Supplementary material to the article: "The Impact of Project-Based Funding in Science: The French ANR Experience"

Nicolas Carayol[◊,‡] and Marianne Lanoe[◊]

[◊] GREThA, Université de Bordeaux - CNRS
[‡] Observatoire des Sciences et Techniques, Paris

January 17, 2017

Appendix A. The data and descriptive statistics

1 The database

Our data come from three different sources: administrative data about all tenured researchers and professors in France, their publications as listed on the Thomson Reuters ISI Web of Science (WoS) database, and the list of researchers who applied to a call for projects from the French funding agency, Agence Nationale de la Recherche (ANR), between 2005 and 2009. More precisely, the data set is built as follows.

- We start with 49,225 tenured researchers and professors associated with a laboratory certified by the French Ministry of Higher Education and Research from around 2010. These files inform us about some individual attributes, such as age, employing institution, laboratory and fine-grained discipline.[1] After some cleaning, we remove those who do not have the different variables correctly filled in. We are then left with 48,328 fully informed persons.

- We next retrieve the full list of publications associated with the researchers from the Thomson-Reuters ISI Web of Science database (surname and first name initials). For this purpose, we applied a thoughtful filter that we developed on the basis of a "seed+expand" approach to deal with the homonymy issue. We present the disambiguation procedure in detail in Appendix E. Some disciplines, however, are not well covered by the publication database (because most of their scientific production is not published in international journals but in books or native language outlets for instance). We thus decided not to consider researchers associated with fine-grained fields (sections) which have a high proportion of non-publishing scholars.[2] Once we collect the publications data (from 1999 to 2012) at the end of the disambiguation process, our sample consists of 31,081 researchers, which represents more than 63% of our initial population (some removals are also due to the disambiguation process, see again Appendix E).

- The last step rests upon the identification of the applicants to ANR calls for projects between 2005 and 2009 in our large sample of professors and researchers. Our initial ANR database consists of 67,812 partners×applications between 2005 and 2009. It provides information about individuals, such as project duration, institutional partner, laboratory, scientific coordinator and their role in the whole project (PI or not) and the value of the grant. After some manual cleaning, this list is reduced to 54,852 partners×applications due to the removal of all partners×applications that are not properly informed or do not concern a public research institution. We refer to this as the whole set of applications because we are not concerned with applications coming from companies in this study, and (nearly) all academic institutions in France are public. This set is described straight below in the next subsection, because it is interesting in its own. The list of scientific coordinators of these partners×applications is then matched with our administrative list of researchers based on their surname and first name. We first use an exact matching approach, while, in a second stage, we apply a fuzzy matching to overcome spelling errors for the remaining identities, which is associated with a systematic comparison of individual characteristics in order to control for homonymy issues. Out of our sample of 31,081 professors and researchers, 10,722 are identified as applicants to at least one ANR application during the period 2005-2009. Among these applicants, 5,786 were awarded a grant at least once (see Table 1).

[1] The "sections". See below and the main article for an explanation of what these sections are.

[2] It turns out that this removes all those in the humanities and most social sciences.

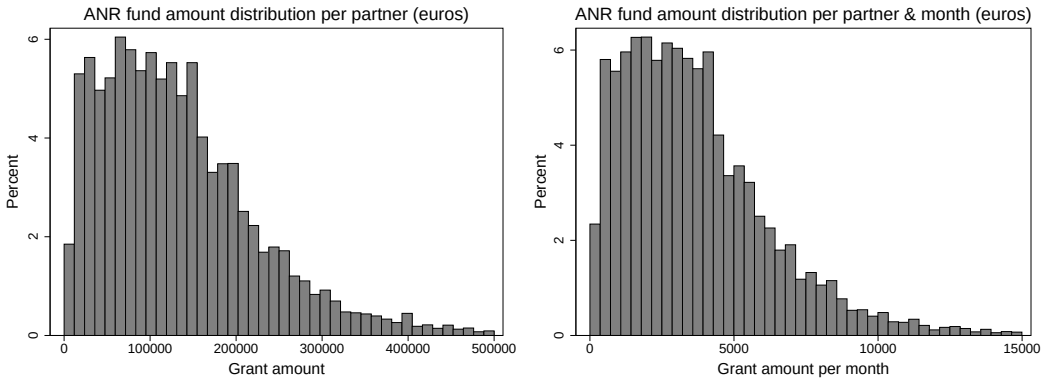
Table 1: The final sample of 31,081 researchers, the applicants and the granted

	number	share
Non applicants	20,359	65.50
Applicants	10,722	34.50
Granted	5,786	18.62
Total	31,081	100.00

2 Descriptive statistics on the whole set of applications

The left graph of Figure 1 shows the histogram of ANR grant values assigned to all the funded projects in the period 2005-2009. The distribution is a log-normal shape, with a plateau around the median (equal to 136,000; the mean is 180,000) and a long right tail. The shape of the monthly subvention distribution (right graph of Figure 1) is similar to the previous one, with a mean of 4,800 and a median of 3,700.

Figure 1: Histogram of the amount of ANR funding



Note: The right tail of the distribution is cut to preserve confidentiality.

The yearly number of applications, number of fundings, total amount of the grants attributed and total cost of the subsidized projects are presented in Table 2. The selection rate is 30% on average, which leads to an accumulated grant amount of 2.4 billion euros over the period 2005-2009. The selection rate was significantly higher in 2005, which was the first year after the creation of the agency, due to a lower number of applications. The estimated total cost of the funded projects is almost four times higher than the amount of the bestowed ANR grant.

Table 2: Applications from the public sector and funded partner×project by years (in # and amounts in million euros)

-	# Applications	# Granted	Grants amounts	Total Cost
2005	5,616	3,553	417 me	1,510 me
2006	12,881	4,188	511 me	2,160 me
2007	11,655	3,496	483 me	1,910 me
2008	9,769	3,315	522 me	2,040 me
2009	14,931	2,890	470 me	1,890 me
Total	54,852	17,442	2,403 me	9,510 me

3 Descriptive statistics on the final sample

The yearly number of applications and amount of funding granted are presented in Table 3. The matching between the administrative list of French researchers and the ANR applications data set allowed for the proper identification of 46.2% of the applications and 45.5% of the granted ones. The related ANR budget represents 46.9% of its total outlay over the period.

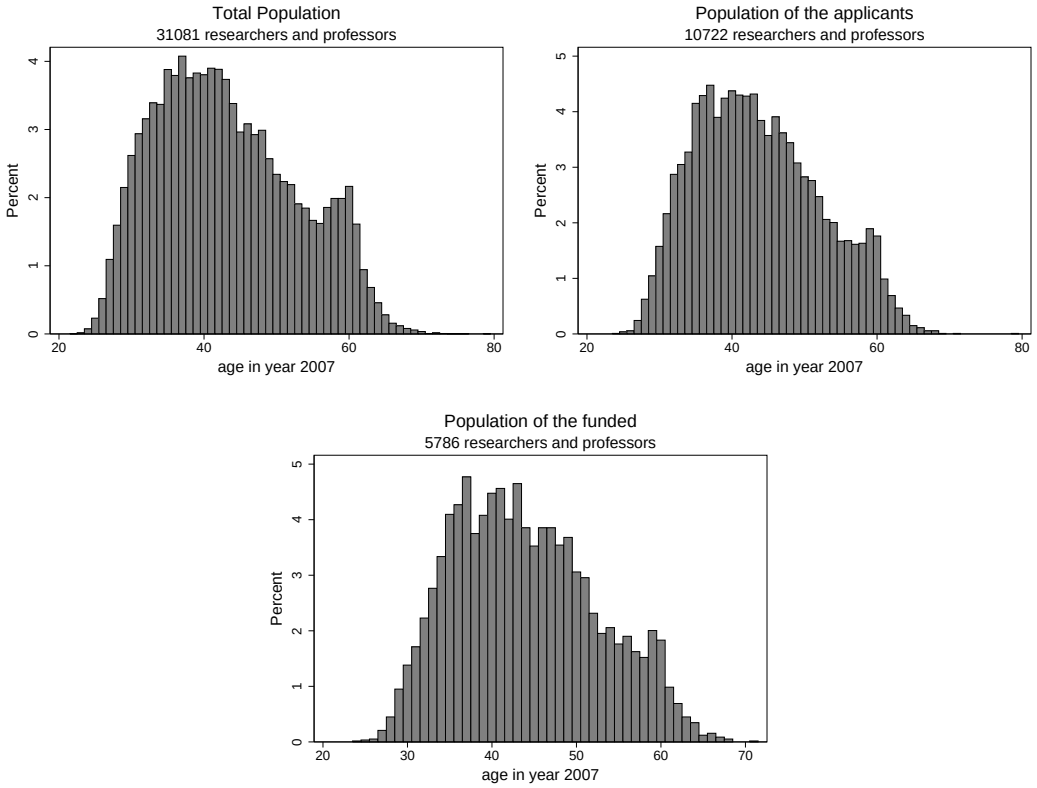
Table 3: Applications from the public sector and funded partner×project by year for our final sample (in # and amounts in million euros)

-	# Applications	# Granted	Grants amounts	Total cost
2005	5,422	2,605	185 me	476 me
2006	8,072	2,125	237 me	1,100 me
2007	4,994	1,473	243 me	984 me
2008	3,559	1,000	231 me	985 me
2009	3,317	742	232 me	856 me
Total	25,364	7,945	1,128 me	4,400 me
Prop	46.2%	45.5%	46.9%	46.3%

Note: The total number of applications is higher than the number of applicants (10,722) because they applied 2.37 times on average over the period. Some of the applicants also received multiple fundings (this accounts for 5,786 applicants granted).

The age distributions of the researchers and professors for the three samples are presented for the three samples in Figure 2. We can see therein that the distributions are quite similar, although the 35 to 50 years of age group are, in proportional terms, slightly more numerous in the sample of grant recipients and the total applicants compared with the overall population.

Figure 2: Age histograms for the total population, the applicants and those funded



In Table 4, we can see that full time researchers (denoted by CR and DR) and full professors (PR) are proportionally over-represented among the applicants. Senior researchers (DR) tend to be even more represented when considering the allocated grants. This is mainly due to the participation and success of CNRS researchers, who represent the vast majority of researchers in our database (see Table 5). Although assistant professors and full professors (72,2% of the population, denoted by the "UNIV" acronym) prevail in our sample, this group has the lowest level of funded individuals: 1.4 granted out of 10 individuals, compared with other groups, which have at least 3 granted out of 10 individuals (with the exception of IRD). The 10,722 applicants identified in our final sample applied 2.37 times on average, corresponding to 25,364 applications over five years.

Table 4: Researchers' and professors' status in the three samples

-	Full sample		applicants		granted	
	#	%	#	%	#	%
CR	5,290	17.02	2,335	21.78	1,260	21.78
DR	3,340	10.75	2,153	20.08	1,462	25.27
MCF	13,887	44.68	2,679	24.99	1,115	19.27
PR	8,564	27.55	3,555	33.16	1,949	33.68
Total	31,081	100.00	10,722	100.00	5,786	100.00

Note: "MCF" (for maître de conférence) is the equivalent of assistant professor (with tenure), whereas "PR" stands for full professor. "CR" stands for chargé de recherche, and "DR" stands for directeur de recherche. Both statuses represent positions that are dedicated full-time to research activity. "DR" corresponds to a senior researcher position. Note that, in France, all these statuses confer a civil servant position and therefore imply tenure.

Table 5: Researchers and professors' employing institutions in the three samples

-	Full sample		Applicants		Granted	
	#	%	#	%	#	%
CNRS	6,580	21.17	3,473	32.39	2,114	36.54
INRA	380	1.22	185	1.73	113	1.95
INRIA	146	0.47	82	0.76	58	1.00
INSERM	1,290	4.15	668	6.23	396	6.84
IRD	235	0.76	81	0.76	41	0.71
UNIV	22,450	72.23	6,233	58.13	3,064	52.96
Total	31,081	100.00	10,722	100.00	5,786	100.00

Note: "UNIV" stands for universities. The CNRS is a public institution, which supports research in any scientific field. The remaining institutes are specialized ones: the INRA is the national agronomic research institute, the INRIA is the national research institute of computer science and automation, and the INSERM is the national institute for health and medical research. The IRD is the national institute for development. It focuses on the interactions between men and their environment in emerging countries.

In Figure 3, we see that the distribution of the number of applications is skewed to the right, with most professors and researchers not applying. Among the ones who apply, most apply only once, while some are applying many times. The applications in our final sample are equally divided between thematic and non-thematic programs on the whole period (see Table 6). In the first years, after the creation of the ANR, the number of applications for thematic programs was higher. The importance of the two types of programs gradually balances before reversing in 2008. In 2009, non-thematic applications represented nearly two thirds of all applications. On looking more in

detail the applications to the seven specific thematic programs that were launched,[3] we observe that the number of applications is highest for the Biology and Health theme. We also note significant variations between years for a given program in terms of the number of applicants.

