Final Report for Period: 01/2010 - 12/2010
Submitted on: 02/04/2011
Principal Investigator: Shapira, Philip .
Award ID: 0738126
Organization: GA Tech Res Corp - GIT
Submitted By:
Shapira, Philip - Principal Investigator
Title:
MOD Measurement and Analysis of Highly Creative Research in the US and Europe

Project Participants

Senior Personnel

Name: Shapira, Philip
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Rogers, Juan
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Youtie, Jan
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Galope, Reynold
Worked for more than 160 Hours: Yes
Contribution to Project:
Research assistance in data development and analysis for matching HCRs. Also involved in process of CV collation and collection.

Name: Tang, Li
Worked for more than 160 Hours: Yes
Contribution to Project:

Undergraduate Student

Name: Bidgood, Anne
Worked for more than 160 Hours: Yes
Contribution to Project:
Undergraduate research into knowledge networks in non-CV data sources and assistance with data coding.

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Otto-Friedrich-Universitaet Bamberg

Dr. Thomas Heinze of the Otto-Friedrich-Universitaet Bamberg, Germany, is our European collaborator in this project. Note: Originally, Dr. Heinze was with the University of Twente, Netherlands. However, following Dr. Heinze's move to the University of Bamberg in April 2008, we moved the designated European subcontractor relationship in the project to Otto-Friedrich-Universitaet Bamberg, with NSF concurrence and approval. Dr. Heinze has been involved in the ongoing work of the project, including sample selection of the matching group for the highly creative scientists, particularly for the European side of the sample.

Other Collaborators or Contacts

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

The SciSIP project, 'MOD Measurement and Analysis of Highly Creative Research in the US and Europe,' NSF Award 0738126 investigates the features of the meso level (e.g., team, organizational, institutional) of the research environment that enable and foster highly creative research in nanotechnology and human genetics in the US and Europe. It will also examine the influence of career patterns. The study contributes to the methodology of science studies by further developing and extending curriculum vita (CV) analysis. The identification of meso level factors in the research environment has broader implications for research and human resource management, and the design and implementation of funding schemes. The use of comparative fields extends the range of impact to two different emerging fields. Public datasets containing variables related to the creative researcher nominees will be made available for use by others.

During the second year of the project (January 1-December 31, 2009) several critical research activities have been completed. We implemented a robust method for identifying a matching control group for the highly creative researchers (HCRs) in the US and Europe identified in the CREA I study (51 in nanotechnology or NT; 25 in human genetics or HG). After trialing several different methods, we found that a theory-driven matching procedure produced the best results in matching the HCR group and the non-HCR. This resulted in a matching sample that comprised of NT = 463 and HG = 249. We developed a protocol to request CVs of the HCR matches, then requested those CVs and updated CVs of HCRs. These procedures occurred simultaneously in the US and in Europe and continued through the first two months of 2009. Based on responses received, we were able to match all HCRs with most optimal non-HCR respondents. We then initiated a process of coding CV data. A coding scheme was developed, with more than 60 potential variables. These variables were related to hypotheses (such as number of job changes or early-career awards) and to general background variables (such as age and gender). This highlighted that a number of key variables of interest (including PhD supervisor, research awards, and academy memberships and prizes) were not universally available via CVs. We have sought to identify missing information using secondary sources, including web searches, dissertation abstracts, and searches of award databases. Subsequently, we have initiated a follow-up process to validate the information directly with the HCRs, non-HCRS, and CRS. This process was begun towards the end of 2009 in both the US and Europe and will continue into the New Year. We have undertaken a significant effort to code and validate data, but judge that it is important to make these investments, searches and follow-ups to ensure that we have a fully accurate and validate data set for the HCRS and controls. The variable detail and accuracy we are securing in terms of researcher career development will form a unique and unparallel dataset. Additionally, we have also developed secondary institutional and field variables, including organizational publications and researcher networks.

Towards the end of 2009, we requested a No-Cost Extension to the project. In order to complete planned activities. In part, this NCE was necessary because of the late start of the project in 2008, due to changes in the European partner organization and associated administration. In December 2009, we received confirmation that NSF has approved a 1 year NCE, through to 12/31/10.

During the third year of the project (January 1-December 31, 2010) several critical research activities have been completed. The initial coding of summer 2009 highlighted that a number of key variables of interest (including PhD supervisor, research awards, academy memberships, and prizes) were not universally available via CVs. In 2010, we initiated a verification process through multiple rounds of email-based surveys in

order to complete missing data. These surveys were administered to US and European researchers from December 2009 through April 2010. We received further data from 40 percent of the European matches and HCRs and 25 percent of the US matches and HCRs as a result of this survey. This verification process called attention to some issues of which the project team would not have otherwise been aware. One example from the European side is the DSc, which we learned is a thesis rather than an education degree in the same way that a PhD. Based on these CVs, we updated our data of HCRs, HCRs matches, and CRs. We also validated the publications of HCR and HCR match researchers in both the US and EU.

In the year of 2010, we developed and validated six linked Microsoft Access databases: (1) US highly creative researchers (HCRs), (2) US matches to HCRs, (3) US creative researchers, (4) European HCRs, (5) European matches to HCRs, (6) European creative researchers with their publications.

To supplement this CV-based information, we extracted Web of Science publications for the researchers. We imported them into the VantagePoint software, cleaned them, and used software macros to create several measures of interdisciplinarity - integration (the extent that articles cite different journal subject categories) and specialization (the extent that authors publish in journals from diverse subject categories). These results were exported to Excel files for each researcher in the Microsoft Access databases. We then engaged in further coding to harmonize differences between US and European degrees (for example, the German-style diploma) and created a STATA dataset from these records.

Our final CREA 'database', which consolidated publication data and CV information are saved in three formats: Access relational dataset, Excel file with coding book, and a flat file of STATA data ready for statistical analysis.

In addition to labor-intensive data collection and cleaning, the US and European project principals held several telephone and video conferences during the year discussing coding schemes and research directions. In September 2010, US project principals Shapira, Youtie, Rogers met with the EU project principal Heinze in Darmstadt Germany, prior to participation in the 2010 Society for the Study of Nanoscience and Emerging Technologies (S.NET) Conference.

As indicated in our original proposal, and as requested by NSF SciSIP, this project has posted an anonymized public dataset comprised of selected pre-coded and readily understandable variables to the Georgia Tech SMARTech repository for use by the wider research community. The link is at: http://hdl.handle.net/1853/36717

Variables in this dataset and the codebook are shown below:

Variable Name: Description
Person_ID: Unique ID of researcher
HCR_MATCH: Researcher type: Highly creative researcher or matched researcher. 1 for HCR; 0 for Match
HG_NANO: Research domain: 1 for Human genetics ; 0for nanotechnology
EU_US: Country affiliation in 2005 when the survey was conducted. 1 for EU; 0 for the US
NumEduInst: Number of institutions from bachelor's degree to Ph.D. degree
Phd_Year: Year of getting the 1st highest degree (Ph.D. or MD)
PostDoc: Dummy variable for Postdoc experience. Yes 1; No 0
EarlyJobAcademic Dummy variable, working experience in academic institution within the first 6 years of terminal degree. Yes 1; No 0
YearGrn,_NonUniv: Starting year of the first non-university award
NumGrnTyp: Number of different types of grants received

In 2010, the project supported two doctoral graduate research assistants, Reynold Galope from the Georgia Tech-Georgia State Joint PhD Program in Public Policy, and Li Tang from the Georgia Tech PhD Program in Public Policy. Both students have assisted in database development, matching methodology development, coding, and initial analysis. Galope later uses the matching methods for his doctoral dissertation. In February 2009, the project submitted a supplemental REU request to support an undergraduate to conduct research into knowledge networks in non-CV data sources. This request was subsequently approved, and since fall 2009 we supported Anne Bidgood, a Georgia Tech undergraduate in Industrial & Systems Engineering, to work with the project and receive research training and mentoring.

