Y_{i,0}\right) - \frac{1}{\left|N^1_{NT}\right|} \sum_{j \in N^1_{NT}} \omega_j \left(Y_{j,1} - Y_{j,0}\right) \\ & - \left(\frac{1}{\left|N^2_T\right|} \sum_{i \in N^2_T} \omega_i \left(Y_{i,1} - Y_{i,0}\right) - \frac{1}{\left|N^2_{NT}\right|} \sum_{j \in N^2_{NT}} \omega_j \left(Y_{j,1} - Y_{j,0}\right) \right), \end{aligned} \tag{5}$$

where N_T^p is the set of persons who received funding of type $p \in \{1,2\}$, N_{NT}^p is the set of controls for the funded individuals of type p. The weights ω_j are defined as previously. The first part of the right side of the equation refers to the difference between the treated and control groups of type 1 programs, whereas the second part is the same difference for type 2 programs. The differencial impact of program 1 over program 2 equals the difference between those two terms. It is estimated using a similar regression as before, but now considering the coefficient of a triple interaction between post-funding dummy, treatment dummy and program 1 dummy.

Tests

Eight different designs have been retained and tested for calculating propensity scores.[16] All logit estimations regress the treatment dummy on individual variables such as age and on a series of variables characterizing the publication profiles at the time of funding. The latter includes the number of publications in the last three years to assess recent research intensity; the number of citations over the same period in order to account for the direct scientific impact of recent research; the highest impact factor of the journals in which they published to take into account the capacity to publish in prestigious journals; and the total number of citations received over their career so far, to account for their long-run reputation. The designs for calculating propensity scores differ, however, in several respects. Publication trend variables in the years preceding treatment are included in some logit regressions so as to capture the recent dynamics of scientific production before treatment. Some designs exclude all non-applicants, while others select controls from within the reference population as a whole. Some, however, require the controls to be in the same section as the treated individuals,[17] while others do not. Since the quality of the research environment is one of the selection criteria, we have also considered the inclusion of laboratory variables among the regressors.[18] In some designs the thematic and non-thematic programs are considered jointly while in others logit regressions are performed by program type, basically assuming that the selection mechanisms of thematic and non-thematic programs are distinct, based on different weights given to the observables, and even on different observables.

The difference-in-difference identification relies on the parallel path hypothesis, that is, the

[16] And for each design, the three weighting methods (five nearest neighbors, kernel and inverse probability of treatment weighting) have been tested. That makes 24 estimations for each outcome variable. Details are presented in the online Supplementary Material (Appendix C)

[17] Then propensity scores are computed separately for each section.

[18] A potential issue is that laboratory variables are observed after funding (approximately around year 2010). Therefore, if the consequence of funding is a mobility in a different lab (potentially of a higher quality), the impact of the funding may actually be underestimated (because the treated would be basically compared to controls in higher quality labs). Our results show, however, that the estimations retaining lab variables do not differ significantly and thus that lab variables do not lead to an underestimation.

treated individuals would have had production paths parallel to the ones of their controls if they had not been treated. This hypothesis cannot be tested before and after treatment, since the counterfactual is not available after treatment. However, the parallel path hypothesis can be tested between different periods before treatment (Imbens, 2004; Abaddie, 2005). Therefore, we estimate a hypothetical impact of the treatment on the treated individuals, between two distinct periods before treatment: t_0-3 and t_0-1, where t_0 stands for the year of funding. The goal is to verify that the variations in outcomes of the treated individuals before treatment are not significantly different from those of the controls. Performing such a test on all estimations,[19] we find that the predicted impact of the treatment is always very small and never significant. For robustness purposes, a similar test is performed between t_0-3 and t_0. This placebo test is more helpful in sorting out candidate estimations, and thus in selecting our reference propensity score estimations. The propensity scores that lead to the most parallel paths are those for which the controls are applicants exclusively; the estimations are performed distinctly between program types (thematic vs. non-thematic); the laboratory variables are not included; while production trend variables before treatment are included. Following the recent literature we will use the IPTW weighting scheme as our reference estimations but consider the other weighting methods. In fact, the results with the nearest-neighbors and the kernel matching methods remain essentially the same as the ones obtained with the IPTW. They are presented in Appendix F of the online Supplementary Material. We end up with 5,314 treated individuals and 9,618 controls.[20]