Figure 3: Histogram of the number of applications for all programs (top graphs) and by type of program (thematic or non-thematic, bottom graphs)

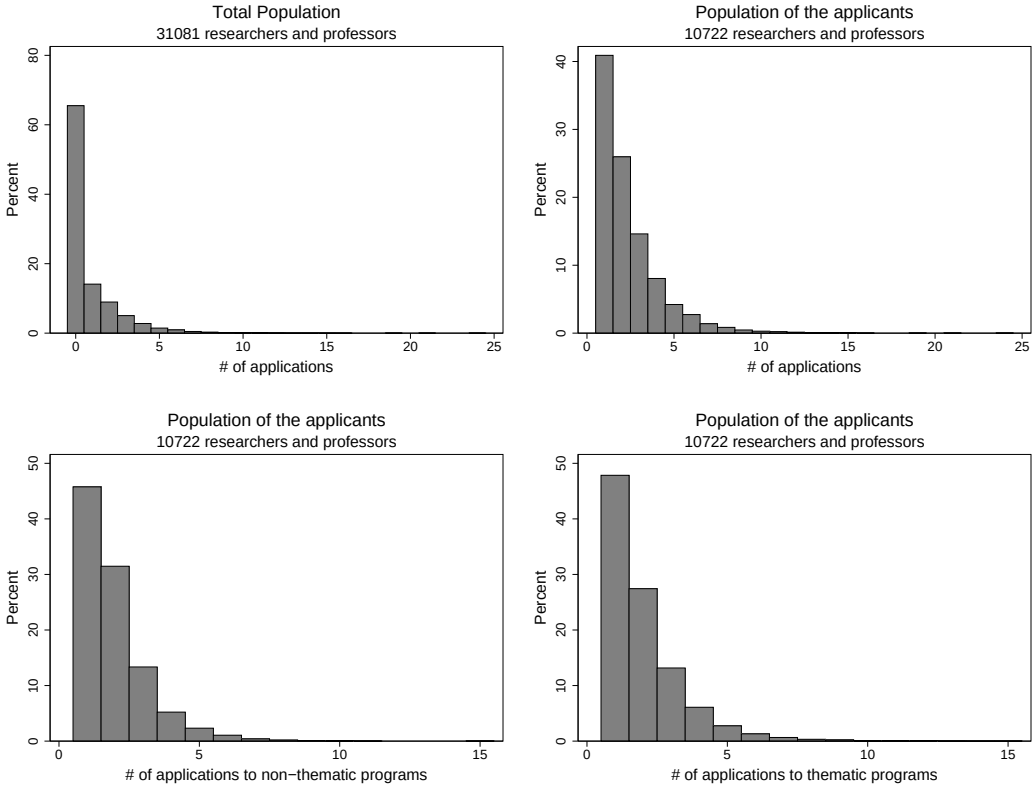


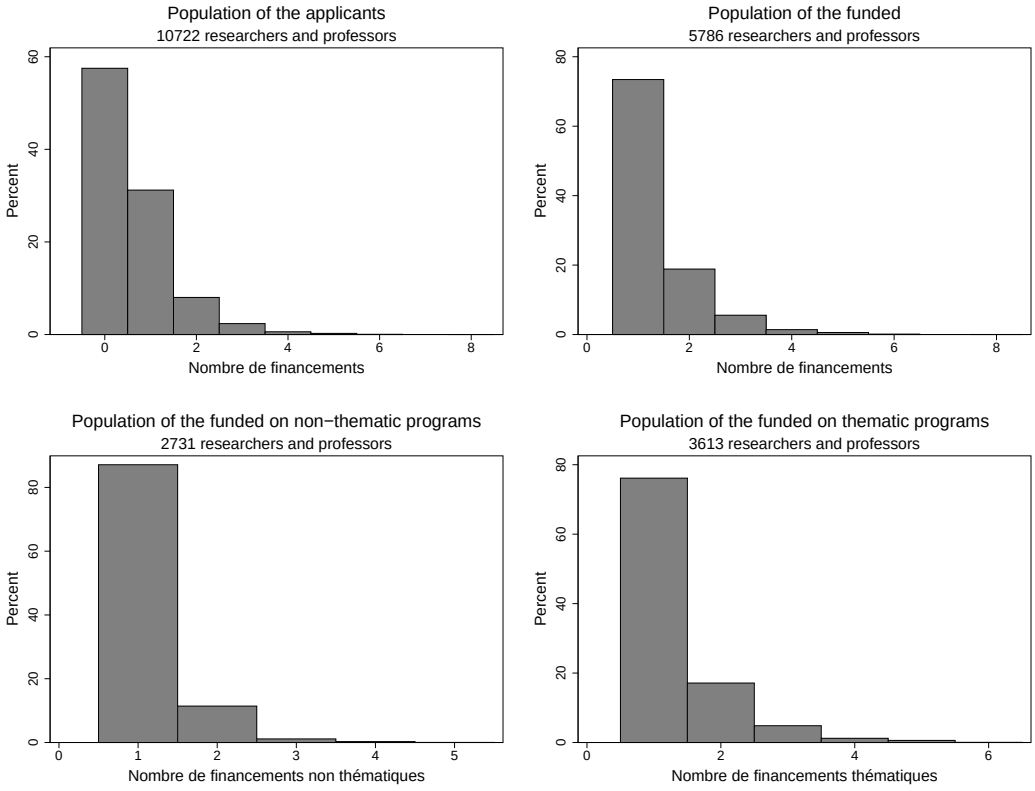
Table 6: Number of applications by year and by program

-	2005	2006	2007	2008	2009	Total
Non-Thematic Programs	2,201	3,860	2,372	1,806	2,172	12,411
Thematic Programs	3,221	4,212	2,622	1,753	1,145	12,953
Biology & Health	1,128	2,165	1,152	630	418	5,493
Ecosystems & Sustainable Development	229	202	151	145	125	852
Renewable Energy & Environment	578	447	400	304	213	1,942
Engineering, Methods & Security	71	136	202	165	66	640
Materials & Information	861	788	200	83	53	1,985
Human & Social Sciences	25	123	84	48	39	319
Information & Communication Sc. & Tech.	329	351	433	378	231	1,722
All Programs	5,422	8,072	4,994	3,559	3,317	25,364

[3]The programs are entitled "Biology and Health", "Ecosystems and Sustainable Development", "Renewable Energy and Environment", "Engineering, Methods and Security", "Materials and Information", "Human and Social Sciences", and "Information and Communication Sciences and Technologies".

The successful applicants received 1.37 grants on average over the relevant period.[4] The grants distribution is also skewed, but with a smaller right tail (Figure 4). More than 70% of the applicants are granted only once over the five years. Almost two thirds of these grants relate to the aforementioned thematic programs, while the remaining one third relates to non-thematic programs (Table 7). As for applications, thematic programs are predominant among all fundings awarded at the beginning of the period, although their share decreases afterwards. However, the rise in non-thematic programs over the period is less pronounced for fundings than for applications: thematic and non-thematic programs balance out in 2009. The number of grants is also unequal between programs, with the same features as the number of applications (see Figure 4).

Figure 4: Histogram of the number of grants for all programs (top graphs) and by type of program (thematic or non-thematic, bottom graphs)



We now focus on the participation in ANR programs according to the scientific disciplines of the applicants. For this purpose, we first allocate sections[5] to large disciplines. This allocation turns out to be complex in a limited number of sections because of the multidimensional nature of some sections. When this issue could not be resolved, allocation is made across multiple disciplines. We observe (see Table 8) that the highest application rate is found in Physics, followed by the Life Sciences (with Chemistry and Applied Biology not far behind).

[4]That is to say, 5,786 funded researchers shared 7,945 grants.

[5]The list of sections is given in Table 10. Researchers could be assigned to one of the 99 different sections, which are specific to their employing institution (if they are professors, it would be the Ministry of Research and Higher Education). The types of research centers in our database are INRA (agronomic research), INRIA (computer science and engineering), INSERM (medical research), CNRS and universities, each of them has its own classification in terms of specialties.

Table 7: Number of grants by year and by program

-	2005	2006	2007	2008	2009	Total
Non-Thematic Programs	1,001	794	581	372	385	3,133
Thematic Programs	1,604	1,331	892	628	357	4,812
Biology & Health	622	576	330	170	108	1,806
Ecosystems & Sustainable Development	193	144	96	88	35	556
Renewable Energy & Environment	279	185	143	115	69	791
Engineering, Methods & Security	17	17	79	61	17	191
Materials & Information	358	279	0	0	0	637
Human & Social Sciences	3	27	22	14	3	69
Information & Communication Sc. & Tech.	132	103	222	180	125	762
All Programs	2,605	2,125	1,473	1,000	742	7,945

By contrast, the application rate for Mathematics is the lowest (less than half the rate for Physics). In some disciplines, such as Life Sciences, Medicine and Engineering, professors and researchers applied more frequently to thematic programs, whereas non-thematic programs were preferred in relation to Physics, Sciences of the Universe, and Mathematics. In terms of the number of granted applications, the highest funding rate is found in Applied Biology and the lowest is found in Mathematics. Results by program go along with those for the applications. The prevalence of grants related to thematic programs is also found in the Life Sciences, Medicine and Engineering. On the contrary, Physics, Sciences of the Universe and Mathematics were more often funded through non-thematic programs. The allocations are fairly balanced between the two types of programs in Social sciences and Chemistry.

Table 8: Allocation of the ANR applications into large disciplines for our final sample

	Researchers	Non-Thematic		Thematic		Total	
Disciplines	#	#	%	#	%	#	%
Life Sciences	6,036	2,423	0.40	3,261	0.54	5,684	0.94
Medicine	3,478	1,055	0.30	1,773	0.51	2,828	0.81
Applied biology - Ecology	1,798	906	0.50	707	0.39	1,613	0.90
Chemistry	3,842	1,835	0.48	1,669	0.43	3,504	0.91
Physics	3,182	1,878	0.59	1,428	0.45	3,306	1.04
Sciences of the Universe	2,202	1,259	0.57	339	0.15	1,598	0.73
Engineering	6,441	1,845	0.29	3,068	0.48	4,913	0.76
Mathematics	2,646	709	0.27	335	0.13	1,044	0.39
Social Sciences	1,524	562	0.37	408	0.27	970	0.64
Total	31,149	12,472	0.40	12,988	0.42	25,460	0.82

Note: The total number of applicants is 31,149 instead of 31,081 because of the multiple allocations of some sections to several disciplines. The number of applications in Social Sciences is low considering we excluded most Human and Social Sciences disciplines from the analysis.

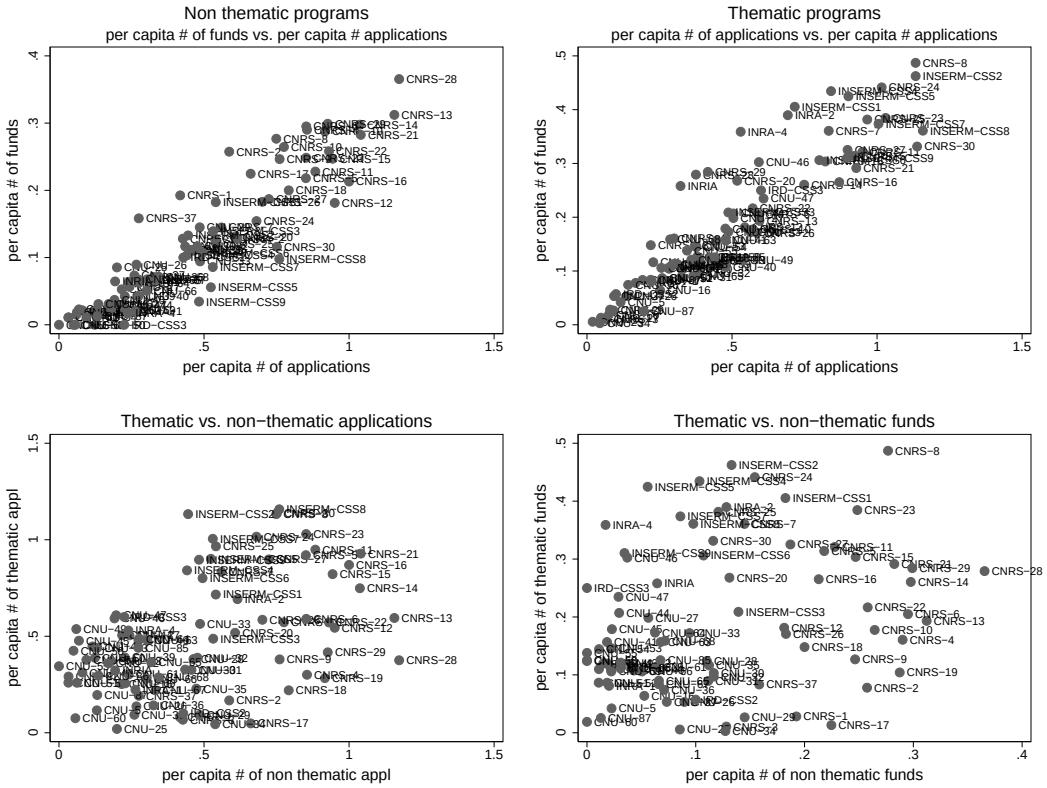
Table 9: Allocation of the ANR granted applications into large disciplines for our final sample

	Researchers	Non-Thematic		Thematic		Total	
Disciplines	#	#	%	#	%	#	%
Life Sciences	6,036	538	0.09	1,132	0.19	1,670	0.28
Medicine	3,478	195	0.06	645	0.19	840	0.24
Applied biology - Ecology	1,798	217	0.12	399	0.22	616	0.34
Chemistry	3,842	428	0.11	508	0.13	936	0.24
Physics	3,182	541	0.17	480	0.15	1,021	0.32
Sciences of the Universe	2,202	341	0.15	156	0.07	497	0.23
Engineering	6,441	467	0.07	1,188	0.18	1,655	0.26
Mathematics	2,646	267	0.10	157	0.06	424	0.16
Social Sciences	1,524	139	0.09	147	0.10	286	0.19
Total	31,149	3,133	0.10	4,812	0.15	7,945	0.26

Note: The total number of applications is 31,149 instead of 31,081 because of the multiple allocations performed when the section relates to several discipline. The number of applications in Social Sciences is low considering we excluded some Human and Social Sciences disciplines from the analysis.

In Figure 5 we investigate the rate of participation (number of applications and number of awards per capita) at the section level (which corresponds mainly to a subdiscipline and an employing in-

Figure 5: Intensity of the participation in thematic and non-thematic programs at the specialties level (for sections with more than 25 researchers)



stitution). We find a linear relationship between the rate of applications and the rate of funding, for both thematic (top-right) and non-thematic (top-left) programs. Some sections benefit from small positive bias in terms of success rate (points that are on the left of the non-represented fitted straight line that could be drawn). Most are CNRS sections for non-thematic programs and INSERM/INRA/INRIA sections for thematic programs. When we consider the joint participation rates of sections in the two types of programs (bottom graphs of Figure 5), the results vary significantly depending on the sections. Some sections favor a particular type of program, while others indicate a fairly balanced participation between the thematic and non-thematic programs (both for applications and fundings).

Lastly, Figure 6 depicts histograms of the size of laboratories, in terms of the number of tenured researchers or professors, in the three samples. In the majority of cases, these academics' laboratory staff is made up of 10 to 70 employees, while some of them exceed 200 tenured staff members. There is no obvious difference in size between the three samples' distributions.