In 2010, the project supports three graduate research assistants at the University of Bamberg: David Pithan, Steffi Heinecke, and Tobias Philipp. All students have assisted in data entry, coding, and initial analysis. Philipp finished a diploma thesis on a topic related to the project.

Findings:

Analysis of the results of our project is not fully complete (the dataset of rich and complex, and we are continuing to work through the analysis

in 2011). Nevertheless, we have some preliminary findings related to the early career development of HCRs which distinguish them from their matches. Analytic priority has been given to a paper on early and mid-career mechanisms. This paper focuses on profiling the distinctive early and mid career characteristics of highly creative researchers. An extensive literature review on the early career characteristics of highly creative research has completed. Two competing theories are explored in this paper: Focused hypothesis (drawing on Merton's accumulation hypothesis) versus a Networked counter hypothesis (drawing on Hollingsworth's integrative diversity hypothesis) are tested. The results show that postdoctoral training experience is a good predictor to distinguish HCRs and matched researchers. Receiving education in multinational context increases the likelihood of doing creative research for European researchers but not US researchers. Compared to their counterparts, researchers who spent fewer years getting their doctorate or other terminal degree and moving to tenured position or senior position are more likely to become highly creative researcher. This analysis and paper writing was begun towards the end of 2010 and continued into the New Year. The resulting paper, in the final stages preparation and titled 'What Early Career Characteristics are Distinctive in Highly Creative Researchers,' is planned for submission to Research Policy.

Training and Development:

The project has continued to improve research skills and training for students (and faculty) in several methods, including advanced techniques of control group matching, bibliometrics, and CV analysis and coding. Graduate and undergraduate students have been supported.

Outreach Activities:

Presentations on the CREA project included:

? 'Understanding and Stimulating Highly Creative Scientific Research,' Presentation at Innovation Seminar Series, Manchester Institute of Innovation Research, Manchester Business School, UK, February 2, 2009. (Shapira)
? Factors enhancing career opportunities: group size, mentorship, and network position,. Presentation at Conference: Women in Top Research Positions, University of Hamburg, Germany, February 13, 2009 (Heinze)
? 'Understanding and Stimulating the Organization of Highly Creative Research. Measurement and Analysis,' Presentation, Panel on The Evolution of Knowledge Production: Exploring Creativity, Innovation, and Networks, American Association for the Advancement of Science, Annual Meeting, Chicago, February, 15, 2009. (Shapira)
? A Matching Protocol for Constructing Comparison Groups for Highly Creative Researchers. 9th Science and Technology in Society Conference, AAAS, Washington D.C. March 28-29, 2009. (Galope)
? Understanding and Stimulating the Organization of Highly Creative Research. Measurement and Analysis - U.S. and Europe,' Science of Science Policy (SciSIP) PIs Workshop, National Science Foundation and American Association for the Advancement of Science, Washington DC, March 24, 2009. (Shapira)
? Organizational and institutional influences of scientific creativity. Colloquium, University of G?ttingen, Germany, June 29, 2009 (Heinze)
? Understanding and Stimulating Highly Creative Research: Measurement and Analysis - U.S. and Europe. Special Session: Developing a (Social) Science of Science and Innovation Policy, American Sociological Association, Annual Meeting, San Francisco, August 10, 2009. (Youtie)
? Understanding and Stimulating the Organization of Highly Creative Research. Measurement and Analysis - U.S. and Europe,' Atlanta Conference on Science, and Innovation Policy, October 2-3 2009. (Shapira)
? Studying Scientific Creativity: Methodological Challenges. Research Policy Institute, University of Lund, May 8, 2009 (Heinze)
? Institutional Influences on Creativity in Scientific Research. Lecture at the Department of Psychology, University of Lund, Sweden, May 7, 2009 (Heinze)
? Explaining research breakthroughs using a organizational sociology perspective. Colloquium, University of Wuppertal, Germany, December 18, 2009 (Heinze)
? Institutional Conditions for Creative Research. Presentation at the Innovation Seminar, German University of Administrative Sciences Speyer, June 7, 2010 (Heinze)
? Mechanisms of Institutional Renewal. Spring Meeting of the Section 'Science and Technology Research' of the German Sociological Association, University of Bamberg, April 22-23, 2010 (Heinze)

Journal Publications

Books or Other One-time Publications

Youtie, J.; Shapira, P.; and Rogers, J., "Blind Matching versus Matchmaking: Comparison Group Selection for Highly Creative Researchers", (2009). Conference procedings, Submitted

Collection: IEEE Xplore
Bibliography: Atlanta Conference on Science and Innovation Policy

T. Heinze and P. Shapira, "Path-breaking Science: Institutional Foundations of Creative Research", ( ). Book, Under contract with Oxford University Press
Bibliography: Under contract and in preparation (publication scheduled for 2012)

Web/Internet Site

Other Specific Products

Product Type:

Journal Article

Product Description:

Heinze, T., Shapira, P., Rogers, J.D., and Senker, J.M., Organizational and institutional influences on creativity in scientific research, Research Policy, 2009, 38, 610-623.

Sharing Information:

Published in journal "Research Policy" and widely available via online databases

Product Type:

Paper

Product Description:

Heinze, T., Bauer, G. (forthcoming): Creativity capabilities in nanoscale research. Longitudinal population level evidence. Annales d?Economie et de Statistique, Special Issue edited by Lynne Zucker and Richard Freeman

Sharing Information:

Forthcoming in Annales d?Economie et de Statistique

Product Type:

Working Paper

Product Description:

Youtie, J; Shapira, P; Rogers, J; Heinze, T. What Early Career Characteristics are Distinctive in Highly Creative Researchers.

Sharing Information:

Working paper, under revision for journal submission in Spring 2011.

Contributions

Contributions within Discipline:

Our efforts to develop a matching control sample for a group of highly creative researchers has, we believe, results in useful methodological insights. These insights will be useful to other researchers in the filed. We have shared our insights by producing a methodological paper on matching highly creative researchers (published in IEEE Explorer).

Insights from the NSF CREA project have informed collaborations and contacts with researchers at Arizona State University (E. Corley) and the University of Wisconsin (D. Scheufele) examining risk perceptions and research practices of nanoscientists.