The conditional independence assumption on which our identification strategy is based implies that the treatment dummy is independent of the variables included in the logit model, conditional on the propensity score. Therefore, if the estimated propensity scores are correct, we expect that controls and treated individuals do not significantly differ with respect to the explanatory variables when they have similar propensity scores. Dehejia and Wahba (1999) suggest a balance test that builds upon this property. It consists in a comparison of means between treated individuals and controls for each variable included in the logit regression.[21] All the logit specifications pass this balance test. Our preferred propensity score estimations also pass this test once the individuals have been grouped into several strata of the propensity scores.[22]

[19] All parallel path tests are presented in the Appendix D of the online Supplementary Material.

[20] These numbers are obtained when using the IPTW method. We have 5,304 treated individuals and 7,983 controls when using the five-nearest neighbors method; and 5,305 treated individuals and 9,618 controls when using the kernel one.

[21] When possible, we use the procedure proposed by Becker et Ichino (2002).

[22] Those balance diagnosis tests (standardized differences of the means tests for each covariate of the logit regression) are performed for seven strata of the propensity score (separately for the non-thematic and thematic programs because the preferred propensity score estimations are also performed separately). All tests are presented in Appendix E of the online Supplementary Material.

4 The impact of project funding

Figure 1 shows that ANR funds persons who have an increasing publication trend, which starts before the year of first funding (at $t = 0$) and subsequently expands. Figure 1 also reports the properly weighted performances of their controls (dashed blue line). We see that the performances of the controls are slightly lower than those of the treated individuals. This is because the pre-treatment difference-in-differences placebo test has given priority to the similarity in trends with the treated individuals. Sets of controls which were more similar in outcome levels have been discarded because they do not satisfy all the placebo parallel paths tests so well.

And as expected, non-funded applicants have very similar trends to the funded agents until the year of funding (included). In fact, it turns out that the trends diverge only starting from the second year after funding. This is somewhat reassuring as it is sometimes claimed that researchers have often nearly completed their project when applying. If these projects are also more likely to be funded, then a positive impact could be partially driven by this phenomenon. However, in that case, divergence should occur early after the funding date, something that is not observed here. This does not mean that anticipated projects are not more likely to be funded but that the conditions we imposed for the selection of controls seem to have sorted out such an effect.

Outcome variables The main conditional difference-in-differences results are shown in Table 1. We find that receiving an ANR fund increases publications by 3.5% according to the preferred estimation. When the impact factor of the scientific journals in which articles are published is taken into consideration, receiving an ANR fund increases production by 8.3%. The impact of funding is strongest when citations are considered: a 15.2% increase is found. As we have seen that age can influence scientific productivity, we are worried that these results may be slightly biased by age differences between treated individuals and controls. However, there are no significant age differences between funded individuals and controls at the time of application (43.66 vs. 43.37). Moreover, unreported similar regressions but controlling for age and age squared exhibit no significant change in the results.

These results are more significant than those obtained for NIH grants (7% impact on citations) by Jacob and Lefgren (2011), though the mean amount of the funds allocated in our sample is far less than the average NIH N01 grant. Jacob and Lefgren (2011) report a 1.7-million-US-dollar NIH N01 grant on average as compared to an average ANR grant of less than 0.14 million euros, and an average total cost of 0.55 million euros. This difference may be due to the specificity of the biomedical sciences in the US for which the availability of funds and the variety of funding sources may induce a displacement effect (as the authors themselves argue). Such an effect occurs if the funded individuals expend less energy in

obtaining more funds than the unsuccessful applicants taken as controls. The plausibility of that explanation is reinforced by the fact that alternative sources of project funding than the ANR at the national level were relatively limited at the time of the study.[23] Note that our results are quite similar to those obtained in Gush et al. (2015), who use a different methodology, and data from a different country.