Figure 6: Histogram of the size of the laboratories (number of tenured researchers or professors) in the three samples

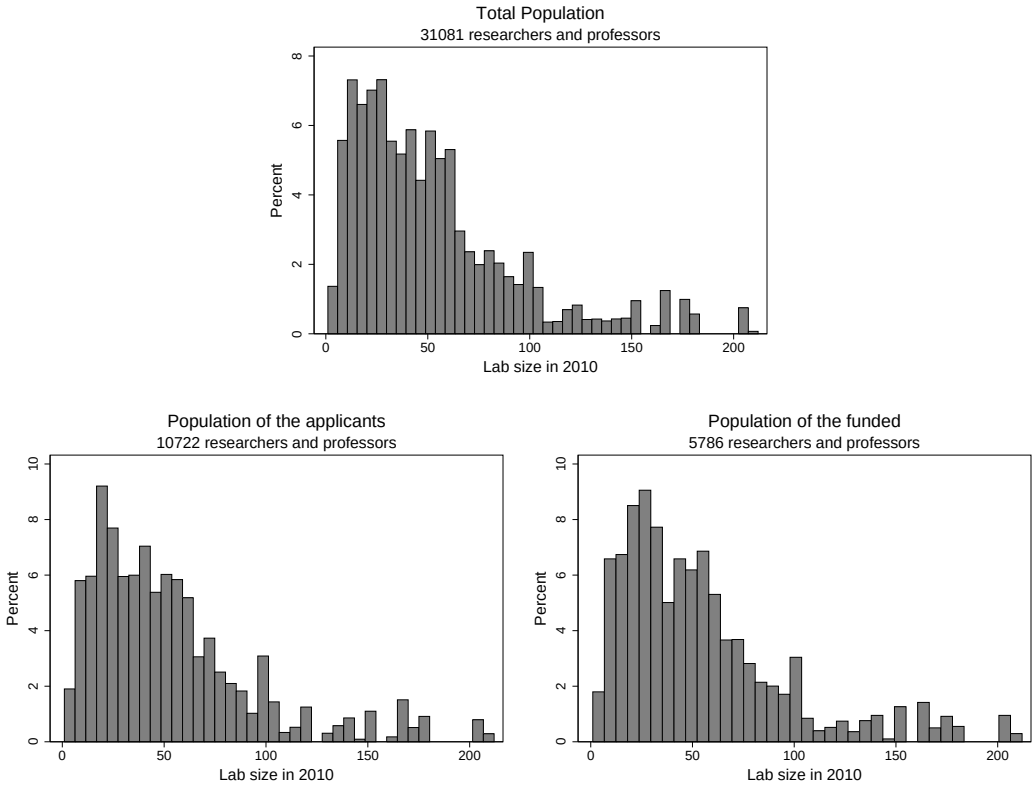


Table 10: List of sections assigned to our final sample of researchers, according to the research institute

Section	
CNRS-1	Mathématiques et interactions des mathématiques
CNRS-10	Milieux fluides et réactifs : transports, transferts, procédés de transformation
CNRS-11	Systèmes supra et macromoléculaires : propriétés, fonctions, ingénierie
CNRS-12	Architectures moléculaires : synthèses, mécanismes et propriétés
CNRS-13	Physicochimie : molécules, milieux
CNRS-14	Chimie de coordination, interfaces et procédés
CNRS-15	Chimie des matériaux, nanomatériaux et procédés
CNRS-16	Chimie du vivant et pour le vivant
CNRS-17	Système solaire et univers lointain
CNRS-18	Terre et planètes telluriques : structure, histoire, modèles
CNRS-19	Système Terre : enveloppes superficielles
CNRS-2	Théories physiques : méthodes, modèles et applications
CNRS-20	Surface continentale et interfaces
CNRS-21	Bases moléculaires et structurales des fonctions du vivant
CNRS-22	Organisation, expression et évolution des génomes
CNRS-23	Biologie cellulaire : org et fonc de la cellule, pathogènes et rel hôte/pathogène
CNRS-24	Interactions cellulaires
CNRS-25	Physiologie moléculaire et intégrative
CNRS-26	Développement, évolution, reproduction, vieillissement
CNRS-27	Comportement, cognition, cerveau
CNRS-28	Biologie végétale intégrative
CNRS-29	Biodiversité, évolution et adaptations biologiques
CNRS-3	Interactions, particules, noyaux du laboratoire au cosmos
CNRS-30	Thérapeutique, médicaments et bio-ingénierie : concepts et moyens
CNRS-37	Économie et gestion
CNRS-4	Atomes et molécules, optiques et lasers, plasmas chauds
CNRS-5	Matière condensée : organisation et dynamique
CNRS-6	Matière condensée : structures et propriétés électroniques
CNRS-7	Sciences et technologies de l'information
CNRS-8	Micro et nano-technologies, élec, photo, électroma, énergie élec
CNRS-9	Ingénierie des matériaux et des structures, mécaniques de solides, acous

Table 10 Continued

Section	
CNU-16	Psychologie, psychologie clinique, psychologie sociale
CNU-25	Mathématiques
CNU-26	Mathématiques appliquées et applications des mathématiques
CNU-27	Informatique
CNU-28	Milieux denses et matériaux
CNU-29	Constituants élémentaires
CNU-30	Milieux dilués et optique
CNU-31	Chimie théorique, physique, analytique
CNU-32	Chimie organique, minérale, industrielle
CNU-33	Chimie des matériaux
CNU-34	Astronomie, astrophysique
CNU-35	Structure et évolution de la terre et des autres planètes
CNU-36	Terre solide : géodynamique des enveloppes supérieure, paléobiosphère
CNU-37	Météorologie, océanographie physique de l'environnement
CNU-39	Sciences physico-chimiques et technologies pharmaceutiques
CNU-40	Sciences du médicament
CNU-41	Sciences biologiques
CNU-42	Morphologie et morphogenèse
CNU-43	Biophysique et imagerie médicale
CNU-44	Biochimie, biologie cellulaire et moléculaire, physiologie et nutrition
CNU-45	Microbiologie, maladies transmissibles et hygiène
CNU-46	Santé publique, environnement et société
CNU-47	Cancérologie, génétique, hématologie, immunologie
CNU-48	Anesthésiologie, réanimation, médecine d'urgence, pharmaco et thérapeutique
CNU-49	Pathologie nerveuse et musculaire, pathologie mentale, handicap et rééducation
CNU-5	Sciences économiques
CNU-50	Pathologie ostéo-articulaire, dermatologie et chirurgie plastique
CNU-51	Pathologie cardiorespiratoire et vasculaire
CNU-52	Maladies des appareils digestif et urinaire
CNU-53	Médecine interne, gériatrie et chirurgie générale
CNU-54	Développement et pathologie de l'enfant, gynéco-obsté, endocrino et reprod
CNU-55	Pathologie de la tête et du cou
CNU-56	Développement, croissance et prévention
CNU-57	Sciences biologiques, médecine et chirurgie buccales
CNU-58	Sciences physiques et physiologiques endodontiques et prothétiques

Table 10 Continued

Section	
CNU-60	Mécanique, génie mécanique, génie civil
CNU-61	Génie informatique, automatique et traitement du signal
CNU-62	Energétique, génie des procédés
CNU-63	Génie électrique, électronique, photonique et systèmes
CNU-64	Biochimie et biologie moléculaire
CNU-65	Biologie cellulaire
CNU-66	Physiologie
CNU-67	Biologie des populations et écologie
CNU-68	Biologie des organismes
CNU-69	Neurosciences
CNU-85	Pharmacie en sciences physico-chimiques et ingénierie appliquée à la santé
CNU-86	Pharmacie en sciences du médicament et des autres produits de santé
CNU-87	Pharmacien sciences biologiques, fondamentales et cliniques
INRA-1	Biologie fondamentale
INRA-2	Médecine
INRA-3	Biologie/Ecologie appliquée
INRA-4	Chimie
INRA-6	Science de l'Univers
INRA-8	Mathématiques
INRA-9	Sciences humaines et sociales
INRIA	Sciences de l'ingénieur et mathématiques
INSERM-CSS1	Bases biochimiques, moléculaires et structurales du vivant
INSERM-CSS2	Génétique, épigénétique, cancérologie
INSERM-CSS3	Biologie cellulaire, développement, vieillissement
INSERM-CSS4	Physiologie et physiopathologie des syst card, vasc, pulm, néphro et musc
INSERM-CSS5	Physiologie et physiopathologie des systèmes endoc, dig, ostéo-artic et cut
INSERM-CSS6	Neurosciences, cognition, santé mentale
INSERM-CSS7	Microbiologie, immunologie, infection
INSERM-CSS8	Technologies pour la santé, thérapeutiques, biotechnologies
INSERM-CSS9	Santé publique, santé des populations : épidémio, biostat, shs
IRD-CSS1	sciences physiques et chimiques de l'environnement planétaire
IRD-CSS2	sciences biologiques et médicales
IRD-CSS3	sciences des systèmes écologiques

Appendix B. Outcome variables

In this section, we present how we built the different outcome variables used in the analysis.

1 Production variables

- Yearly number of contributions to articles published in WoS journals, with each paper being weighted by the inverse of the number of its authors:

$$VC_i^t = \sum_{j \in J_t} \frac{1\{i \Leftarrow j\}}{n(j)}, \tag{1}$$

where J_t denotes the set of published paper in year t, $1\{.\}$ is the indicator function equal to one if the condition into brackets is verified and zero otherwise, the expression "$i \Leftarrow j$" means i is the author of j and $n(j)$ denotes the author number of the article j. In the main paper and in the tables, we refer to this variable as the ***Volume***.

- Yearly number of articles published in WoS journals, with each paper being adjusted by the impact factor of the journal and by the inverse of the number of authors:

$$IFC_i^t = \sum_{j \in J_t} \frac{1\{i \Leftarrow j\} \times IF(j)}{n(j)}, \tag{2}$$

where $IF(j)$ denotes the (three-years) impact factor of the journal where publication j has been published. In the main article and in the tables, we refer to this variable as ***Impact Factor***.

- Yearly number of articles published in WoS journals, with each article being adjusted by the number of citations in the three-year moving window $(t, t+2)$ and by the inverse of the number of coauthors:

$$CITC_i^t = \sum_{j \in J_t} \frac{\{i \Leftarrow j\} \times C(j)}{n(j)}, \tag{3}$$

where $C(j)$ denotes the number of citations received by article j from articles published in the three-year moving window $(t, t+2)$. In the main body of the article and in the tables, we refer to this variable as ***Citations***.

2 Other outcome variables

- Mean number of authors by article in a given time period τ:

$$COA_i^\tau = \frac{\sum_{j \in J_\tau} 1\{i \Leftarrow j\} \times n(j)}{\sum_{j \in J_\tau} 1\{i \Leftarrow j\}}. \tag{4}$$

- Number of distinct coauthors recorded for the considered individual COD_i^τ.

- Number of new coauthors: number of distinct coauthors observed in period τ who did not appear among previous coauthors if i: $CODN_i^\tau$.

- Number of international collaborations: number of published articles with at least one author with a professional address located outside France for a given period τ:

$$INT_i^\tau = \sum_{j \in J_\tau} 1\{i \Leftarrow j\} \times 1\{j \leftarrow inter\}, \tag{5}$$

where "$j \leftarrow inter$" means that paper j results from an international collaboration.

- The novelty of i's scientific production is given by

$$NOV_i^\tau = 90th\left(n_{kct} \,|(k,c) \in K_i^\tau, t \in \tau\right), \tag{6}$$

where $90th\,(.)$ gives the 90th percentile of the distribution characterized into parentheses and with K_i^τ the set of keywords×fields obtained by the articles published by agent i in a given period of time τ. n_{kct} is the novelty of keyword k, in field c and at year t. It is defined as:

$$n_{kct} = -log\frac{N_{kct}}{N_{ct}}, \tag{7}$$

where N_{ct} denotes the number of (non-distinct) keywords to be found in the articles published in field c and year t and N_{kct} be the number of times the keyword k is is actually used in that year t and field.

Appendix C. Specification of the selection model

We first discuss the basic principles used to build the selection model, before presenting the eight specifications that are retained. All these specifications will be compared in the next section.

Principles

We consider two different sets of persons in which to pick controls:

- the first set consists of all the researchers and professors in our whole cleaned data set who did not get an ANR grant in the period 2005-2009, that is 25,295 persons (31,081 researchers and professors, of which 5,786 received a grant);
- the second one is a subset of the first set, which consists of researchers and professors who applied to an ANR call for proposal between 2005 and 2009, but received no funding. It comprises 4,936 persons (10,722 applicants,of which 5,786 received a grant).

Although the second group size is much smaller compared to the first one, its members are characterized by the same self-selection in terms of applying for a grant as those who were successful in doing so. Moreover, these individuals (as with the grant recipients) could have been subject to variations in their performance before applying in order to increase their chances of being selected. If researchers increase the number of authored publications before the application date, the use of the first set as control group (all non-recipients of funding) would underestimate the mean effect of funding (when using difference-in-differences method). The second group (applicants), however, has the disadvantage in terms of offering much less potential controls. Hence, non-applicants can possibly be depicted as better controls than unsuccessful grant applicants.Two types of information can be used to explain the selection process:

- Individual variables. Personal characteristics of the researchers, observed at the date of the application are likely to influence the selection of the project by the ANR, as well as future scientific production. Age is well known to affect scientific production over the career path. Scientific production first increases before eventually decreasing later in some fields. Since it is also likely to affect selection into treatment, we thus use the age of the researcher, together with the squared age to capture a possible non linear effect. We also consider several production measures built from publication data to account for scientific activity, impact and audience. We use the number of articles published in the three previous years to account for the intensity of the recent research effort, the number of citations received in the same period to control for the recent impact of the authors' research in the recent past, the maximum impact factor of the journals in the same period to consider the ability to publish in large audience journals and the number of citations received over a longer period (recorded from 1999) in order to account for the long-term scientific reputation. Finally, we introduce, in some specifications, the production variation before the application year.[6] It is intended to account for the scientific production dynamics just before funding, while all the other publication variables explaining the treatment are averaged over the previous years.
- Laboratory variables. Given that the research environment quality is explicitly examined in the ANR evaluation process, laboratory attributes are likely to affect the selection of the applications (as well as the propensity to apply). They also influence the production outcomes (see for instance Carayol and Matt, 2006). We select variables that measure the intensity of scientific production, the reputation in terms of citations at the laboratory level and the size

[6]These variables are calculated by taking the difference of the production measures in the level between $t-3$ and $t-1$, with t as the application date.

of the laboratory. These variables are not included in the first specifications of the selection model because they correspond to the configuration of the laboratories in year 2010. Though mobility is limited, laboratory memberships could have changed since the application year. In theory, the model should not include covariates observed after the application date because it may bias our estimates. For instance, a grant recipient could have moved between the grant awarding year and 2010. Indeed, the recipient may now be member of a laboratory with better performance than the one he was affiliated to at the date of the application, either because the laboratory was able to employ new staff as a consequence of the grant or because the funding influenced the mobility of the recipient, which could increase the weight given to controls affiliated to laboratories with better quality. If this frequently occurs, it could result in an underestimation of the mean effect of the grant (because controls are selected in relation to better quality laboratories). However, as shown above, the inclusion of these laboratory variables does not affect our results significantly.