Contributions to Other Disciplines:

The methods, techniques, and findings from the CREA project have attracted interest from natural scientists, research managers, and policy makers, as they consider strategies for improving research performance.

Contributions to Human Resource Development:

Students engaged in the project have gained significant expertise in
data mining and propensity matching. The graduate students most involved have been are Luciano Kay (PhD Student, Georgia Tech Public Policy) (to 2009), Reynold Galope (PhD Student, Joint Program in Public Policy, Georgia Tech and Georgia State University) (to 2010) and Li Tang, PhD Student, Georgia Tech Public Policy) (2010).

Contributions to Resources for Research and Education:

Contributions Beyond Science and Engineering:

Presentations at AAAS, at NSF, and at other locations have generated interest in the methods and findings of the project and informed broader concerns about ways to improve research performance and management.

Conference Proceedings

Categories for which nothing is reported:

Any Journal

Any Web/Internet Site

Contributions: To Any Resources for Research and Education

Any Conference

The CREA Project: Measuring and Analyzing Highly Creative Scientific Research
www.cherry.gatech.edu

Working Paper

Blind Matching versus Matchmaking: Comparison Group Selection for Highly Creative Researchers

Philip Shapira[1,2,*], Jan Youtie[3], and Juan Rogers[1]

November 2009

[1] School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332-0345, USA.
[2] Manchester Institute of Innovation Research, Manchester Business School, University of Manchester, Manchester, M13 9PL, UK.
[3] Enterprise Innovation Institute, Georgia Institute of Technology, Atlanta, GA 30332-0640, USA

* Corresponding author - Email: pshapira@gatech.edu

Acknowledgement

We gratefully acknowledge support from the National Science Foundation (Science of Science & Innovation Policy) under award number SBE-0738126. We also draw on earlier research funded by the European Commission under award number EU-NEST/CREA-511889. Thanks to Reynold Galope and Stephen Carley for their work assembling the data and to Thomas Heinze for his helpful guidance. The results are solely the responsibility of the authors.

Abstract

This research examines approaches for constructing a comparison group relative to highly creative researchers in nanotechnology and human genetics in the US and Europe. Such a comparison group would be useful in identifying factors that contribute to scientific creativity in these emerging fields. Two comparison group development approaches are investigated. The first approach is based on propensity score analysis and the second is based on knowledge from the literature on scientific creativity and early career patterns. In the first approach, the log of citations over the years of activity in the domains under analysis produces a significant result, but the distribution of matches is not adequate at the middle and high ends of the scale. The second approach matches highly creative researchers in nanotechnology and human genetics with a comparison group of researchers that have the same or similar early career characteristics were considered: (1) same first year of publication (2) same subject category of the first publication, (3) similar publication volume for the first six years in the specified emerging domain. High levels of diversity among the highly creative researchers, especially those in human genetics, underscore the difficulties of constructing a comparison group to understand factors that have brought about their level of performance.

Keywords: Scientific Creativity, Propensity Score Matching, Publications, Citations

1. Introduction

Creative capabilities are an important cornerstone of progress in science and technology, and also a precondition for advances in other societal domains. The desire to know more about the factors that contribute to research creativity is given impetus by the substantial changes seen over the last three decades in the institutional and organizational conditions under which scientific research is conducted. In the debate as to whether the individual genius or the broader environment are responsible for some of the major discoveries (Simonton, 1999; Merton, 1973), it is clear that policies have changed from long-term disciplinary grants directed towards individual researchers to competitive project funding for research centers, networks, and cross-disciplinary teams. Efforts to promote scientific creativity and excellence in the face of increasing competition from China and other rising global locations calls fresh insights about the factors that can stimulate and sustain highly creative research which, in turn, require improved measures for assessing and distinguishing highly creative work.

One of the issues in examining highly creative work and distinguishing the factors that facilitate it is need for construction of a comparison group. Highly creative researchers are by nature a selective group that operates in a selective setting, so disentangling their characteristics from environmental attributes can be challenging. Development of a good comparison frame would enable matching of highly creative researchers with a paired set of regular researchers to understand the effects of relevant observed characteristics and reduce systematic differences in unobserved characteristics. This approach would allow for addressing of confounding selection biases. But highly creative researchers are difficult to match because they are by definition non-normal.

Two paths from the literature are suggestive for addressing this situation. The first emphasizes theory-based attributes of highly creative research. Productivity is one such attribute. Simonton's (2004) work argues that the more prolific a researcher is, the greater the likelihood that this output will eventually a produce high impact contribution because of the application of the constant probability law to the relationship between quantity of publications and quality in terms of citations. This argument is popularized in Gladwell's (2008) account of the amount of early career hours logged (in excess of 10,000), which is often coupled with access to specialized equipment and assistance, in the backgrounds of some of the most highly successful inventors. Based on this line of reasoning, highly creative researchers could be compared to a pool of researchers with similar levels of productivity or other relevant attributes to understand important differences and similarities. Heinze and Bauer (2007) have done this type of match of highly creative researchers in nanotechnology and human genetics based on publication output along with citations, linkages with unconnected scientists, and multidisciplinarity. Their analysis finds that while productivity is an important distinguishing attribute of highly creative scientists, so too is the ability to link disconnected scientists across disciplines.

A second path focuses on understanding the factors of high impact research in the context of evaluation of a particular program. The focal program is usually a program that makes awards to eminent or highly regarded researchers. The challenge in this type of research is that such programs by definition honor a highly selective set of the "best" individual researchers and thus are subject to selection bias in efforts to understand how these awardees differ from the

population of researchers. In particular, this bias makes it difficult to construct a comparison group because who are those not selected are likely to differ in observed, if not unobserved, ways. One way to address this deficiency is to comprise the treatment group of unsuccessful but very highly rated applicants to the programs. The National Research Council's (2006) evaluation of the Markey Scholars program conducted just this type of matching. This evaluation compared successful applicants to two classes of unsuccessful applicants: those who were "top ranked" and whose applications were given high rankings, and those considered "competitive" and whose applications received slightly lesser rankings. While one might expect the unsuccessful applicants to differ significantly from successful ones, the study's anecdotal reviews of first group of top ranked but unsuccessful applicants concludes that this top-ranked but unsuccessful group is nearly identical to that of the successful awardees. The results of this evaluation indicate that the awardees and highly-rated but non-awardee group did not differ much on measures such as faculty position or publication success, but the successful awardees were more apt to have been at top universities, received tenure and been promoted, and received more research grants.

An evolution of these two approaches involves statistical matching of target and comparison groups to account for selection biases. This approach uses techniques such as propensity score matching to statistically create an appropriate matched pool using a set of available information of pertinent attributes. (Rosenbaum and Rubin, 1984; Rubin 1997) A model is created with the treatment and control group membership as the dependent variable conditional on a set of independent variables. The propensity score matching will yield a balanced design of treatment and control groups that have the same or similar conditional probabilities relative to the independent variables in the model. The model must produce a distribution of propensity scores that has enough balanced observations in each group. (Lee 2006) Unobserved differences are not accounted for in propensity score matching, unlike in the case of randomized experimental designs. Pion and Cordray (2008) use propensity score matching, along with the aforementioned approach of constructing comparators from highly rated but not awarded applications, to understand the impact of the Career Award in Biomedical Sciences (CABS). Their effort to identify factors distinguishing CABS awardees from highly rated but not awarded applications did not prove useful because of the heterogeneity of unsuccessful applicants. The propensity score analysis of CABS was able to isolate a small set of attributes that distinguished awardees from comparators, including articles appearing in top-ranked journals, attaining faculty positions, and receiving early R01 grants. However, the analysis was challenged to achieve balance due to the clustering of awardees in the top quintiles and comparators in the bottom quintiles.