Collaboration patterns The literature has recently documented a long-run increase in the size of research teams proxied by the number of co-authors of the articles (Wuchty et al. 2007). We now document a hypothetical impact of project funding on team size. Coordinators may have incentives to delegate research tasks because they experience rising time constraints and because they have more financial resources to staff their teams. We find (see Table 1) a positive but limited impact of funding on the average number of authors per paper (2.2%). However, the impact of ANR funding on the total number of co-authors is significantly larger (9.8%). Thus, project-based funding increases the network of collaborators of the funded individuals more than it does the size of their research teams. This increase seems essentially due to the turnover of coauthors, as treated individuals have 6.7% more new collaborators than controls. This could be due to a higher capacity to hire PhD students or postdocs that eventually become coauthors on specific projects. It could also indicate that the funded individuals become more attractive as coauthors on the academic "collaboration market". To disantangle the two effects we would need to characterize further the collaborators of the treated individuals and controls, which is very difficult because of data limitations. We can however proxy the international span of their individual networks by counting the number of publications for which the authors gave at least one professional address outside France. Funding is found to increase the number of such articles by 4.2%, a result which is positive and significant albeit below the impact of funding on publication volume. This supports the idea that the two effects are at play: the funded individuals increase their networks by hiring, and also by collaborating more with independent colleagues. Moreover, this shows that ANR funding, which is mainly organized on a national basis, does not decrease the internationalization of collaborations but increases it, though to a limited extent.

Problem choice In principle, funding should open up new opportunities to tackle original problems. However, one of the biggest concerns about project-based funding is its capacity to sustain innovative and risky research. Azoulay et al. (2011) study impact differences between two US funding institutions with different grant designs. They find that the scientists supported on a program specifically funding researchers diversify their research through novel

[23] At the European level, the ERC was launched in 2007. It had however a limited budget in period 2007-2009: less than 1.7 billion euros for the whole of Europe. We matched the PIs of ERC grants in this period with our list of French professors and researchers, but found only a few scientists in the two lists.

research lines more than control scientists supported by programs which specifically fund projects (NIH-grants). Is it because the latter are bound to their initial proposals while the former can constantly redesign their goals? Or is it because, as research funding agencies often report, it is very difficult to induce jurys and external referees to support projets that are risky. If this is correct, applicants are not encouraged to submit risky proposals and thus funding should not lead them to engage in more novel research. To check this, we study the impact of funding on the novelty of the research projects based on the atypicality of their keywords (variable Novelty defined above). We find that funding has a positive but very small and not significant impact on the novelty of the research problems. Therefore, funded scientists are not likely to investigate more novel research lines.

5 Impact and the design of funding programs

We run a series of estimations in order to differentiate the impacts of funding depending on the variants or contexts of such "treatments". These investigations, conducted according to triple-difference estimations, allow us to shed light on the conditions under which project funding turns out to be more efficient. Our first investigations deal with the differentiated impacts according to the type of programs launched. The second series of results concern more the role of the individual characteristics of the recipients.

Thematic or non-thematic

The first question concerns the compared efficiency of the programs through which funds are allocated. Over the considered period, the ANR ran two main types of programs. The programs of the first type are thematic, meaning that specific calls are launched, mostly in new fields of research for which the agency has diagnosed a specific need or opportunity for its financial support. Because these calls are targeted, only subsets of possible recipients can apply. In the non-thematic programs, any possible project, a priori, goes. Therefore, non-thematic programs are likely to be characterized by a higher degree of competition. Observed success rates support this statement (37% in the thematic programs vs. 25% in the non-thematic ones). Therefore, self-selection is likely to be more pronounced in the non-thematic programs, and indeed, unreported analyses show that the applicants to non-thematic programs outperform on average applicants to the thematic programs. Which of the two types of programs should have a larger impact on scientific outcomes? On the one hand, we expect that thematic programs may make a big difference on targeted sub-populations. On the other hand, non-thematic programs are likely to have a larger impact, because the competition is stronger and because of their openness to any research idea.

We find (see Table 2) that thematic and non-thematic programs barely differ in terms

of their impact on the volume of scientific production: a 2.8% difference in favor of non-thematic programs, only significant at the 10% level. However, non-thematic programs turn out to be significantly more efficient when the impact factor of the journals or the number of direct citations are taken into account. Non-thematic programs lead to a 11.1% increase when articles are weighted by the average journal impact, and up to a 20.3% increase when articles are weighted by direct citations. These differences between program types are even larger than the overall impact of ANR funding. They strongly support non-thematic versus thematic programs as regard their scientific impact.