Some other additional relevant covariates are also considered. We use them in various forms (exact matching or explanatory covariates).

- Scientific fields. Given that the study covers scientific fields with heterogeneous publication profiles, we investigate whether the regression has to be implemented by scientific specialty. For this purpose, we investigate an exact matching with the section that also allows us to control the employer type and employment type (professor or researcher). This comes down to considering whether the conditions of selection can change from a specific section to another one. It ensures that a grant recipient will never have a control from a different field. This, however, implies a reduced set of treated and controls in each model. Some sections count a very limited number of members and thus do not have enough treated or controls left to obtain consistent coefficients in the logit regression. Another disadvantage of performing an exact matching on the section variable is that the implementation turns out to be very complicated, given the large number of sections involved. Therefore, in some models, we rather perform exact matching on aggregated, thematically close sections.

- Program type. The selection process can actually follow slightly different logics according to the type of program considered. In particular, the selection processes of the thematic and non-thematic programs may differ. An exact matching with the program type may allow us to consider different weighting schemes of the ANR selection process, according to the type of program.

- Application year. The process for allocating ANR grants has not necessarily been the same across the years, especially in a context of the gradual establishment of the ANR. In particular, 2005, the first year of activity of the ANR, is characterized by a much higher selection rate than other years.

The selection models

We now present here the eight different specifications that we selected (denoted PSM1-PSM8) and applied to the selection model. See Table 11 for an overview and Tables 12 to 14 for a detailed description of all models.

PSM1 The specification of the model includes individual covariates, which influence both the selection process and our outcome variables, such as age, as well as some measures relating to the scientific production in the three previous years (number of publications adjusted for authorship, number of citations received and the maximum impact factor of the journal). The propensity score is estimated by exact matching in terms of the section and the year of an ANR program. The control

group is the whole set of French researchers who did not receive a grant from the ANR during the period 2005-2009, that is, 25,390 researchers were observed for each year. We do not consider all sections×years with less than five funded researchers. For some sections×years where the maximum likelihood algorithm of the logit regression does not converge, we discard these groups from the analysis. We decided not to apply any modification in the specification for each model in order to avoid introducing any uncontrolled bias.

PSM2 The specification is similar to PSM1, with the inclusion of additional explanatory covariates related to the laboratory of each researcher.

PSM3 The control group is limited to the subset of the 4,936 unsuccessful ANR applicants.[7]We assume that this control group allows us to control the self-selection bias (the decision to respond to an ANR call). As the size of the control group is severely reduced, this specification is no longer based on exact matching on section×year. Instead, we aggregate sections on a disciplinary basis (see Table 16 for a description of the sections grouped together into disciplines).

PSM4 The specification is similar to PSM3, with the inclusion of explanatory variables related to the laboratory of the researcher.

PSM5 We assume here that the ANR selection process can be driven by various determinants, according to the type of program. Instead of using the same covariate specification for each group (formed according to section×year or large disciplines), we use a different specification of the model for thematic and non-thematic programs. This is also explained by the difficulty in finding a uniquely good specification for both program types. Compared with the previous specifications, some continuous covariates (such as the production measures) are transformed into categorical covariates. Information related to the laboratories are not included in the set of explanatory covariates. The discipline is represented by grouping sections into large fields (see groups of sections in Table 16).

PSM6-PSM8 These specifications are analogous to PSM3-PSM5, with the introduction of the additional "trend" variables. We define two measures of the production evolution before the year of application. The first one refers to the difference in the level of production between $t-3$ and t (or $t-1$), whereas the second one refers to the growth rate (percent variation) of the output between $t-3$ and $t-1$.

$$var\_X\_bef = X_{t-1} - X_{t-3}$$

$$var1\_X\_bef = X_t - X_{t-3}$$

$$\%\_var\_X\_bef = \frac{X_{t-1} - X_{t-3}}{X_{t-3}}$$

Where X denotes one of the three production resumes (volume, citations or impact Factor). The additional trend covariates used in the PSM6-PSM8 models are:

- $var\_citations\_bef$ in PSM6
- $\%\_var\_citations\_bef$ in PSM7
- $var1\_IF\_bef$ and $\%\_var\_IF\_bef$ in PSM8 non-thematic programs

[7]The control group is built using the unsuccessful ANR applicants' subset from PSM3 to PSM8.

- $var1\_art\_bef$ and $\%\_var\_citations\_bef$ in PSM8 non-thematic programs

Table 11: Synthesis of PSM estimations

	PSM1	PSM2	PSM3	PSM4	PSM5	PSM6	PSM7	PSM8
Restriction on the controls								
all researchers without funding	X	X						
only applicants without funding			X	X	X	X	X	X
Exact matching								
section	X	X						
field & research institute			X	X		X	X	
theme of the program					X			X
Covariates explaining the treatment								
individual covariates	X	X	X	X	X	X	X	X
laboratory covariates		X		X	X		X	X
trend covariates						X	X	X

Table 12: List of covariates used for the propensity score estimation in the PSM1-PSM4 models

individual covariates	Description
age	age of the individual at the time of the application
age2	age squared of the individual at the time of the application
articles_bef3	number of articles published during the three previous years before the application (adjusted for coauthorship)
citations_bef3	number of citations received during the three previous years before the application (adjusted for coauthorship)
maxIF_bef3	maximum impact factor of published articles during the three previous years before the application
d_citations	1 if the individual belongs to the top 10% in terms of the number of citations adjusted (from 1999) by section at the date of the application, 2 belongs to the following 20%, 3 if in the next 30% and 4 if in the next 40%
laboratory covariates	
nber_art_lab	mean number of articles (by researcher) published by the laboratory members during the three previous years before the application year
max_cit_lab	number of articles (adjusted by the impact factor) published by the top member of the laboratory during the three previous years before the application year
d_size_lab	dummy for the number of researchers in the laboratory in 2013; 1 if the laboratory belongs to the top 25% of the biggest laboratories in France, 2, 3 and 4 for the following quartiles
covariates used for exact matching	
section	section related to the classification of the research institute at which the researcher is affiliated
year	year of the ANR selection of project recipients
discipline_gr1	grouping of sections of a similar field given the classification of the research institute (cf table 3)

Table 13: List of covariates used for the propensity score estimation in the PSM5 model (non-thematic-programs)

Non-thematic programs	
covariates	Description
age	age of the individual at the time of the application
d_articles_bef3	1 if the individual belongs to the top 10% of the field in terms of the number of publications adjusted for coauthorship (during the three previous years), 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%.
d_citations_bef3	1 if the individual belongs to the top 10% of the field considering the number of citations adjusted for coauthorship (during the three previous years), 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%.
d_maxIF_bef3	1 if the individual belongs to the top 10% of the field considering the maximum impact factor of the journal (during the three previous years), 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%
d_citations	1 if the individual belongs to the top 10% of the field in terms of the number of citations adjusted for coauthorship (from 1999) at the date of the application, 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%
discipline_gr2	large disciplines (cf table 4)
institute	research institute where the researcher is affiliated (CNRS-UNIV-INRIA-INRIA-IRD-INSERM)

Table 14: List of covariates used for the propensity score estimation in the PSM5 model (thematic programs)

	Thematic programs
covariates	Description
age	age of the individual at the time of the application
d_articles_bef3	1 if the individual belongs to the top 10% of the field in terms of the number of publications adjusted for coauthorship (during the three previous years), 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%.
d_citations_bef3	1 if the individual belongs to the top 10% of the field considering the number of citations adjusted for coauthorship (during the three previous years), 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%.
d_maxIF_bef3	1 if the individual belongs to the top 10% of the field considering the maximum impact factor of the journal (during the three previous years), 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%.
d_citations	1 if the individual belongs to the top 10% of the field in terms of the number of citations adjusted for coauthorship (from 1999) at the date of the application, 2 if belonging to the following 20%, 3 if in the following 30% and 4 if in the following 40%
program theme	theme of the thematic program (Biology & Health, Sustainable Energy, ...)
year	year of the application
prog*year	interaction between the program theme and the application year

Table 15: Groups of sections of a similar field, given the classification of the research institute (used in PSM3 and PSM4 through *discipline_gr1*)

Groups of sections (by research institute)
CNU-25 -26
CNRS-11 -12 -13
CNU-37 -35 -36
CNRS-23 -20 -21
CNRS-26 -25 -27 -24 -28
CNU-68 -65 -66 -41
INSERM-CSS8 -CSS7
CNU-40 -39
CNU-52 -43 -45 -57 -56 -50 -46 -44 -53 -49 -51 -54 -42 -47 -55 -48 -58
INSERM-CSS1 -CSS3 -CSS6 -CSS5
CNRS-38 -31
CNRS-4 -2 -3
CNU-29 -30
CNRS-40 -36
CNU-4 -3 -1 -2
CNU-5 -6
CNRS-39 -31
CNU-23 -24
CNU-7 -71
CNU-73 -13 -14 -15 -10 -8 -12 -9 -11
CNU-76 -18 -17 -72 -77
CNRS-5 -6
CNRS-17 -15
CNRS-18 -16
CNU-27 -61

Table 16: Groups of sections of a similar field (used in PSM5 through *discipline_ gr2*)

Discipline	Groups of sections
Life sciences	CNRS-20 -21 -22 -23 -24 -25 -26 -27 -28 -29 -30 ; CNU-39 -40 -41 CNU-64 -65 -66 -67 -68 -69 INRA-1 -3 INSERM-CSS2 -CSS4 -CSS7 -CSS8
Medical research	CNU-42 -43 -44 -45 -46 -47 -48 -49 -50 -51 -52 -53 -54 -55 -57 -58 -85 -86 -87 INRA-2 INSERM-CSS1 -CSS3 -CSS5 -CSS6
Chemistry	CNRS-15 -16 -17 -18 -19 CNU-31 -32 -33 INRA-4
Physics	CNRS-2 -3 -4 -5 -6 CNU-28 -29 -30
Universe science	CNRS-11 -12 -13 -14 CNU-34 -35 -36 -37 INRA-6
Engineering	CNRS-10 -9 CNU-60 -62 INRIA
Mathematics	CNRS-1 CNU-25 -26 INRA-8
ICST	CNRS-7 -8 CNU-27 -61 -63
Human & social sciences	CNRS-37 CNU-16 -5 INRA-9 INSERM-CSS9
Others (IRD)	IRD-CSS1 -CSS2 -CSS3

References

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

Appendix D. A parallel path test before treatment

The conditional difference-in-difference model is valid if the parallel trend assumption is verified. It states that the outcome variable for the treated should have experienced (after the treatment date) the same progress on average, in the absence of treatment, as the controls who have the same probability of assignment into treatment $p(x)$. It can be written as follows:

$$E(Y_{t+\tau} - Y_t|T = 1, P(X)) = E(Y_{t+\tau} - Y_t|T = 0, P(X)), \tag{8}$$

where Y is the outcome variable observed at the year of application t and, in a later year, at $t + \tau$, while T denotes the decision of the ANR to select the project or not and $P(X)$ is the propensity score. The parallel path assumption in Equation (7), however, cannot be tested directly because the counterfactual outcome of the treated is not available. That said, we can compare the outcome paths of the treated and the controls before treatment. That is we set up a parallel path test on the period before the attribution of the treatment. We assume that individuals who follow parallel trajectories right before the assignment are also likely to share parallel paths afterwards (all other factors being equal). Our objective is to check whether the production difference between $t - 3$ and t is significantly different (in weighted means) between the treated and the controls for each specification of the selection model (PSM1-PSM8) and for each matching method. The test is based on a difference-in-differences model before the application year. We want to check whether the variation in outcomes during the three years before the selection of grant recipients (between $t - 3$ and $t - 1$) is significantly different between the controls and those who received grants. If the results show a significant difference, it would disprove our assumption of a parallel trend between controls and treated. The results are presented in Table 17. A robustness check is presented in Table 18 in which we compare outcomes between $t - 3$ with t.

The main results of the parallel path tests between $t - 3$ with $t-1$ are the following:

- Only PSM1 and PSM2 specifications do not pass the tests (a significant difference of trajectories between treated and controls).[8]
- PSM3 to PSM8 exhibit very weak and insignificant differences in the production dynamics between groups for the three outcome measures.[9]

We then repeat the tests by comparing the outcomes of year $t - 3$ with the outcomes of year t.

- Only PSM8 returns non-significant differences between groups, whatever the technique used to form the control group.
- PSM5 specification also passes the test when implemented using the five nearest neighbors technique.

Although it is complicated to order the different models according to the quality of the results obtained from the test, we can assert that the PSM8 specification provides the most relevant estimation, as, for any of its weighting schemes and for any one of the three outcome measures, the parallel path hypothesis before treatment is never violated.

[8]The difference is also significant for the PSM3 and PSM4 specifications of the IPTW model, when the citations measure is the outcome.

[9]Only PSM3 and PSM4 in the IPTW method exhibit significant differences.