These approaches highlight the challenges in efforts to match highly creative researchers with a relevant population to identify distinguishing factors for investigative purposes and often subsequent policy development and implementation. Highly creative researchers have unique characteristics that affect their distribution of observations along most dimensions. The very features which distinguish them as highly creative also make them difficult to compare with the broader population of researchers. Approaches that rely on the central limit theorem do not apply because highly creative researchers do not follow a normal distribution. To understand what differentiates highly creative researchers, matching these researchers to a comparison frame and how one sets up the matching matters. This work informs and advances efforts to create a matching frame to understand the factors that encourage highly creative research. We present results from two approaches. The first is based on statistical matching models and the second

draws from the literature on early career creativity. We use publication data from the Web of Science in nanotechnology and human genetics domains to explore these approaches. Results suggest that current attributes are less useful than early career characteristics for developing matching frames and that statistical models suffer from inherent heterogeneities across the populations.

2. Data and Methods

The main research question guiding this study is: how can we develop a matched comparison group for subsequent study of the factors that distinguish highly creative researchers in nanotechnology and human genetics? The specific objective is to develop a matched comparison group of researchers to pair with an existing dataset of highly creative researchers (HCRs), which would then in a subsequent analysis receive an email request for a copy of their curriculum vita (CV). This CV would then form the basis for measurement of career trajectory and "meso-level" level factors of the organization to be used to distinguish highly creative researchers from their matched comparator. Because of this subsequent email-based CV request, the comparison group would require several matches for a given highly creative researcher to accommodate nonresponse to the email request.

The major challenge inherent in this objective is that highly creative researchers have the potential to be so far out on the tail of any research novelty's distributional measure that they become difficult if not impossible to match. But the extent of this challenge depends on how the concept of a highly creative researcher is defined and operationalized. In this study, we use the listing of highly creative researchers in Europe and the US in nanotechnology and human genetics pioneered in Heinze et al. (2007). This study's conceptual definition of highly creative research is that "highly creative research is work that is both novel and which has major implications or potential" (Heinze et al. (2006, p. 16). This definition is then operationalized as a select group of highly nominated and/or multiple prize winning researchers. These researchers were identified in the Heinze et al. work through a survey of some 300 peers and gate keepers including highly published researchers and journal editors. This survey requested respondents to provide up to three nominated researchers along with a description of their research accomplishment and justification of why the research is considered highly creative. Nominations were also coupled with a search of winners of nearly 100 prizes relevant to the two target fields.

The two target fields – nanotechnology and human genetics – were chosen to enhance the comparative nature of the work. Human genetics is a comparatively more discipline-embedded field with a longer established history going back to the middle of the 20th century. In contrast, nanotechnology is an emerging interdisciplinary field (Porter and Youtie, 2009; Rafols and Meyer 2009) with a more recent time horizon dating from the microscopy discoveries in the 1980s. These distinctive attributes have implications for the distribution research attributes among highly creative researchers themselves.

It was determined that we would use the publication record of the highly creative researchers in their respective fields (nanotechnology or human genetics) as the basis for developing a matched comparison group. The publication record came from a multi-module Boolean search strategy

for each field that draws on journal names and titles/keywords/abstracts in the Web of Science's Science Citation Index (SCI) from 1990-2006. (Porter et al., 2008; Heinze et al, 2007)

This decision poses two challenges. The first challenge concerns truncation of the publication record. Because both of the target technological areas are emerging fields, they do not encompass the full research activity of any of the highly creative researchers. Moreover, the extent of truncation of publishing activity varies considerably; some researchers' publication records are almost fully covered by the emerging field as we have operationalized it in our study, while others have rather few articles in the target field. An initial examination of this truncation effect indicated that the effect was greater in the case of human genetics. We posited that the setting of the early threshold to 1990, while arguably appropriate for nanotechnology given the microscopy discoveries of the 1980s that enabled nanoscale manipulation, was not as appropriate for the more established field of the human genetics field. Therefore we extended the early threshold for human genetics from 1990 to 1970. We also added five additional genetics journals that were not in the original human genetics Boolean search in Heinze et al 2007 and filtered articles in these journals for inclusion of the term "human." The results yielded nearly 126,000 human genetics publication records extracted from SCI along with 407,000 nanotechnology records. Truncation of the full publication record of the highly creative research is observed (See Table 1). In the case of nanotechnology, nearly 40% of the 50 highly creative researchers have more than half of their total publication record included in the nanotechnology domain as defined in this study, and more than three-quarters of these researchers have 25% of their records included. In the case of human genetics, however, only 12% of the 25 highly creative human genetics researchers have more than half of their total publication record included in the human genetics domain as defined in this study, and forty percent of these researchers have a quarter of their records included. Many of these underrepresented researchers in human genetics had publications that related to genetics in plants for example, but not to the more specific field of human genetics. Still it is reasonable to assume that an emerging field would not necessarily include all of a researcher's publication records, but that the field would have sufficient representation in the publication domain for analytic purposes.[1]

[1] In the nanotechnology domain, Barthlott, W; Neinhuis, C; and Belcher A. have published in journals that are not well covered by the domain definition used in this study. In the case of Barthlott and Heinhuis the search strategy developed by Porter et al. (2008) specifically excluded nanoflora and nanofauna while these highly creative researchers focused their work in this area. The search strategy excluded nanoflora and nanofauna because it sought a definition of nanotechnology that emphasized engineered science and technology rather than simply descriptions of small items in nature. In the case of Belcher, he publishes in oncological nursing journals; this is rather specialized to warrant inclusion in aforementioned nano search strategy. Belcher also does not have many publications in his full WOS/SCI record.