One of the potential advantages of the thematic programs over non-thematic ones could be related to the investigation of new problems. As thematic programs focus primarily on new and promising areas of science, we expect that they more strongly sustain the investigation of new research problems. This is precisely one of the main goals of the thematic programs. However, no significant differences were found in this respect between the two types of programs.

Interestingly though, thematic programs are clearly stimulating coauthor turnover as the recipients of thematic program funds have 17% more new collaborators than the recipients of non-thematic programs. These new collaborations could be an indicator of sound accomplishments that would be delayed in time because investigating new and more original problems takes more time. That would not yet be observable three years after funding. However, at this stage, no evidence in this study supports such a prospect.

Impacts along the career path

Estimating the impact of fund allocation at different career stages is an important policy issue. We thus ran estimations similar to the preceding ones, allowing us to differentiate the impact on younger scientific coordinators (equal to or less than 43 years old, the median age) from that on older ones. Results are reported in Table 3. We find non significant differences in the volume of publications and when articles are weighted by the journal's average impact factor. However, an important and significant difference is found in terms of citations: the impact on younger coordinators is 9.5% higher than that on older ones. This implies that the impact in terms of citations for younger coordinators is more than twice that observed among older scientists. This result is pretty strong and has significant policy implications. Further, no significant differentiated effect on collaborations or on novelty is observed. Funding only increases the team size of the older scientists slightly more than that of the younger scientists (2.8%, significant at the 10% level only).

We now differentiate the impact according to the publication profiles of the treated individuals at the time of funding. Our goal is to investigate whether some publication profiles

are more likely to be positively impacted by the funding policy than others. Treated individuals and controls are ranked within each discipline according to the number of citations received by their articles published in the preceding three years,[24] and are categorized in either one of the four largest deciles or in the remaining six deciles. In the triple difference approach, the performances of the top 10% are taken into reference. It is found that the treated individuals who are in the top 10% are never those on which the impact is the largest. Largest impacts are found in terms of publication volume when the treated individuals are in the second to the fourth deciles only, which are significantly larger than those of the first decile (from 8.2% to 11.8% larger). Similar statements can be made in terms of impact factor and citations, though coefficients are less significant. This can be explained by the fact that the top professors and researchers may have access to other sources of funds. Though the committees should select applicants who have strong publication records, the impact is not likely to be the largest when the funds are targeted to those who can obtain funds elsewhere, at the European level, for instance. Note that this statement is in terms of elacticities, not in absolute terms (number of citations for instance). A lower impact in terms of elasticity on top-10% performers may well correspond to a larger impact in absolute outcomes. On the other side of the distribution, when the treated individuals are not in the four largest deciles, the impact is likely to be significantly lower, not on the volume of publications, but both when the impact factor of the journal is considered and for citations. When, for instance, in the six lower deciles, the treated individuals have an average impact in terms of citations reduced by 9%, that is no longer significantly different from zero.

6 More results

We run a series of similar triple-difference estimations to address more residual but still interesting questions. The main results are presented in this section whereas their associated tables are to be found in Appendix F of the online Supplementary Material.

PI or not PI In a project, do partners free-ride on the PI who bears most of the between-partners coordination costs? Or, conversely, does the project PI free-ride on the partners' scientific coordinators, using their labor force to increase his or her scientific production? We find no significant difference according to the status of the treated individuals in the project, who can be either PI or partner scientific coordinator. Thus it seems that the benefits and costs of coordinating multipartner projects counterbalance each other. Gains of assuming the PI role are also not observed in collaborations. Unreported estimates show that the PI role has no effect on team size, number of coauthors and number of new coauthors. These

[24] We have used alternative performance variables to rank them, such as the number of articles, or even when such articles are weighted by the journal impact factor. Results are qualitatively similar.

results highlight how burdensome the PI role is. At the time of the proposal, assembling together all partners' contributions. At the time of the project, coordinating the work of all partners. The specific rules of the ANR, which give broad autonomy to the institutional partners and thus less power to the PI of the projects, probably does not help reduce such coordination costs. Another explanation is that it is still complex (though not impossible) in France to use project funds to reduce coordination costs or at least buy back teaching time, for instance.