Table 17: Parallel path test : Difference-in-differences estimates of the mean effect of treatment on various production variables (calculated from $t-3$ to $t-1$)

	δ^{5nn}		δ^{kernel}		δ^{iptw}	
	PSM1	PSM2	PSM1	PSM2	PSM1	PSM2
Volume	.00262	.00592	.00827	.00770	.01265**	.01098**
	(0.48)	(1.03)	(1.62)	(1.44)	(2.34)	(2.05)
Citations	.01915	.02959**	.02533**	.03738***	.03446***	.03269***
	(1.55)	(2.36)	(2.18)	(3.17)	(2.86)	(2.78)
Impact Factor	.01941**	.01777.**	.01788**	.02075**	.01665**	.01754**
	(2.31)	(2.02)	(2.27)	(2.52)	(2.16)	(2.18)

	δ^{5nn}			δ^{kernel}			δ^{iptw}		
	PSM3	PSM4	PSM5	PSM3	PSM4	PSM5	PSM3	PSM4	PSM5
Volume	.00333	.00202	-.00650	.00533	.00353	-.00700	.00723	.00969	-.00518
	(0.48)	(0.29)	(-0.95)	(0.83)	(0.53)	(-1.1)	(1.06)	(1.17)	(-0.82)
Citations	.02358	.01399	-.00451	.01862	.0137	.00291	.02546*	.02523*	.00233
	(1.57)	(0.93)	(-0.30)	(1.31)	(0.92)	(0.21)	(1.73)	(1.68)	(0.17)
Impact Factor	.00781	-.00131	-.00571	.00496	-.00257	-.00192	.00595	.00106	-.00627
	(0.77)	(-0.13)	(-0.65)	(0.51)	(-0.26)	(-0.24)	(0.61)	(0.10)	(-0.78)

	δ^{5nn}			δ^{kernel}			δ^{iptw}		
	PSM6	PSM7	PSM8	PSM6	PSM7	PSM8	PSM6	PSM7	PSM8
Volume	.00074	.00442	-.00954	.00108	.00472	-.00809	.00372	.00742	-.00727
	(0.11)	(0.65)	(-1.39)	(0.17)	(0.72)	(-1.27)	(0.53)	(0.87)	(-1.13)
Citations	.01295	.00322	-.00454	.00719	.00328	-.0012	.0088	.01153	-.00268
	(0.86)	(0.21)	(-0.3)	(0.5)	(0.22)	(-0.09)	(0.58)	(0.75)	(-0.19)
Impact Factor	.00184	-.00989	-.00515	.00319	-.00925	-.00383	.00224	-.00148	-.00588
	(0.18)	(-0.96)	(-0.6)	(0.33)	(-0.94)	(-0.47)	(0.23)	(-0.14)	(-0.72)

*Note: Conditional difference-in-difference results. Dependent variables in Log. Robust standard errors in parentheses, clustered at the project level. Significance levels: 0.01: ***, 0.05: **, 0.10: *. Observations are weighted according to the inverse probability of treatment.*

Table 18: Parallel path test : Difference-in-differences estimates of the mean effect of treatment on various production variables (calculated from $t-3$ to t)

	δ^{5nn}		δ^{kernel}		δ^{iptw}	
	PSM1	PSM2	PSM1	PSM2	PSM1	PSM2
Volume	-.03863***	-.03716***	-.05787***	-.05582***	-.04238***	-.042***
	(-5.4)	(-5.01)	(-8.79)	(-8.13)	(-6.54)	(-6.33)
Citations	-.11242***	-.11238***	-.1183***	-.12586***	-.1108***	-.11078***
	(-7.64)	(-7.42)	(-8.55)	(-8.89)	(-8.09)	(-7.84)
Impact Factor	-.07165***	-.07123***	-.085***	-.08648***	-.0697***	-.06944***
	(-6.91)	(-6.65)	(-9.01)	(-8.65)	(-7.81)	(-7.28)

	δ^{5nn}			δ^{kernel}			δ^{iptw}		
	PSM3	PSM4	PSM5	PSM3	PSM4	PSM5	PSM3	PSM4	PSM5
Volume	-.01788**	-.00986	-.00336	-.0145*	-.01339	-.00207	-.01615*	-.00941	-.00320
	(-1.97)	(-1.02)	(-0.47)	(-1.7)	(-1.43)	(-0.32)	(-1.87)	(-1.06)	(-0.49)
Citations	-.05572***	-.03816*	-.02149	-.04503***	-.04739**	-.02437*	-.04416**	-.04537*	-.02464*
	(-3.03)	(-1.99)	(-1.4)	(-2.62)	(-2.54)	(-1.75)	(-2.54)	(-2.5)	(-1.78)
Impact Factor	-.02595**	-.01843	-.00817	-.02275*	-.02203	-.00657	-.02197*	-.00984	-.00492
	(-2.01)	(-1.38)	(-0.88)	(-1.88)	(-1.72)	(-0.78)	(-1.81)	(-0.69)	(-0.57)

	δ^{5nn}			δ^{kernel}			δ^{iptw}		
	PSM6	PSM7	PSM8	PSM6	PSM7	PSM8	PSM6	PSM7	PSM8
Volume	-.01875**	-.01241	.00078	-.01363	-.0139	-.00101	-.01202	-.00751	-.00139
	(-2.06)	(-1.31)	(0.11)	(-1.61)	(-1.54)	(-0.15)	(-1.34)	(-0.85)	(-0.21)
Citations	-.04165**	-.03541*	-.01868	-.03777**	-.03951**	-.02018	-.03229*	-.03657**	-.02246
	(-2.24)	(-1.85)	(-1.23)	(-2.16)	(-2.13)	(-1.45)	(-1.8)	(-1.98)	(-1.58)
Impact Factor	-.02285*	-.01457	-.00626	-.02355*	-.01771	-.004.	-.02059*	-.00773	-.00484
	(-1.73)	(-1.1)	(-0.69)	(-1.94)	(-1.4)	(-0.47)	(-1.67)	(-0.56)	(-0.56)

*Note: Conditional difference-in-difference results. Dependent variables in Log. Robust standard errors in parentheses, clustered at the project level. Significance levels: 0.01: ***, 0.05: **, 0.10: *. Observations are weighted according to the inverse probability of treatment.*

Appendix E. Balance diagnostics

We now present the balance tests applied to the eight specifications of the previously presented selection model (PSM1-PSM8). Such test (Austin, 2011) builds upon the idea that, if the CIA assumption (Equations 2-3) holds, treated and controls should share a similar distribution of their observables X after weighting. That is to say, for each level of the estimated propensity score, the distribution of the attributes X is conditionally independent of treatment status. If this balancing property is satisfied, i.e. covariates in X are balanced between treated and control subgroups for all propensity scores, then we can reliably assume that the conditional assignment into treatment is random. The difficulty in finding several treated and controls for each level of $p(X)$ makes this assumption untestable in this way. Instead, we first implement a balance test after weighting, without any conditioning of the propensity score. We next refine the test by defining intervals of the propensity score, in which the same test can be performed.

For each specification of the model and for each weighting technique used, we calculate the standardised difference (in %)(Austin (2009)), which stands for the remaining bias between groups, a follows:

$$standt.bias = \frac{\bar{x}_{T=1} - \bar{x}_{T=0}}{\sqrt{\left(s^2_{T=1} - s^2_{T=0}\right)/2}} \times 100, \tag{9}$$

where $\bar{x}$ and s^2 respectively denote the weighted mean and variance of the covariates among the treated ($T = 1$) and the controls ($T = 0$).[10]

Figures 7 to 12 report the distribution of the estimated propensity score (with density and box plot) for thematic programs and non-thematic programs in the PSM8 model, with the nearest neighbors, the kernel and the IPTW weighting methods. We observe that controls tend to have a lower probability to be treated compared to the real grant recipients. After weighting, the propensity score is similarly distributed between treated and controls, but even more strikingly with the nearest neighbors approach.

Then, going into more details, we look the balance of the covariates used in the selection model in Figures 13 to 18. The standardized difference is reported for each covariate in line, for thematic and non-thematic programs and for the two matching methods after matching. We see that the bias between groups has been severely reduced after matching, while the standardized difference is low for each specification (far below the 11% threshold usually retained in the literature). All the other specifications we present in this paper (PSM1-PSM8) satisfy such balancing properties.[11]

We next refine the balance test in terms of dividing the range of the estimated propensity scores into several strata where the balancing property holds (Rosenbaum & Rubin, 1984; Dehejia & Wahba, 1999; Austin, 2009). Following the algorithm used in Dehejia & Wahba (2002) , we proceed as follow:

- A limited number of intervals is chosen so that we find an equal mean value of the propensity score for the treated and control subsamples.
- We implement the covariates balance test in each previously defined stratum of $p(x)$. If the equality of the means of a covariate between the treated and control subsamples does not hold, we reduce the size of the interval or finally change the specification of the model (in introducing interaction terms, for example).

For each of the thematic and non-thematic specifications of the PSM8 model, we are able to divide the propensity scale into seven strata, into which all covariates are balanced.

[10] Equation (18) is used to calculate the standardised mean for a continuous variable. The calculation is slightly different when we refer to a categorical variable.

[11] Other balance tests are not presented due to space constraint, but are available upon request from the authors.

Figure 7: Density and box plot of the estimated propensity score before and after matching for the PSM8 thematic programs with the 5 nearest neighbors matching method

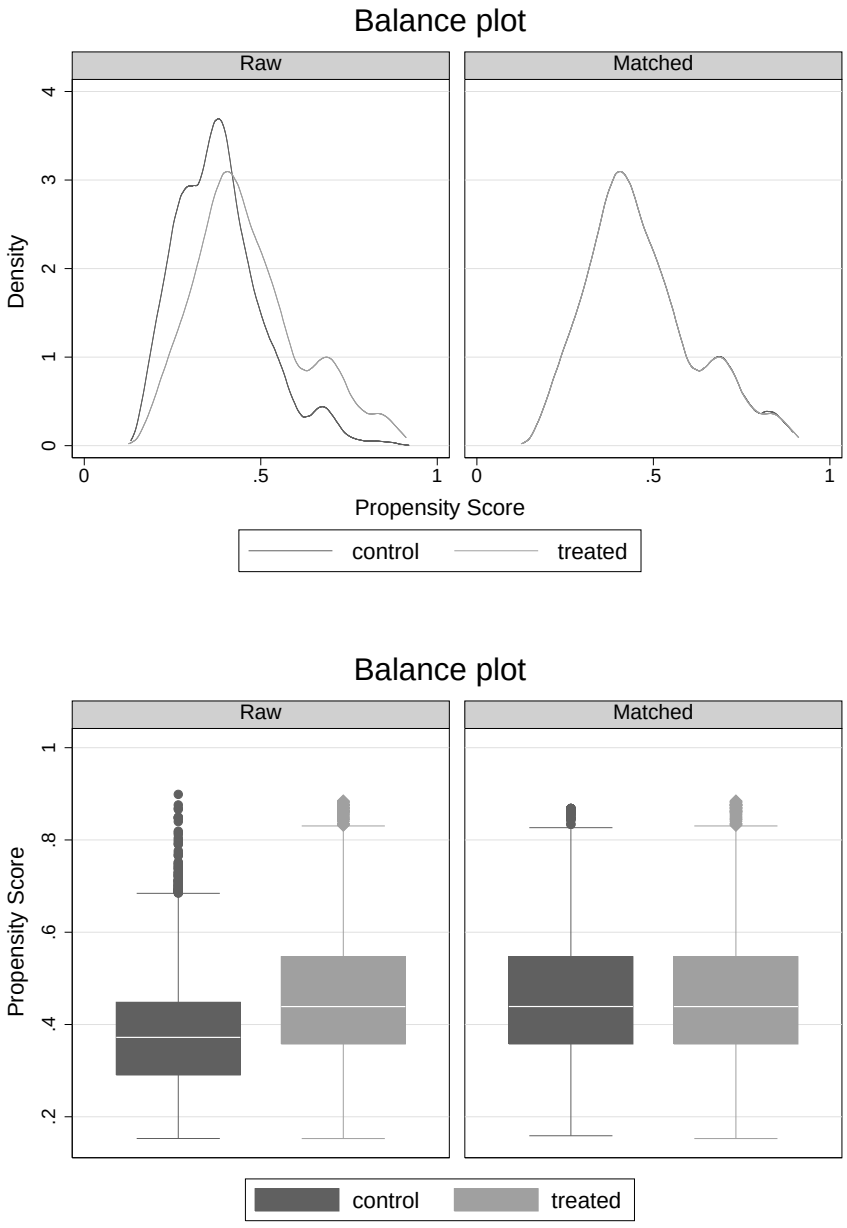
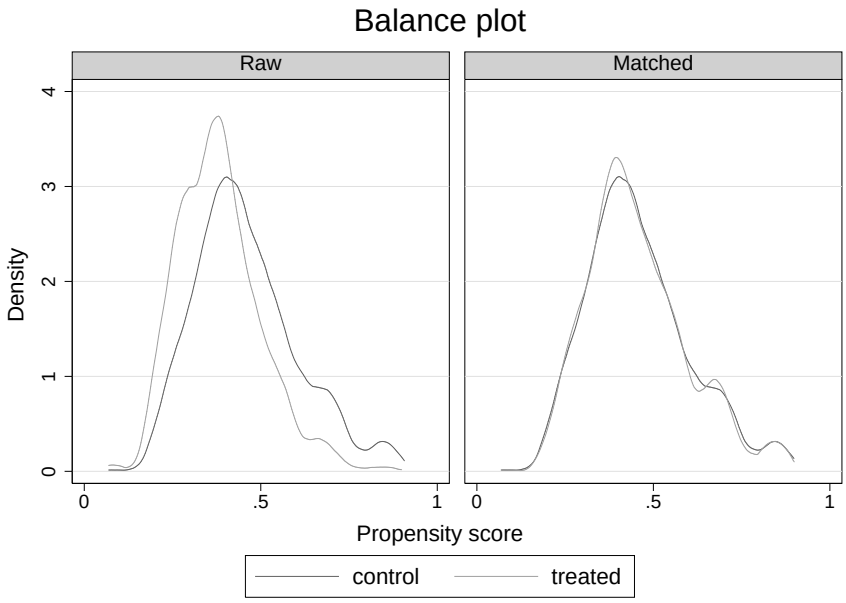


Figure 8: Density and box plot of the estimated propensity score before and after matching for the PSM8 thematic programs with the kernel matching method