Table 1a: Coverage of Highly Creative Researcher's Full SCI Publication Record in Nanotechnology Subset

Highly Creative Researcher ID	Nanotechnology Dataset	Full WOS-SCI Record	Percent Coverage
102	206	256	80.5%
147	348	458	76.0%
101	201	284	70.8%
124	61	88	69.3%
151	127	184	69.0%
129	287	423	67.8%
136	126	186	67.7%
106	21	34	61.8%
141	59	96	61.5%
123	203	335	60.6%
111	118	195	60.5%
132	36	61	59.0%
120	117	204	57.4%
133	54	97	55.7%
103	179	344	52.0%
140	205	396	51.8%
126	199	386	51.6%
121	184	358	51.4%
115	16	32	50.0%
119	165	355	46.5%
144	119	258	46.1%
145	146	331	44.1%
112	95	217	43.8%
105	122	280	43.6%
138	93	222	41.9%
104	128	313	40.9%
114	66	168	39.3%
127	88	235	37.4%
128	18	50	36.0%
142	250	761	32.9%
137	66	212	31.1%
113	30	97	30.9%
148	40	136	29.4%
134	66	225	29.3%
122	98	342	28.7%
110	56	196	28.6%
125	75	263	28.5%

139	66	242	27.3%
143	149	600	24.8%
146	52	229	22.7%
130	119	554	21.5%
116	61	295	20.7%
150	103	573	18.0%
118	6	36	16.7%
131	6	41	14.6%
117	78	631	12.4%
107	10	111	9.0%
109	2	33	6.1%
108	10	213	4.7%
135	2	51	3.9%
149	325	1106	29.4%

N of cases=51

Table 1b: Coverage of Highly Creative Researcher's Full SCI Publication Record in Human Genetics Subset

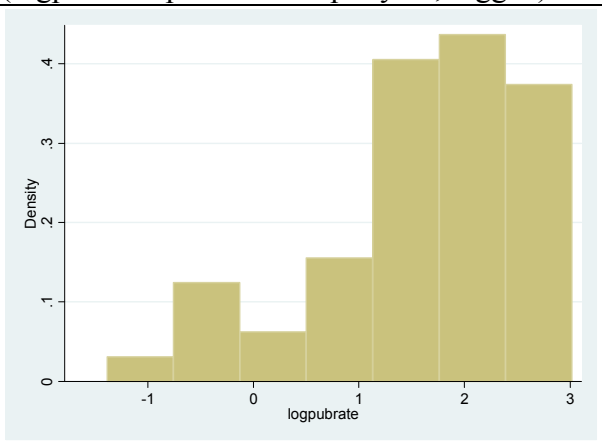
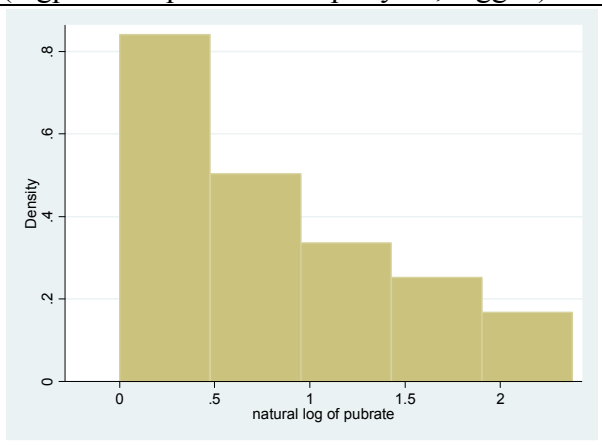
Authors – HCR ID	Human Genetics Dataset	Full WOS-SCI Record	Percent Coverage
225	216	389	55.50%
212	85	155	54.80%
202	30	59	50.80%
224	47	108	43.50%
217	102	251	40.60%
205	101	292	34.60%
219	75	242	31.00%
206	115	376	30.60%
215	42	160	26.30%
222	6	23	26.10%
218	12	52	23.10%
211	35	218	16.10%
216	14	113	12.40%
223	14	162	8.60%
204	26	315	8.30%
214	27	348	7.80%
220	17	261	6.50%
209	19	309	6.10%
213	11	191	5.80%
221	2	40	5.00%
210	7	144	4.90%

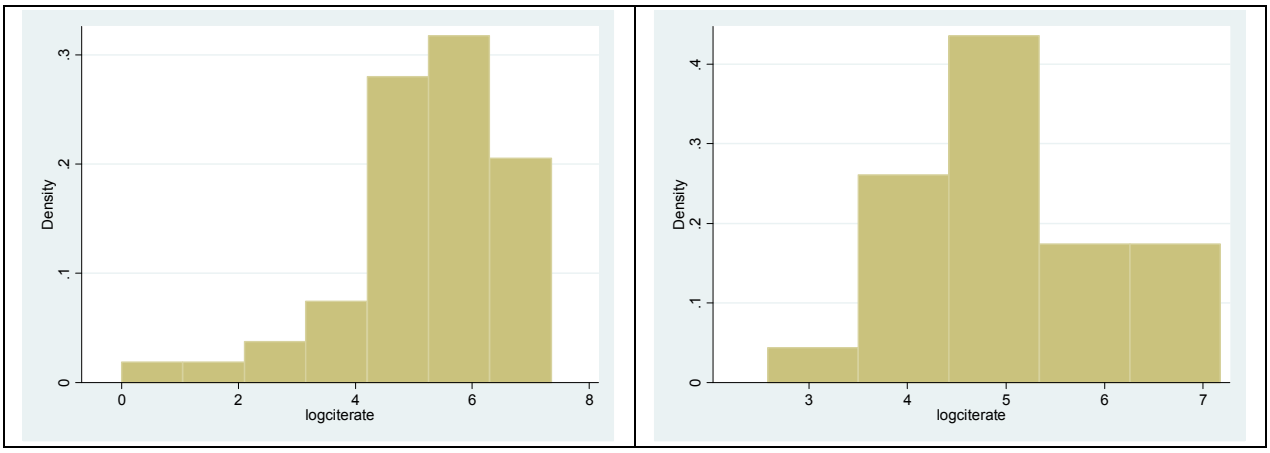
203	6	130	4.60%
201	5	266	1.90%
208	2	127	1.60%
207	27	2048	1.30%

N of cases=25

The second challenge is that the two distributions of publications of US and European highly creative researchers in the nanotechnology and human genetics domains exhibit different patterns of homogeneity and heterogeneity. Figure 1 presents histograms of publication and citations measures for highly creative researchers in nanotechnology and human genetics alongside one another. The nanotechnology publication and citation distributions associated with highly creative researchers in nanotechnology show signs of some clustering of researchers along the right hand side of the x-axis. In contrast, the human genetics distribution appears more spread out and heterogeneous. To some extent the differences could be a reflection of the larger sample size in the nanotechnology highly creative researcher subsample. Still, these distributional differences can influence the ability to identify matches for the highly creative researchers in each group.

Figure 1. Histograms of Publication Distributions of Highly Creative Researchers in the Nanotechnology and Human Genetics Domains: Publication Counts and Citations*

Nanotechnology Publications (logpubrate=publications per year, logged)	Human Genetics Publications (logpubrate=publications per year, logged)
	
Nanotechnology Citations (logciterate=citations per year, logged)	Human Genetics Citations (logciterate=citations per year, logged)



*Number of cases (highly creative researchers): nanotechnology=50; human genetics=25.

3. Results

We address the need for a matched comparison group, while taking on board the aforementioned methodological challenges, through two approaches. The first is a statistical approach based on propensity score modeling. The second is a "theory-based approach" grounded in the literature on early career patterns that emphasizes productivity and disciplinary structure (Simonton, 2004; Heinze and Bauer 2007; Burt 1999 see below).