Year effect No significant difference is found according to the year of funding. This result may seem surprising, bearing in mind that the agency was created in 2005. We guess that the agency has significantly increased its capabilities over the time period considered. We also know that the level of competition has been fairly different across years. For instance, the rate of success of the first year was much higher than that of the second year (48% as compared to 26%). In a sense, the fact that we find no significant difference between years is reassuring vis-à-vis our estimation methodology - tending to show that appropriate controls have been found for each year.

Scientific field effect When interaction with the scientific discipline is considered, we find that the impact of receiving funds is never significantly larger than in the life sciences, which is the reference. The only exception applies to Information and Communication Sciences and Technologies, where the impact is greater by 8.8% on citations and by 6.2% on the number of articles. Note that significance levels are however low (in particular for citations) and should thus be treated with caution.

7 Conclusions and discussion

In this article we have taken advantage of the recent French experience in which a new institution for project-based funding was created in 2005. This institution operates on a large scale, having distributed funds to research projects whose cumulated total costs approach ten billion euros over the five years covered by the study. Moreover, a certain level of variation in programs' rules and recipients' characteristics allows us to investigate the relative efficiency of variants of project funding. The results are not specific to one field of science, as all disciplines of hard and natural sciences are concerned (as well as some social sciences).

We identify the impact of receiving a research grant essentially by comparing the research production trajectories of the scientific coordinators of the funded projects with those of control groups. The controls are selected and weighted thanks to propensity scores that model the treatment on observables. Because the data on the whole reference population (not only

on applicants) as well as several useful variables potentially explaining selection are available, we can compare how various sets of controls pass parallel paths tests. The "best" set of controls according to those tests picks controls among applicants exclusively, models treatment by program types, and includes past publication performances at the time of treatment as well as recent trends. This suggests future studies should have similar information to obtain satisfactory control sets.

Concerning the global efficiency of project-based funding, our study concludes that a grant increases the number of publications weighted by citations by about 15%. That result is larger than what was previously observed in Jacob and Lefgren (2011). However, as our study is not limited to a specific scientific field and as few alternative opportunities for project-based funding were available at the time of the study in France, our results are less prone to be affected by a displacement effect (negative bias). This suggests that our quantification of project-based funding is the closest to the real effect.

Further, we also find that funding has a positive effect on the size of collaborators' network and on the turnover of collaborators. Although the agency under investigation operates on a national basis, it does increase international collaborations. Funding thus has a significant and positive impact on the scope of collaboration networks. One concern remains, however, since project funding does not affect the novelty of the research problems that are tackled by the funded individuals. This is a serious issue often raised by funding agencies themselves which would need further investigation.

Some of the most striking practical results of our study concern the differentiated impacts with respect to the types of programs and the characteristics of the recipients. We find that when programs have no specific thematic bounds, so that they are open to wider competition, they have a much larger impact. The difference between the impact of non-thematic and thematic programs appears even larger than the impact of funding. This nominal advantage of non-thematic programs is not offset by an increased novelty of the research performed by recipients of thematic grants. The latter however tend to collaborate much more with new people, which could suggest that such an effect exists but takes more time and so has not yet been observed. Last but not least, the funds allocated to younger applicants have much larger impacts than those allocated to older applicants. This strongly supports the idea that project-based funding should keep a large door open to younger applicants. If confirmed by other studies, these results may provide some guidelines for improving project funding in science.

Aknowledgements

Nicolas Carayol is Professor of Economics at the Université de Bordeaux, Research Fellow of the CNRS laboratory GREThA and Scientific Delegate at the Observatoire des Sciences et Techniques of the HCERES in Paris. Email: nicolas.carayol@u-bordeaux.fr. Marianne Lanoë is a PhD student at the Université de Bordeaux and Research Fellow at GREThA (Email: marianne.lanoe@u-bordeaux.fr). We would like to thank the research teams at OST-HCERES and GREThA and, in particular, C. Mouradian, E. de Turkheim and G. Filliatreau. We acknowledge the financial support of the ANR, deliberately very limited to avoid any form of conflict of interest. Special thanks go to Philippe Freyssinet, Michael Matlosz and Anne Portier from the ANR, for their advice, fruitful exchanges, and their complete respect of the independence of the authors, who are the only ones responsible for any remaining errors.