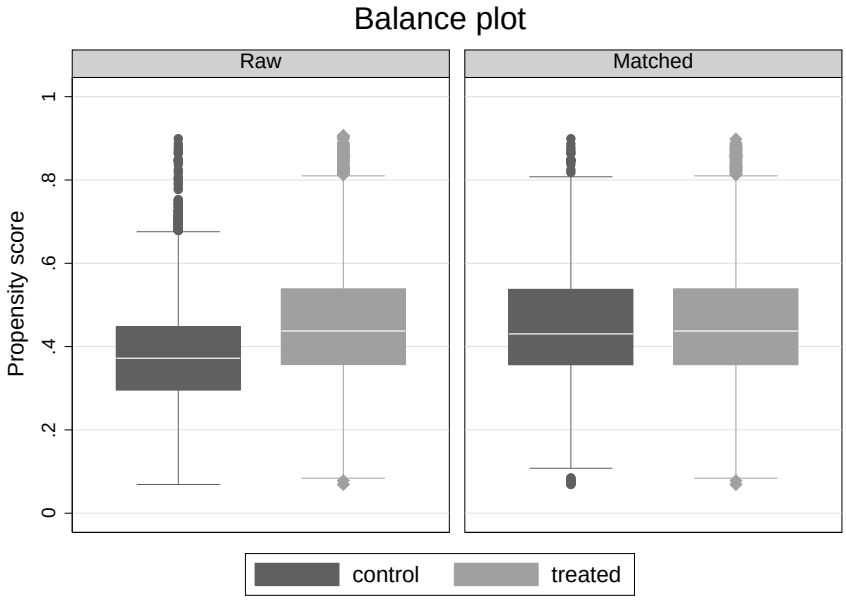
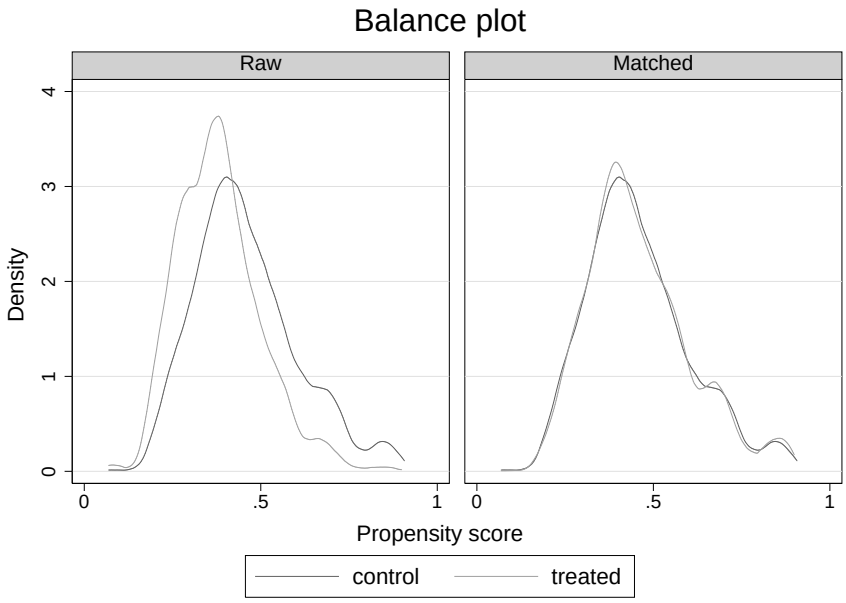


Figure 9: Density and box plot of the estimated propensity score before and after matching for the PSM8 thematic programs with the 5 IPTW matching method



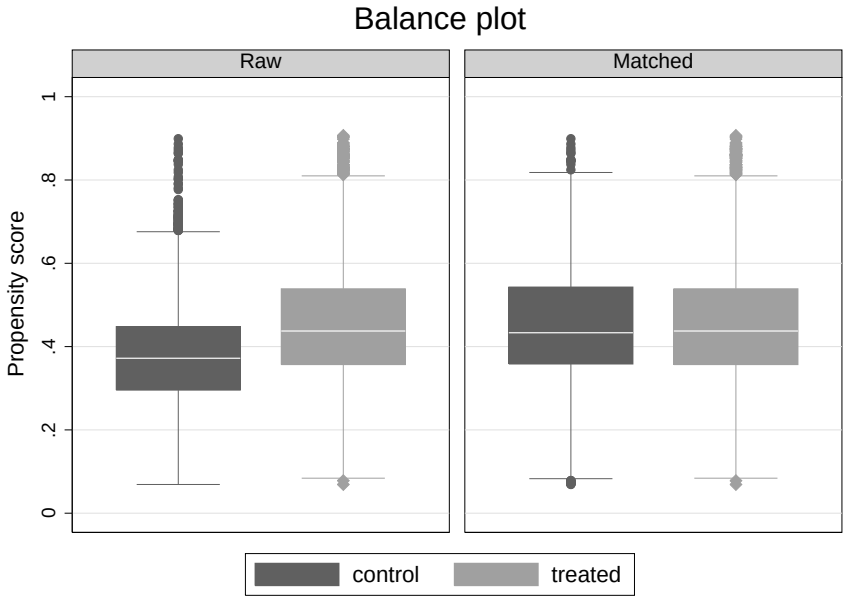


Figure 10: Density and box plot of the estimated propensity score before and after matching for the PSM8 non-thematic programs with the 5 nearest neighbors matching method

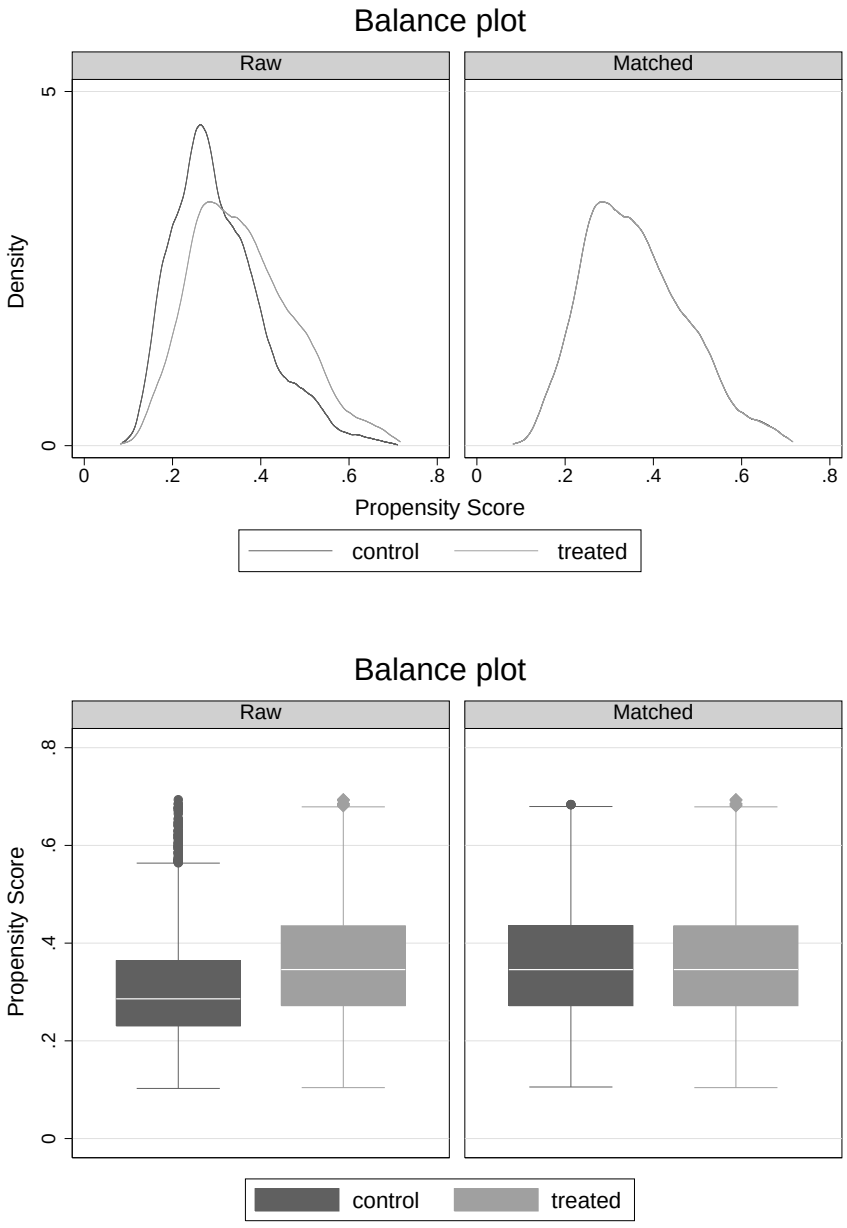
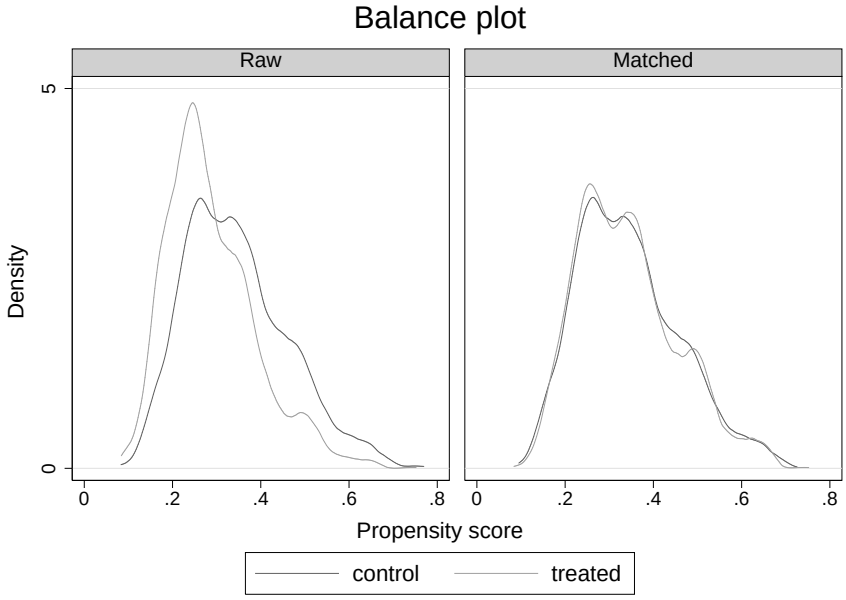


Figure 11: Density and box plot of the estimated propensity score before and after matching for the PSM8 non-thematic programs with the kernel matching method



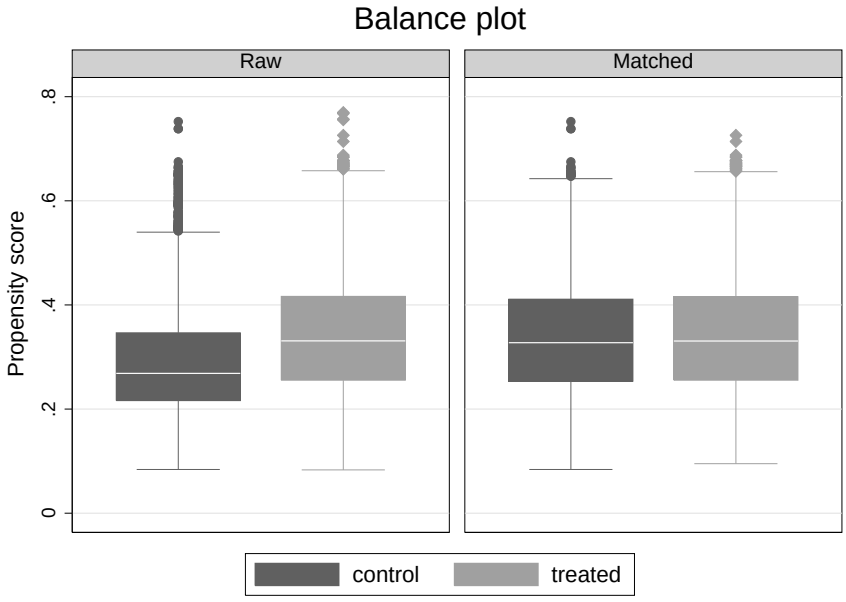
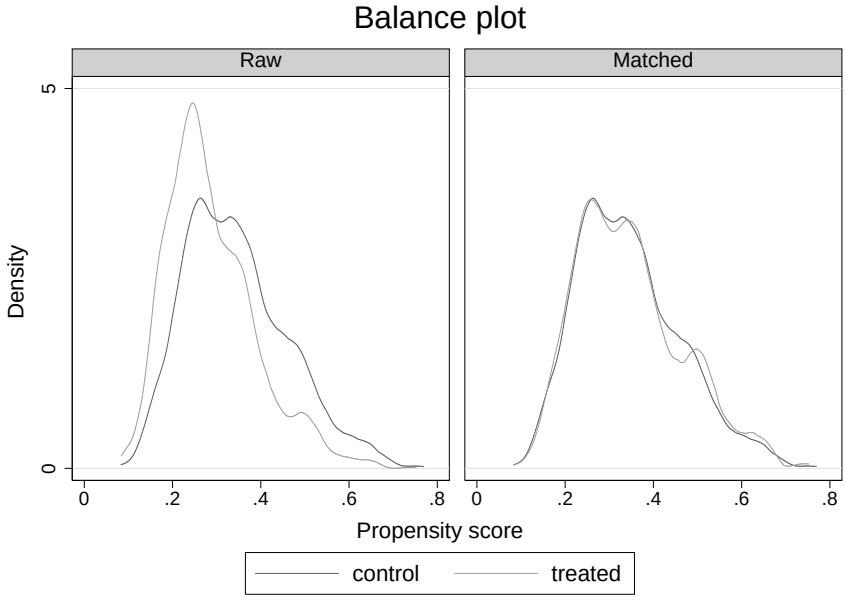


Figure 12: Density and box plot of the estimated propensity score before and after matching for the PSM8 non-thematic programs with the IPTW matching method