Propensity score matching is a statistical approach in the manner of the classic experimental design. Propensity score matching compares a "treatment group" which in this case is highly creative researchers, with a relevant control group of researchers, with the caveat that assignment to these two groups is not random as in the classic experimental design. As previously discussed, this method is designed to reduce differences in observed characteristics between the two groups and is often used to evaluate program participation or other similar kinds of treatments. (Busom and Fernandez-Ribas, 2008) In this case, we are not evaluating are particular treatment, rather we seek to find matches between highly creative researchers and a comparison group, then – in a subsequent analysis - measure organizational and career mobility attributes of each to identify any differential influences of these types of meso level factors. Ideally, a matching process should use all observable characteristics for pairing treatment and control group researchers to reduce bias. The observable characteristics in this case are those within the researcher's publication record. This need for full specification of observable characteristics poses an issue, however, because some aspects of the publication record may be important for subsequent analysis of meso level factors, for example, co-publication networks. Thus, we seek to focus on publication record characteristics that will not preclude their subsequent use in analysis of meso level factor influences on creativity because they were used to effect the matching. For this matching we have focused on the citation, which is the number of times a paper has been cited aggregated to the author level. Citations are often considered to represent the influence and quality of a researcher's work on a scientific field, albeit not without issues such as self-citing, negative-citing, referee-inclusions, time lags, and the like.(Garfield, 1973; Narin and Hamilton, 1996; Kostoff, 2002; Glanzel, Thijs, and Schlemmer, 2004; Dosi, Llerena, and Labini, 2005; Aksnes, 2006) The challenges with using citations in analysis are well known and include (1) they are time related in that earlier articles have more opportunity to receive citations than do

recent articles, and (2) they are not normally distributed but rather follow a power curve with the majority of articles having no citations at all. (Newman, 2005; Clauset, Shalizi, and Newman 2007). We address these issues by estimating the "citation rate" or the natural log of the total number of citations of an author divided by the number of years of nanotechnology or human genetics publications of this author in the appropriate database.

Using this logged citation rate variable, we estimate the propensity score or probability of being a highly creative researcher. We perform this estimation to identify and match researchers outside this highly creative group that would have had a similar chance of being among the highly creative researchers. This analysis is performed with samples of 1,000 (and subsequently with a sample of 20,000) potentially matching researchers in nanotechnology and human genetics. All authors with fewer than two publications are excluded from these databases under the rationale that because article productivity is distributed with a long tail, there would be a number of authors with a single publication who would not likely match the highly creative researchers in this sample given the associations between productivity and creativity in previous studies (Simonton 2004; Heinze and Bauer 2007).

Propensity score modeling results are shown in Tables 2a and 2b. Initially, we estimated propensity scores with samples of 1,000 potential matches to highly creative researchers. The resulting propensity scores were divided into seven intervals in the case of nanotechnology and x intervals in the case of human genetics to optimally satisfy the balancing property of the algorithm. The 1,000 case analysis did not identify many good matches across the distribution. Among highly creative researchers in nanotechnology, only 12% fell into the lowest interval while more than 70% fell into the top three intervals. However, among the comparison group, 94% fell into the lowest interval and less than 1% into the highest interval. The pattern in human genetics was different still, with the highly creative human geneticists showing little clustering at the top intervals and some spread in the middle intervals, while the matched researchers were clustered in the lower intervals. We initially tried to address this lack of match by increasing the samples by a factor of 20, but this did not much change the results because power law distributions of citations and other similarly spread variables do not follow the Central Limit Theorem's assumptions of convergence toward normality under large sample size conditions. (Katz et al., forthcoming) We also tried other specifications that involved the introduction of additional variables: overall publication counts per year, number of journals, number of co-authors, and number of publications in Science and Nature. These specifications did not improve upon the use of citation rate and in many cases created out-of-balance situations. In sum, the propensity score approach we used was not judged useful for developing a matched sample in this situation.

Table 2a. Number of blocks of controls for matching highly creative researchers: Nanotechnology

Block*	Inferior of Prob(highly creative researcher)	Number of controls	Number of highly creative researchers	Total
Controls=1,000				
1	0	936	6	942
2	.1	30	2	32
3	.2	23	2	25
4	.3	5	5	10
5	.4	3	12	15
6	.6	2	10	12
7	.8	1	14	15
Controls=20,000				
1	0	19,110	5	19,115
2	.006	329	1	330
3	.012	255	2	257
4	.025	147	1	148
5	.05	88	5	93
6	.1	47	13	60
7	.2	18	10	28
8	.4	6	10	16
9	.6	0	4	4

*The optimal number of blocks is reported based on the algorithm developed by Becker and Ichino. The balancing property of the propensity score is satisfied.

Table 2b. Number of blocks of controls for matching highly creative researchers: Human Genetics

Block*	Inferior of Prob(highly creative researcher)	Number of controls	Number of highly creative researchers	Total
Controls=20,000**				
1	0	14,421	2	14,423
2	.001	2,105	3	2,108
3	.002	1,676	2	1,678
4	.003	996	6	1,002
5	.006	504	5	509
6	.012	202	3	205
7	.025	78	3	81
8	.05	16	1	17
9	.1	1	0	1
10	.2	1	0	1

*The optimal number of blocks is reported based on the algorithm developed by Becker and Ichino. The balancing property of the propensity score is satisfied.
**Analysis for human genetics 1,000 control sample has insufficient variation to support pscore analysis.

Table 3. Citation-based determinants of highly creative research: Marginal effects on the probability

Explanatory Variable	Marginal Effects	logL	Pseudo-R2	N
Logciterate (nano)	.94*** (.09)	-85.7	.58	1,051
Logciterate (nano)	.82*** (.06)	-162.7	.54	20,051
Logciterate (human genetics)	.51*** (.07)	-154.8	.19	20,025

Dependent variable: probability of being a highly creative researcher. Standard errors in parentheses.
* Significance at the 10% level ** Significance at the 5% level *** Significance at the 1% level.

Thus we move to the second approach, which is oriented around early career patterns. It is conjectured that highly creative and comparison researchers may have similar early career research patterns in the timing, quantity, and subject area of their initial publications. Later on they may diverge because of various characteristics including a hypothesized set of meso-level institutional and career mobility factors. The following early career characteristics were considered: (1) same first year of publication (2) same subject category of the first publication, (3) similar publication volume for the first six years (six years was chosen because an examination of the spread of articles suggested that this length of time was sufficient for amassing an early career record). The first category represents the importance of event-history research into creativity in terms of how certain time periods have been especially important in generating pathbreaking findings (Allison, 1984) such as the launch of Sputnik (Stokes, 1997) as well as how the timing within a scientific career is relevant for understanding creative events (Simonton, 1999). The second category represents the importance of disciplinary affiliation in understanding scientific creativity. Innovation is often thought to occur at the nexus of organizational boundaries. (Burt, 2004) one of which is the academic discipline. The Institute for Scientific Information (ISI) journal Subject Categories (SC) is a standard proxy for academic disciplines, and differences in cross-disciplinary linkages have been found by examining the citation patterns of articles in different SCs, with mathematics found to be less cross-disciplinary in its citation patterns than physics for example (Porter et al, 2009; Porter and Youtie 2009). The third category underscores the previously mentioned link between creativity and productivity (Simonton, 2004). In addition to these three criteria, we also consider continental affiliation — whether the researcher is in the US or Europe — to ensure a match of early career context.