References

Abadie, A., 2005. Semiparametric difference-in-differences estimators, The Review of Economic Studies 72(1), 1-19.

Arora, A., David, P. A., Gambardella, A. 2000. Reputation and competence in publicly funded science: estimating the effects on research group productivity, 141-176: in Encaoua, D., Hall B., Laisnay F., Mairesse, J., The Economics and Econometrics of Innovation, Springer US.

Arora, A., Gambardella, A. 2005. The impact of NSF support for basic research in economics, Annales d'Economie et de Statistique 79-80, 91-117.

Austin, P. C., 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies, Multivariate Behavioral Research 46(3), 399-424.

Azoulay, P., Graff Zivin, JS., Manso, G., 2011. Incentives and Creativity: Evidence from the Academic Life Sciences, RAND Journal of Economics 42 (3), 527-554.

Becker, S., Ichino, A., 2002. Estimation of average treatment effects based on propensity scores, The Stata Journal 2(4), 358-377.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates?, Quarterly Journal of Economics 119, 249-275.

Carayol N., Lanoë, M., 2013. The impact of blue sky project based funding of academic research, mimeo.

Carter, G.M., Winkler, J.D., Biddle A.K., 1987. An evaluation of the NIH Research Career Development Award. Santa Monica, Calif, Rand Corporation.

Cochran, W. G., Chambers, S. P., 1965. The planning of observational studies of human populations, Journal of the Royal Statistical Society Series A (General), 234-266.

Cochran, W. G., Rubin, D. B. (1973). Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A, 417-446.

Dehejia, R., Wahba, S., 1999. Causal effects in non experimental studies: reevaluating the evaluation of training programs, Journal of Statistical American Association 94(448), 1053-1062.

Fang, F.C., Bowen, A., Casadevall, A., 2016. NIH peer review percentile scores are poorly predictive of grant productivity, eLife 5, e13323.

Frölich, M., 2004. Finite-sample properties of propensity-score matching and weighting estimators, The Review of Economics and Statistics 86(1), 77-90.

Gush, J., Jaffe, A.B., Larsen, V., Laws, A., 2015. The effect of public funding on research output: the New Zealand Marden Fund, NBER Working Paper #21652.

Heckman, J. J., Ichimura, H., Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The review of economic studies, 64(4), 605-654.

Hirano, K., Imbens, G. W. 2001. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes Research Methodology 2(3-4), 259-278.

Hirano, K., Imbens, G. W., Ridder, G., 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, Econometrica 71(4), 1161-1189.

Imbens, G. W., 2004. Nonparametric estimation of average treatment effects under exogeneity: A review, Review of Economics and statistics 86(1), 4-29.

Imbens, G. W., Wooldridge, J., 2008. Recent Developments in the Econometrics of Program Evaluation, National Bureau of Economic Research w14251.

Jacob, B. A., Lefgren L., 2011. The Impact of Research Grant Funding on Scientific Productivity, Journal of Public Economics 95(9-10), 1168-1177.

Lehman, H.C., 1953. Age and Achievement. Princeton University Press, Princeton.

Levin, S.G., Stephan, P.E., 1991. Research productivity over the life cycle: evidence for academic scientists. American Economic Review 81, 114–132.

Li, D., Agha, L., 2015, Research funding. Big names or big ideas: do peer-review panels select the best science proposals?, Science 348(6233), 434-8.