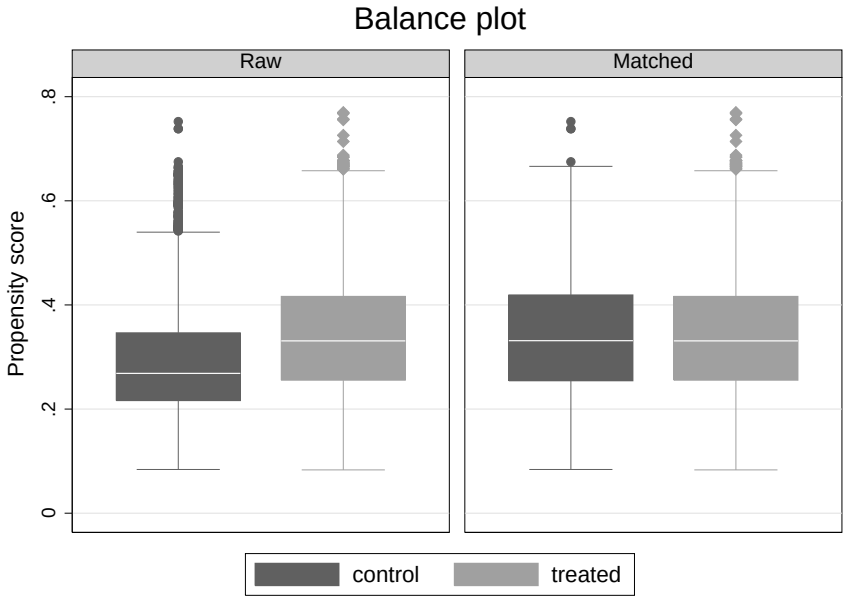
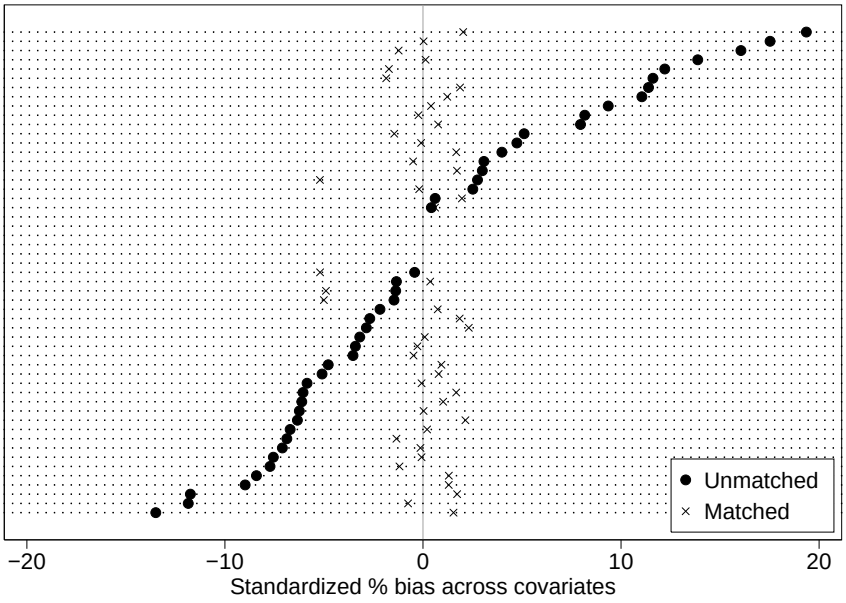
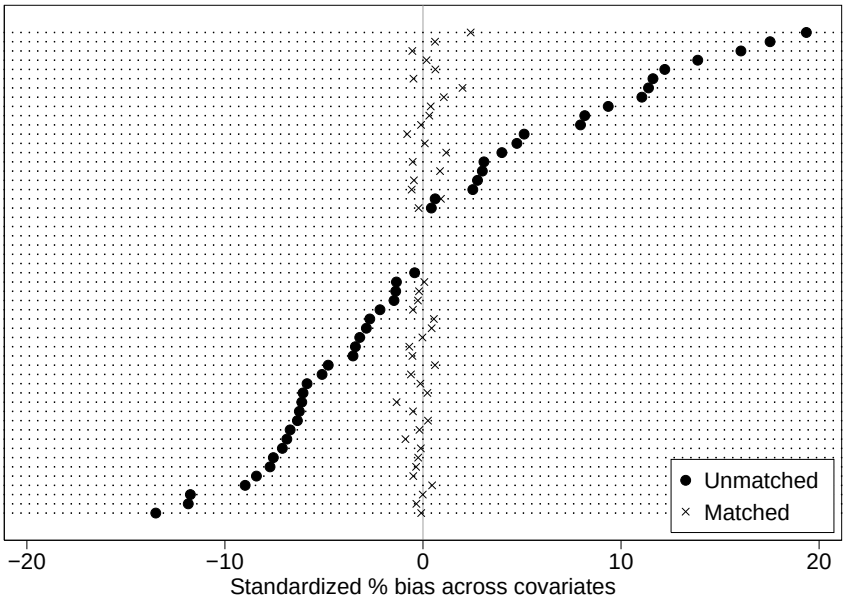


Figure 13: Standardized bias (in %) associated with each explanatory covariates before and after matching for the PSM8 thematic programs with the nearest neighbors matching method



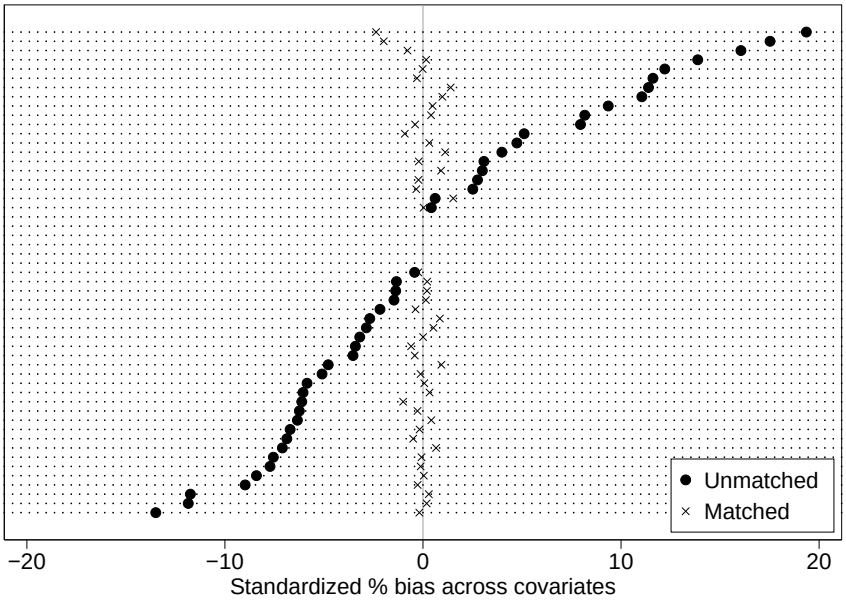
Note: Each dotted line represents an explanatory covariate included in the X vector.

Figure 14: Standardized bias (in %) associated with each explanatory covariates before and after matching for the PSM8 thematic programs with the kernel matching method



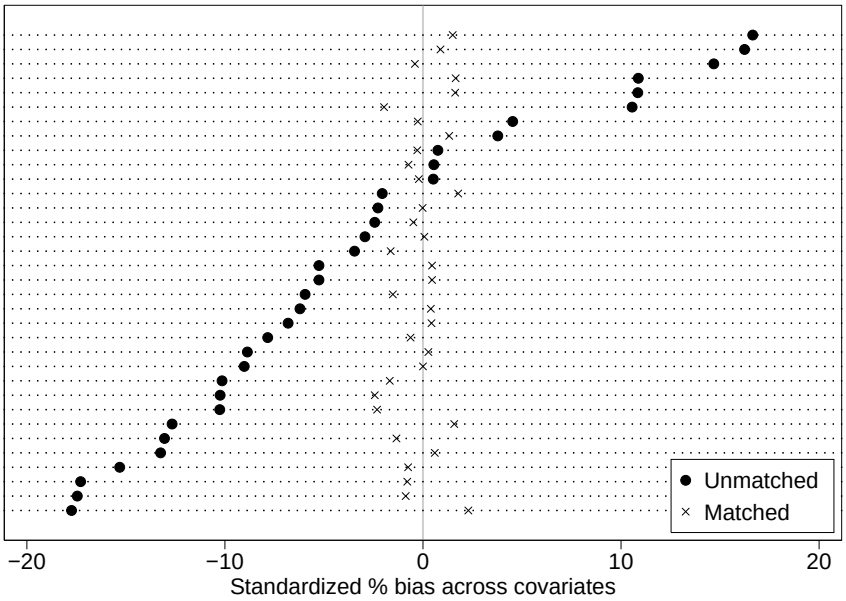
Note: Each dotted line represents an explanatory covariate included in the X vector.

Figure 15: Standardized bias (in %) associated with each explanatory covariates before and after matching for the PSM8 thematic programs with the IPTW matching method



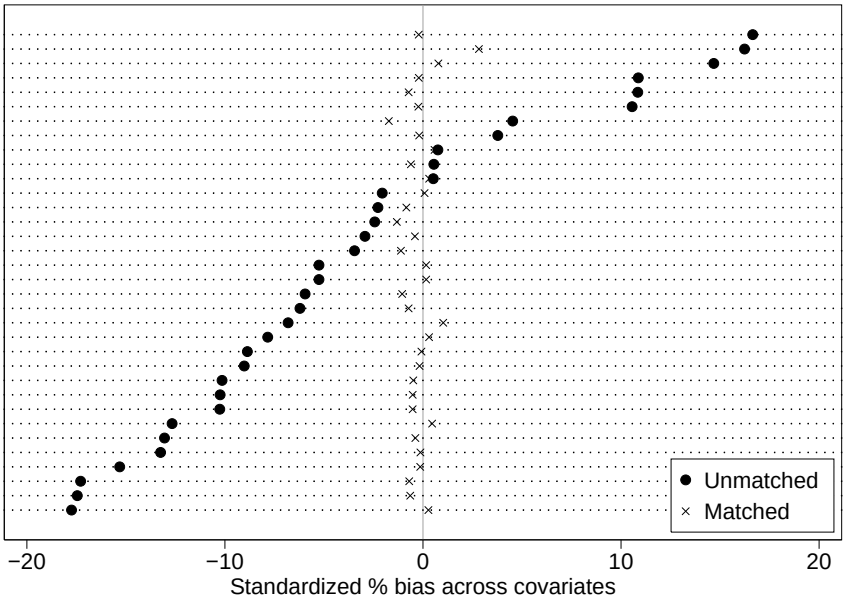
Note: Each dotted line represents an explanatory covariate included in the X vector.

Figure 16: Standardized bias (in %) associated with each explanatory covariates before and after matching for the PSM8 non-thematic programs with the 5 nearest neighbors matching method



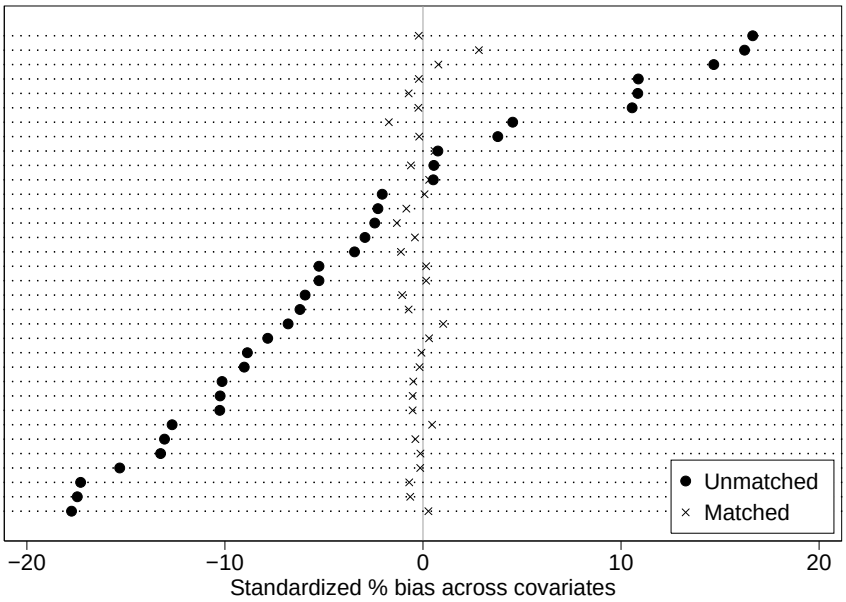
Note: Each dotted line represents an explanatory covariate included in the X vector.

Figure 17: Standardized bias (in %) associated with each explanatory covariates before and after matching for the PSM8 non-thematic programs with the kernel matching method



Note: Each dotted line represents an explanatory covariate included in the X vector.

Figure 18: Standardized bias (in %) associated with each explanatory covariates before and after matching for the PSM8 non-thematic programs with the IPTW matching method



Note: Each dotted line represents an explanatory covariate included in the X vector.

References

Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. Communications in Statistics-Simulation

and Computation, 38(6), 1228-1234.

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.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association, 94(448), 1053-1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics, 84(1), 151-161.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American statistical Association, 79(387), 516-524.

Appendix F. Supplementary estimation results

Table 19: Average treatment effect of receiving an ANR grant on publication outcomes (three years after treatment against three years before).

	δ^{5nn}	δ^{kernel}	δ^{iptw}
Volume	.03738***	.03544***	.03503***
	(4.46)	(4.54)	(4.46)
Citations	.1428***	.15098***	.15254***
	(8.42)	(9.45)	(9.30)
Impact Factor	.08023***	.08206***	.08252***
	(7.0)	(7.64)	(7.53)

*Note: Conditional difference-in-difference 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 either according to the nearest neighbors, to the kernel or 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: *.*

Table 20: Average treatment effect of receiving an ANR grant on collaboration behaviors and novelty (three years after treatment against three years before).

	δ^{5nn}	δ^{kernel}	δ^{iptw}
Average	0.0201**	0.0217***	0.0218***
Team Size	(2.37)	(2.76)	(2.71)
Coauthors	0.0930***	0.0984***	0.0981***
	(6.24)	(7.09)	(7.02)
International	0.0437***	0.0414***	0.0418***
Collaborations	(2.74)	(2.81)	(2.82)
New Coauthors[a]	0.0595**	0.0651***	0.0668***
	(2.54)	(2.97)	(3.03)
New Problems	0.00133	0.00160	0.00183
	(0.79)	(1.00)	(1.14)

*Note: Conditional difference-in-difference 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 either according to the nearest neighbors, to the kernel or 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 21: Differentiated effects of receiving an ANR grant on outcomes (three years after treatment against three years before) according to two different funding schemes: non-thematic versus thematic programs.

	δ^{5nn}	δ^{kernel}	δ^{iptw}
Volume	0.02136	0.02858*	0.02766*
	(1.28)	(1.85)	(1.77)
Citations	0.16543***	0.20122***	0.20275***
	(4.92)	(6.37)	(6.26)
Impact Factor	0.08989***	0.11275***	0.11112***
	(3.94)	(5.28)	(5.10)

*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 either according to the nearest neighbors, to the kernel or 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: *.*

Table 22: Differentiated effects of receiving an ANR grant on collaboration and novelty variables (next three years against previous three years) according to two different funding schemes: non-thematic versus thematic programs.

	δ^{5nn}	δ^{kernel}	δ^{iptw}
Average	0.00873	0.000408	-0.00107
Team Size	(0.51)	(0.03)	(-0.07)
Coauthors	0.0247	0.0199	0.0201
	(0.83)	(0.72)	(0.72)
International	0.0270	0.0282	0.0286
Collaborations	(0.85)	0.96	(0.97)
New Coauthors[a]	-0.199***	-0.163***	-0.170***
	(-3.62)	(-3.23)	(-3.31)
New Problems	-0.000236	0.000391	0.000460
	(-0.07)	(0.12)	(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 23: Differentiated effects of receiving an ANR grant on publication outcomes (next three years against previous three years) according to the age class (below the median age vs. over the median).