Following this approach, we generated 8-10 initial matches for every highly creative research to account for non-response to our email queries for CVs in the subsequent phase of this research. It is important to note that all the authors that satisfy these criteria were eligible for the random sample we drew in the first approach, that is, they are they drawn from the same population. The match sample is thus composed of comparator researchers who have the same earliest year of publication, same subject category, similar publication volume at least at their early years of publishing in nanotechnology or human genetics, and the same continental affiliation as that of

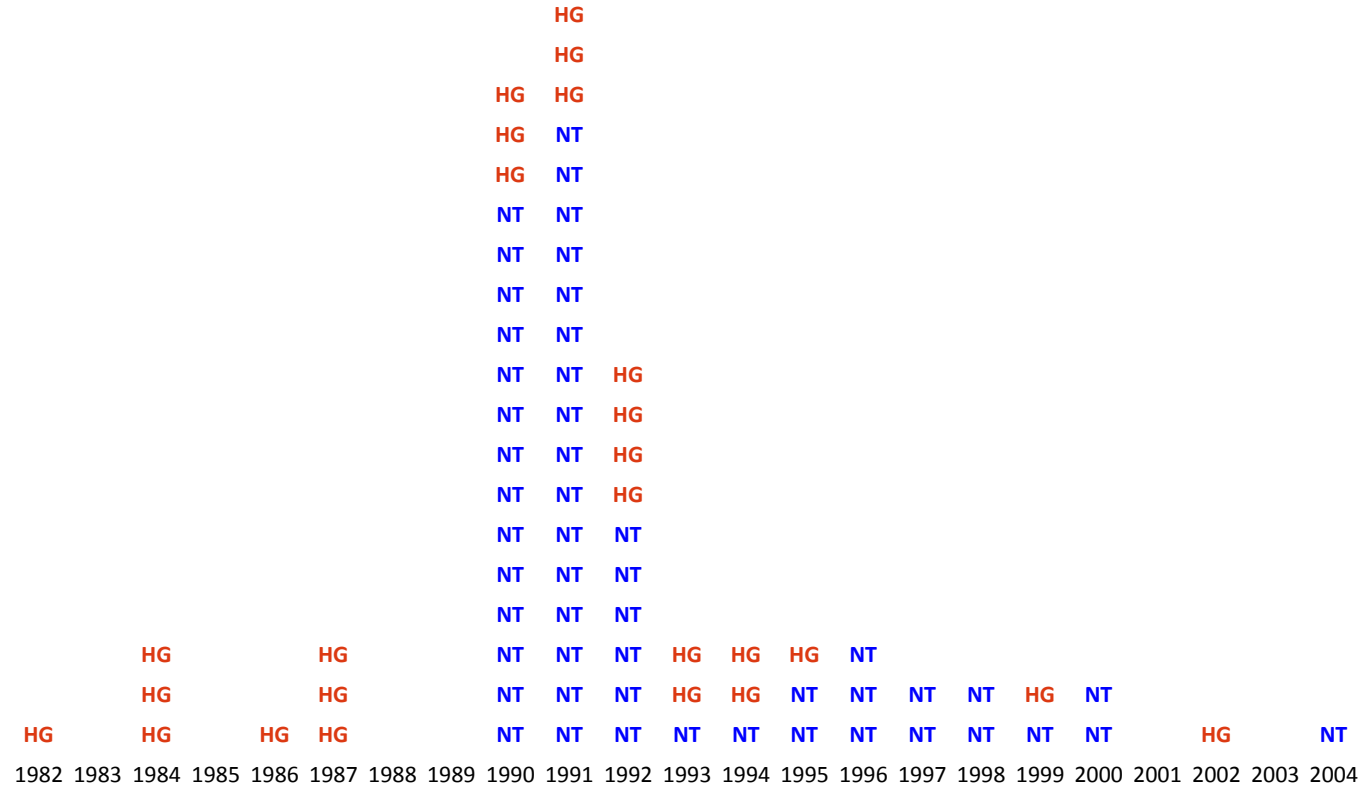
the highly creative researcher with whom they are associated. Because we are matching on four variables, many of the comparator researchers have the exact same early career characteristics as their highly creative researcher counterparts. For instances where there are more than 10 exact matches in the comparator group we have randomly selected 10. For example, one highly creative researcher had 29 exact matches, so we randomly selected 10 of these to populate the comparison group for this researcher. Roughly 20 of the 75 highly creative researchers had fewer than 8-10 exact matches on the four criteria described above. For these highly creative researchers, we expanded the publication counts by one or two publications on either side of the highly creative researcher's count, so if the highly creative researcher's early career publication count was 30 we sought matched researchers with publication counts of 28-32 for example. The final composition of the matching sample is NT = 510 and HG = 247.

Descriptive analyses of these three matching categories follow. The distribution of the first year of publication differs among HCRs between the two domains. Highly creative researchers in nanotechnology are observed to have first years that cluster in the early 2000's while those in human genetics are more heterogeneous across the 22-year timeframe. This difference is statistically significant ($p<.01$) using a Kolmogorov-Smirnov test. Figure 2 visually depicts this distributional difference.

The journal subject categories unsurprisingly also differ by domain. Genetics and Heredity represents for nearly two-thirds of the first publications of HCRs (64%), followed by Biochemistry & Molecular Biology (40%) and Cell Biology (32%).[2] Nanotechnology researcher's first publications are less dominated by one particular subject category. Physics represents for 29% of the first publications, followed by Chemistry (22%) and Materials Science (14%). Multidisciplinary journals such as Science and Nature were more likely to be the first publication of HCRs in nanotechnology, accounting for 16% of first publication journals while there was only one human genetics HCR with a first publication in a multidisciplinary journal. This difference certainly comports with the stated multidisciplinary nature of nanotechnology. (Porter and Youtie, 2009).

[2] A journal can be associated with more than one subject category. Multiple associations are especially common in the biosciences.

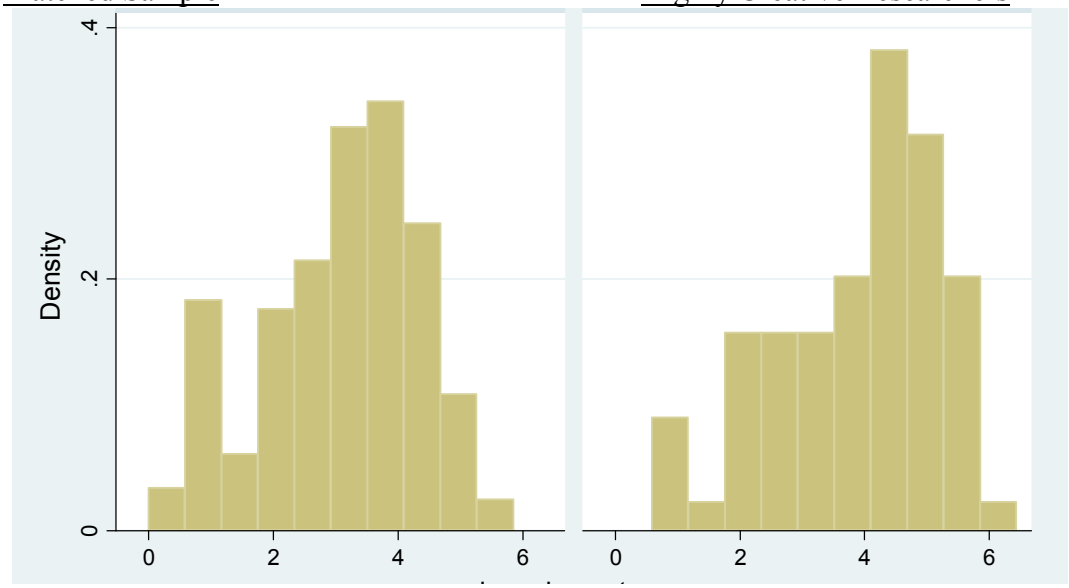
Figure 2. Distribution of Highly Creative Researchers by First Year of Publication (NT=Nanotechnology, HG=Human Genetics)