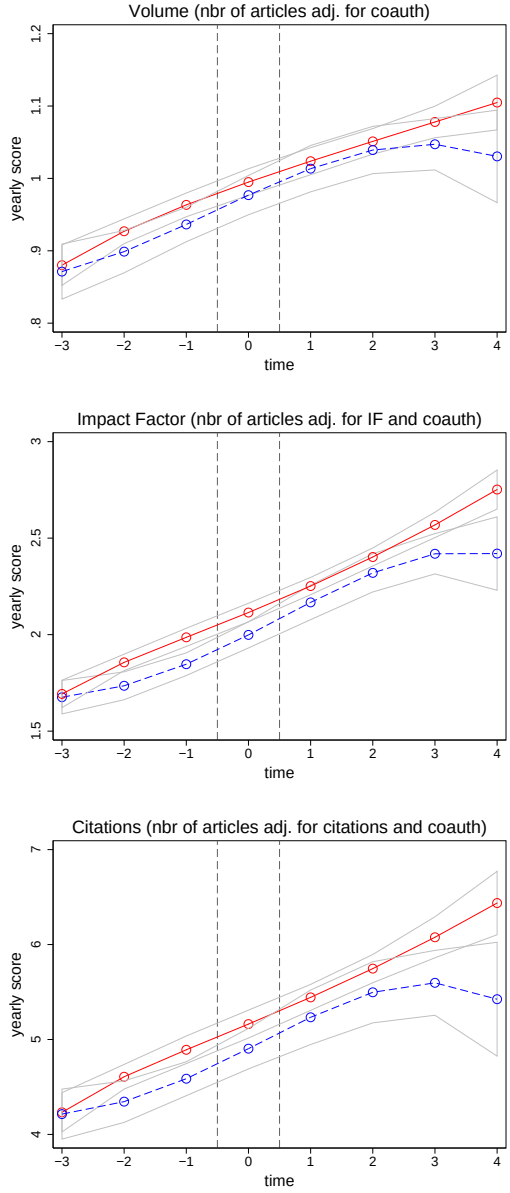
McDowell, J.M., 1982. Obsolescence of knowledge and career publication profiles: some evidence of differences among fields in costs of interrupted careers. American Economic Review 72, 752–768.

National Science Board, 2016, Science & Engineering Indicators 2016, National Science Foundation.

Reijnhoudt, L., Costas, R., Noyons, Ed, Börner, K., Scharnhorst, A., 2014. 'Seed + expand': a general methodology for detecting publication oeuvres of individual researchers, Scientometrics 101, 1403-1417.

Robins, J. M., Hernan, M. A., Brumback, B., 2000. Marginal structural models and causal inference in epidemiology, Epidemiology 11(5), 550-560.
Rosenbaum, P. R., Rubin, D. B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, The American Statistician, 39(1), 33-38.
Rosenbaum, P. R., Rubin, D. B., 1983. The central role of the propensity score in observational studies for causal effects, Biometrika, 70(1), 41-55.
Rosenbaum, P. R., 2002. Observational Studies. New York: Springer.
Stephan, P., 2012. How Economics Shapes Science, Harvard University Press.
Wuchty, S., B. F Jones and B. Uzzi, 2007. The Increasing Dominance of Teams in the Production of Knowledge, Science 316(5827): 1036-1039.
Zuckerman, H.A., Merton, R.K., 1972. Age, aging, and age structure in science. In: Riley, M.R., Johnson, M., Foner, A. (Eds.), A Sociology of Age Stratification: Aging and Society, vol. 3, Russel Sage foundation, New York, Reprinted in: Storer, N.W. (Ed.), 1973. The Sociology of Science: Collected Papers of R.K. Merton, Chicago University, Chicago Press, pp. 497–559.

Figure 1: Yearly scientific outcomes of the funded professors and researchers (red solid line) and of their controls (blue dashed line).



Note: Mean and 95% fractional polynomial confidence intervals are presented. The year of first funding occurs at time = 0. The points of the fourth year after funding consider only the data on the funded individuals (and their controls) in year 2008 and before. For the Citations indicator only, observations of years 2011 and 2012 are excluded because, for those publication years, the three-year citation window is not completely observed. Observations are weighted according to the inverse probability of treatment.

Table 1: Average treatment effect of receiving an ANR grant on publication outcomes and collaboration behaviors (the three years after treatment against the three years before).