	δ^{5nn}	δ^{kernel}	δ^{iptw}
volume	0.0253	0.0202	0.0221
	(1.51)	(1.30)	(1.41)
Citations	0.107***	0.0866***	0.0952***
	(3.32)	(2.86)	(3.09)
Impact Factor	0.0359	0.0228	0.0266
	(1.62)	(1.11)	(1.29)

*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 either according to the nearest neighbors, to the kernel or 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: *.*

Table 24: Differentiated effects of receiving an ANR grant on the collaboration and novelty variables (next three years against previous three years) according to the age class (below the median age vs. over the median).

	δ^{5nn}	δ^{kernel}	δ^{iptw}
Average	-0.0338**	-0.0251*	-0.0279*
Team Size	(-2.05)	(-1.66)	(-1.80)
Coauthors	0.00799	-0.00474	-0.00344
	(0.27)	(-0.17)	(-0.13)
International	0.0421	0.287	0.0280
Collaborations	(1.32)	(0.98)	(0.95)
New Coauthors[a]	0.0035	-0.0416	-0.0418
	(0.07)	(-0.86)	(-0.83)
New Problems	-0.000959	0.000221	-0.0000433
	(-0.29)	(0.07)	(-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 either according to the nearest neighbors, to the kernel or 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 25: Differentiated effects of receiving an ANR grant on publication outcomes (next three years against previous three years) according to the investigator's role (principal investigator vs. partner coordinator).

	δ^{5nn}	δ^{kernel}	δ^{iptw}
volume	-0.0235	-0.0147	-0.0160
	(-1.37)	(-0.91)	(-0.99)
Citations	0.00602	0.0114	0.00500
	(0.18)	(0.36)	(0.15)
Impact Factor	0.00847	0.0219	0.0179
	(0.37)	(1.02)	(0.82)

*Note: Conditional difference-in-difference-in-difference results. 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 project-principal-investigator (PI) dummy, in a fixed effect regression. Observations are weighted either according to the nearest neighbors, to the kernel or 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: *.*

Table 26: Differentiated effects of receiving an ANR grant on publication outcomes according to the position in the citation distribution at the time of funding, on various production variables (next three years against previous three years).

Dependent Variable	Position in the volume distribution	δ^{5nn}	δ^{kernel}	δ^{iptw}
Volume	10-20%	0.0568**	0.0735***	0.0823***
		(1.97)	(2.76)	(3.06)
	20-30%	0.0828***	0.0958***	0.106***
		(2.91)	(3.72)	(4.05)
	30-40%	0.0467*	0.110***	0.118***
		(1.65)	(4.14)	(4.42)
	bottom 60%	0.0201	0.0558**	0.0632**
		(0.69)	(2.11)	(2.38)
Citations	10-20%	-0.00623	0.0885*	0.0989*
		(-0.11)	(1.68)	(1.80)
	20-30%	0.0526	0.0536	0.0716
		(0.96)	(1.05)	(1.33)
	30-40%	-0.0185	0.0804	0.0921*
		(-0.33)	(1.53)	(1.68)
	bottom 60%	-0.140***	-0.0987**	-0.0898*
		(-2.78)	(-1.96)	(-1.70)
Impact Factor	10-20%	0.0376	0.0540	0.0608
		(0.98)	(1.47)	(1.62)
	20-30%	0.0547	0.0585*	0.0675*
		(1.44)	(1.65)	(1.87)
	30-40%	0.0174	0.0812**	0.0880**
		(0.45)	(2.25)	(2.39)
	bottom 60%	-0.0429	-0.0236	-0.0188
		(-1.22)	(-0.71)	(-0.55)

*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 either according to the nearest neighbors, to the kernel or 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: *.*

Table 27: Differentiated effects of receiving an ANR grant on publication outcomes according to the year of funding, on various production variables (next three years against previous three years).

Year	Volume	Impact Factor	Citations
2006	0.0019	0.0108	-0.0099
	(0.06)	(0.26)	(-0.16)
2007	0.0169	0.047	-0.0065
	(0.52)	(1.09)	(-0.10)
2008	-0.0146	0.0317	-0.025
	(-0.43)	(0.72)	(-0.38)
2009	0.0322	0.0303	0.0064
	(0.97)	(0.72)	(0.10)

*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 year considered (the year 2005 is 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: *.*

Impact analysis by field

Table 28: Differentiated effects of receiving an ANR grant on publication outcomes according to the scientific discipline of the applicant (three years after treatment against three years before).

Field of science	Volume	Impact Factor	Citations
Medicine	-0.0266	-0.0624	-0.0366
	(-0.96)	(-1.48)	(-0.61)
Chemistry	-0.0153	-0.0244	0.0176
	(-0.61)	(-0.64)	(0.32)
Physics	0.0362	0.00717	0.0934
	(1.31)	(0.17)	(1.57)
Engineering	0.0190	-0.0266	0.0690
	(0.61)	(-0.61)	(1.02)
Universe Sciences	0.0167	0.0273	0.103
	(0.57)	(0.60)	(1.60)
ICST[a]	0.0616**	0.0120	0.0881*
	(2.50)	(0.37)	(1.74)
Mathematics	0.00617	-0.0213	0.0743
	(0.13)	(-0.38)	(0.84)
Social Sciences	-0.0203	-0.0388	0.00250
	(-0.41)	(-0.71)	(0.03)

*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 scientific discipline of the applicant (life sciences 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: *.*

[a] *ICST refers to "Information and Communication Sciences and Technologies".*

Appendix I. Authors disambiguation

The disambiguation algorithm

In this section, we present the three main stages of the used disambiguation algorithm, as well as some descriptive elements about the implementation of the procedure. To be definitely selected, a document has to pass the seed stage or, if not, the expand stage. We now present these two stages.

Seed stage

The seed stage can be decomposed into four conditions that need to be jointly verified:

- The name(s) and the initial(s) of the scientist should be identified within author identities (presented with a name and first name initials). The matching allows for variation in the name (introduction of a name particle and additional first name initials).
- The publication date of the article should be consistent with the researchers' or professors' age that year. We have retained a minimal age of 24 years and an upper limit of 80 years.
- The declared disciplinary field of the scientist (mentioned on the administrative data under the classification form section) should be consistent with the specialty of the journals in which their papers are published (determined on the basis of the field classification of scientific journals performed by the OST).
- The institution to which the scientist is affiliated should be mentioned in the affiliations of the author(s) of the paper. In order to be able to establish a connection between both information types (they could be spelled differently), the complete denomination of the institution is chosen (e.g., Université d'Aix-Marseille, Université Toulouse III and ENS Paris), as well as considering the fusions between institutions that have been carried out so far (e.g., Université de Bordeaux). The reason why we have not employed the laboratory name to perform the comparison is explained on the basis of the larger complexity to set up a connection between both bases (laboratories are often spelled differently), as well as leave open the possibility that a scientist can be affiliated to more than one research laboratory at the same institution.

Expand stage

The expand stage offers a chance to all the documents that did not pass the seed stage by relaxing some of the previous conditions, while introducing new conditions based on the potential similarity with already validated articles of the same researcher or professor. We consider three types of information:

- Two types of keywords (reported by the authors or attributed by ISI WoS),
- The coauthors (surname and first name initials),
- The reference lists.

The basic idea is that scientists are more likely to use the same keywords, work with the same people and cite the same papers.

Basically, the expand stage works as follows:

1. First, relax the fourth condition of the seed stage (same institution);[12]

[12]This strategy considers that scientists can be mobile and thus allows us to recover the articles published when they were working in another institution during their academic career. It also allows us to consider that authors sometimes misreport their institution.

2. Then validate all candidate papers reporting a keyword (from the authors) used in a previously validated articles by the same researcher or professor.

3. If there are validated articles in step 2, add them to the list of previously selected article and return to 2; otherwise, go to next step;

4. Validate all candidate papers reporting a keyword (attributed by ISI) used in a previously validated articles by the same researcher or professor.

5. If there are validated articles in step 4, add them to the list of previously selected article and return to 4; otherwise, go to next step;

6. Validate all candidate papers that are authored by one of the authors of the articles previously validated by the same researcher or professor;[13]

7. If there are validated articles in step 6, add them to the list of previously selected article and return to 6; otherwise, go to next step;

8. If no article is validated in steps 2, 4 and 6, go to stage 9; otherwise, loop on step 2;

9. Now relax the third condition of the seed stage (same field);[14]

10. Validate all candidate papers that have a reference list sufficiently similar to at least one of the articles previously validated by the same researcher or professor;

11. If there are validated articles in step 10, add them to the list of previously selected article and return to 10; then stop anyway after 30 loops;[15]

12. If there are validated articles in step 11, go to step 2; then stop anyway after two loops;

The similarity between reference lists is based on a score calculated as follows:

$$\alpha_{ij} = \sum_k \frac{1\{i, j \to k\}}{\#\{u \,|u \to k\}},$$

for two papers i and j, of which one is already validated and the other is a candidate paper. The dummy $1\{i, j \to k\}$ takes the value 1 if reference k is cited at the same time by i and j (it is a common reference). The denominator $\#\{u \,|u \to k\}$ is the number of citations that reference k received. It allows us to control the citation frequency of common reference: the more a common reference is cited, the less it should increase the similarity score. We perform the following normalization: $\theta_{ij} = \alpha_{ij}/max_{v=i,j}\{\alpha_{vv}\}$. This normalization is predicated on the maximal similarity that the reference lists of the two papers could reach; that is, the similarity reached if their reference lists were identical, and identical to the one that has the greater self-similarity. The threshold for inclusion is defined as the 98th percentile of all θ_{ij} recorded in the publications of the members in the section.[16]

The collection process is detailed in Table 56, which shows the number of retrieved publications and their related researchers at each stage. The expand stage has been run successively twice, with seven complete loops in the first round and four in the second (each round was followed by 30 loops for the reference list).

[13]Herself being excluded from the author's lists. Moreover, only consider the authors of papers with fewer than 50 authors.

[14]This strategy also considers that scientists can publish in different fields

[15]Since reference loops are quite heavy and can loop a great number of times for only a few validated, we decided it was appropriate to stop after 30 loops.

[16]The 98th percentile was chosen in order to optimize the disambiguation performance.

Table 29: Number of newly retrieved publications at each step and the number of related researchers

Stage	# documents	# authors
SEED	521,817	29,647
EXPAND		
Round 1		
• Keywords & Authors	585,324	29,309
• References	87,953	29,193
Round 2		
• Keywords & Authors	7,963	29,189
• References	7,929	29,160
Total	1,210,986	29,160
FINAL SAMPLE	1,210,867	29,154

Note: The column "#documents" gives the number of new validated papers in each step. The column "#authors" gives the number of authors left in the database (authors with more than 500 retrieved papers are removed). At the end of the disambiguiation process, a total of 1,210,986 publications is allocated to a sample of 29,160 researchers. Afterwards, the sample is reduced to 29,154 researchers, which equates to 1,210,867 documents, once we correct for homonymy issues.

Benchmarking the disambiguation

This section explains the creation of the benchmark and the indicators used to assess the quality of the disambiguation.

We established a list of 353 French researchers who created an ORCID number[17] and can be found in our initial list of researchers.[18]

The performance indicators used are precision and recall. Precision measures the ability to clearly identify the correct documents from among a set using a common author's identity, whereas recall refers to the ability to retrieve as many relevant publications as possible. These two indicators are scored by:

$$PRECISION = \frac{number\,of\,true\,positives}{number\,of\,true\,positives + number\,of\,false\,positives}$$

$$RECALL = \frac{number\,of\,true\,positives}{number\,of\,true\,positives + number\,of\,false\,negatives}$$

where the true positives stand for the relevant recovered publications, the false positives are the papers retrieved by mistake (they belong to another author) and the false negatives gather the relevant papers that should have been collected, but are missing.

In order to enhance the quality of our disambiguation approach, we implemented exclusion conditions at the researcher level. We decided to set an upper bound of 500 validated publications per author. Hence, at any step of the algorithm, if the number of documents recovered by a researcher exceeds this threshold, we consider that our disambiguation approach has not been relevant enough to treat this homonymy issue, such that the researcher is definitely discarded from the analysis. From the benchmark formed in relation to the 291 remaining scientists,[19] we get a recall of 0.90 and a precision of 0.82.

[17] An ORCID number lets researchers verify their own publications set on a voluntarily basis.

[18] A manual checking on the similarity of affiliation has been done to ensure there is no homonymy issue.

[19] Among these 291 researchers, no publication was retrieved for 21 of them during the disambiguation process. The mean age of this benchmark is 42.14, and those individuals reside in a large number of French research center locations. Almost 60% of the sample is affiliated to a university (22% of professors, 37% of associate professors), whereas 40% work as full time researchers.

In Figure 17, we represent the relationship between verified publications vs. retrieved publications for different measures of outputs, finding that the observations are mostly located on or around the first bisector, which suggests that errors are limited.

We compare our results with those of Reijnhoudt et al. (2013), who develop a different seed+expand approach to deal with the disambiguation of authors' names. In this paper, the publications collection is based on the similarity between the individual and article characteristics in the seed stage (affiliation addresses, e-mail addresses), as well as exploiting papers' common features combined with various data sources (WoS, Scopus) in the expand part. Reijnhoudt et al. tested their methodology performance on a sample of 1,400 researchers with verified publication records in the period 2001-2010 ("CWTS 'gold standard"'), drawn from a set of 6,753 Dutch full professors. From this subset, they obtained a recall close to 0.96 and a precision in the range of [84.2 - 88.5] for the three different versions of the expand stage. At first sight, our study seems to perform to a lesser extent, but this gap may be explained as follows. Firstly, their indicators are calculated only in the basis of the publications of those authors who retrieved at least one publication from the seed stage, whereas we include those for whom no publication is recovered during the seed stage. As a consequence, their precision is positively biased because false positives are artificially reduced (the recall remains unchanged). Our second remark relates to the selection of individuals for inclusion in the benchmark. In Reijnhoudt et al.'s study, the verified list of relevant publications was systematic, obtained according to the authors' requests or directly from the administration. Our benchmark is different, limited to professors and researchers who created an ORCID profile. Such profile creation is voluntary and unsolicited, which means it is more subject to selection bias. We suspect that the main reason that spur some of the researchers to create an ORCID number (and verify their publications) is a large and complicated publication profile which could prove to be difficult to disentangle automatically. Typically, this is the case when authors have been mobile in the career or publish in different scientific domains. Thus they want to clarify the authorship of their publications record, which is facilitated by creating an ORCID profile. As a consequence, the use of our specific benchmark is likely to introduce a negative bias on the recall and precision indicators.

Figure 19: Comparison of scores on three indicators, comparing correct vs. retrieved measures for the professors and researchers in the benchmark