Volume	Impact Factor	Citations
0.03503***	0.08252***	0.15254***
(4.46)	(7.52)	(9.28)
Av. Team Size	Coauthors	Internat. Collab.
0.0218***	0.0981***	0.0418***
(2.71)	(7.02)	(2.82)
New Coauthors[a]		New Problems
0.0668***		0.00183
(3.03)		(1.14)

*Note: Conditional difference-in-differences results. Coefficients and standard errors of the interaction term between the post-funding period dummy and the treatment dummy in a fixed effect regression. Observations are weighted according to the inverse probability of treatment. Dependent variables in Log. Robust standard errors in parentheses, clustered at the project level. Significance levels: 0.01: ***, 0.05: **, 0.10: *.*

[a] *Conditional differences results only as this variable counts the new items in the post-treatment period as compared to the pre-treatment period.*

Table 2: Differentiated effects of receiving an ANR grant on outcomes according to two different funding schemes: non-thematic versus thematic (the three years after treatment against the three years before).

Volume	Impact Factor	Citations
0.02766*	0.11112***	0.20275***
(1.77)	(5.08)	(6.24)
Av. Team Size	**Coauthors**	**Internat. Collab.**
-0.00107	0.0201	0.0286
(-0.07)	(0.72)	(0.97)
New Coauthors[a]		**New Problems**
-0.170***		0.000460
(-3.31)		(0.14)

*Note: Conditional difference-in-difference-in-difference results. Coefficients and standard errors of the triple interaction term between the post-funding period dummy, the treatment dummy and the non-thematic-program dummy, in a fixed effect regression. Observations are weighted according to the inverse probability of treatment. Dependent variables in Log. Robust standard errors in parentheses, clustered at the project level. Significance levels: 0.01: ***, 0.05: **, 0.10: *.*

[a] *Conditional differences results only as this variable counts the new items in the post-treatment period as compared to the pre-treatment period.*

Table 3: Differentiated effects of receiving an ANR grant on outcomes according to age dummy: below the median age (43) versus over the median age (the three years after treatment against the three years before).

Volume	Impact Factor	Citations
0.0221	0.0266	0.0952***
(1.41)	(1.29)	(3.09)
Av. Team Size	**Coauthors**	**Internat. Collab.**
-0.0279*	-0.00344	0.0280
(-1.80)	(-0.13)	(0.95)
New Coauthors[a]		**New Problems**
-0.0418		-0.0000433
(-0.83)		(-0.01)

*Note: Conditional difference-in-difference-in-difference results. Coefficients and standard errors of the triple interaction term between the post-funding period dummy, the treatment dummy and the below-the-median-age dummy, in a fixed effect regression. Observations are weighted according to the inverse probability of treatment. Dependent variables in Log. Robust standard errors in parentheses, clustered at the project level. Significance levels: 0.01: ***, 0.05: **, 0.10: *.*

[a] *Conditional difference-in-differences results only as this variable counts the new items in the post-treatment period as compared to the pre-treatment period.*

Table 4: Differentiated effects of receiving an ANR grant on publication outcomes according to the position in the citation distribution at the time of funding (the three years after treatment against the three years before).

	Volume	Impact Factor	Citations
Top-10-to-20% publication performance (vs. top-10%)	0.0823***	0.0608	0.0989*
	(3.06)	(1.62)	(1.80)
Top-20-to-30% publication performance (vs. top-10%)	0.106***	0.0675*	0.0716
	(4.05)	(1.87)	(1.33)
Top-30-to-40% publication performance (vs. top-10%)	0.118***	0.0880**	0.0921*
	(4.42)	(2.39)	(1.68)
Bottom-60% publication performance (vs. top-10%)	0.0632**	-0.0188	-0.0898*
	(2.38)	(-0.55)	(-1.70)

*Note: Conditional difference-in-difference-in-difference results. Coefficients and standard errors of the triple interaction term between the post-funding period dummy, the treatment dummy and the percentile-class-of-the-citations-volume-prior-to-application dummy (mentioned at the right of each line, the top-10% are in reference), in a fixed effect regression. Observations are weighted according to the inverse probability of treatment. Dependent variables in Log. Robust standard errors in parentheses, clustered at the project level. Significance levels: 0.01: ***, 0.05: **, 0.10: *.*