Number of cases=76 HCRs.

The HCRs and their comparison group were matched in terms of having the same or similar numbers of early career publications. Thus it is interesting to examine these two groups in light of the full publication record of the targeted domain that resulted across the entire career of the researchers under examination. Here we find that although the two groups had the same early career publication levels, HCRs had significantly more total publications (mean=86, median=66) and consequently more middle and later year volumes of publications than the comparison group (mean=41, median=27). This difference in total numbers of publications is statistically significant ($p<.01$) using a paired t-test of the logged distribution. (See Figure 3.) The results raise the question, why did the groups' productivity levels differ so dramatically after being the same in the first five years of their domain-specific careers? The explanation for this difference lies in factors beyond publication measures, which is why this matching analysis is a prelude to a subsequent effort that codes and analyzes additional information from the two groups' CVs.

Figure 3. Histogram of Logged Number of Full Career Publications in Targeted Domains: Highly Creative Researchers versus Comparison Group



Matched sample: Mean=41.0 (s.d. 46.4), Median=27; HCR: Mean=85.5 (s.d. 80.0), Median=66. Number of cases=76 HCRs, 757 matched researchers

This matching approach emphasizing the three early career attributes is expected to do a better job at achieving comparability between the HCR group and the non-HCR group than through statistically-based matching. It should be noted that when we apply the original propensity score specification based on the logged citation rate across the full domain-specific publication career of these HCRs and comparators, we similarly find that the comparison group does not provide a fully distributed match for the HCRs across the block distribution. The HCRs are again distributed across the blocks while the controls are clustered on the low end of the distribution.

Table 4. Number of Blocks of Controls and Marginal Effects: HCRs and Early Career Comparison Sample

Block	Inferior of Prob(highly creative researcher)	Number of controls	Number of highly creative researchers	Total
Nanotechnology				
1	0	492	8	500
2	.2	13	6	19
3	.4	3	13	16
4	.6	2	10	12
5	.8	0	14	14
Human Genetics				
	0	229	8	237
	.2	12	9	21
	.4	3	3	6
	.6	3	3	6
	.8	0	2	2

Explanatory Variable	Marginal Effects	logL	Pseudo-R2	N
Logciterate (nano)	.89*** (.09)	-88.9	.48	561
Logciterate (human genetics)	.92*** (.15)	-83.5	.30	272

While statistical tests suggest challenges in full career matching efforts, there are some signs of positive achievement. For example, the productivity distribution of the matched sample in Figure 3 broadly resembles that of the highly creative researchers. This similarity does suggest some degree of success in developing a matched group using early career characteristics.

3. Conclusions

This research contributes to efforts to understand the factors which encourage highly creative research. Previous work in this area has been challenged to construct a sufficiently similar comparison group because of the exceptional performance endemic to highly creative researchers. Most previous work draws on unobtrusive measures such as publication data. That is the case with this study and constitutes a limitation in the lack of information with which to match HCR and comparison groups using the publication record alone. This model specification issue underlies the need for other datasets, which is why we plan to collect and code variables from the CVs of the HCRs and comparison group. On the other hand, it is not uncommon for efforts at framing comparison groups to rely on unobtrusive measures so as to avoid prior influence on the groups.

Another limitation is that of truncation. Since we are not using the full record of the individual, we are only providing information about the target field of interest as defined through keywords and journal names, so distortion is introduced. From the point of view of understanding productivity and creativity, this truncation presents a distorted picture although it is a reasonable convention to use in understanding activities in emerging scientific domains.

The results highlight some of the issues in trying to match highly creative and comparison group researchers. Propensity score matching allowed us to create models which were statistically significant using the researchers' (logged) number of citations of in-domain publications divided by the number of years of active publications within the domain. The logged citation rate variable, while useful in model development, was not able to result in a distribution that could pinpoint sufficient matches in the comparison group, especially at the middle and higher ends of the distribution. The lack of distributional matching was again seen in our application of the propensity score model to a comparison group researchers who had, relative to their most proximate HCR, the exact same (or very similar) number of publications, year of first publication, and journal subject category of first publication. Despite this statistical finding, graphs of the matched and HCR groups did show broadly similar patterns, signifying some level of success in using this early career technique for comparison group creation.

We further note that there is much diversity in the HCR treatment group. The target HCRs are very differ in terms of publication counts, citations, linkages with other researchers, and the like. This extent of difference is especially the case for highly creative human genetics scientists. These scientists do not exhibit homogenous clustering around certain values in the distribution of indicators such as productivity and first year of publications, rather the highly creative human geneticists tend to be widely dispersed across the scales of indicators employed in this analysis. The extent of diversity can make it is difficult to find a "group" among these creative researchers with which to compare. Indeed, Heinze et al (2009) has found from case studies of 20 highly creative researchers in nanotechnology and human genetics that highly creative researchers take distinctive paths to success, while at the same time there are common organizational factors involves such as the size of the group, availability of complementary technical skills, access to extramural resources, and good leadership. It is hoped that having a thoughtfully crafted comparison group will enable systematic identification of these and other factors in terms of their distinctive relationship to scientific creativity in two emerging fields, to the ultimate benefit of university and faculty and industrial R&D management, funding organizations, and national research policy.

References

Aksnes D, (2006), Citation rates and perceptions of scientific contribution. Journal of the American Society for Information Science and Technology, 57 (2), pp. 169-185.

Allison, PD, (1984), Event History Analysis: Regression for Longitudinal Event Data, Newbury, Park, CA: Sage Publications.

Burt, R. S. (2004), Structural holes and good ideas, American Journal of Sociology 110 (2), 349-399.

Busom, I. and Fernández-Ribas, A. (2008) The Impact of participation in R&D Programs on R&D partnerships, Research Policy, 37 (2), pp. 240-257.

Clauset, A. Shalizi, C., and Newman, M., 2007, Power-law Distributions in Empirical Data. SIAM Review, to appear (2009). (Preprint at arxiv:0706.1062).

Dosi, G., Llerena, P., and Labini, M., 2005, June, Evaluating and Comparing the innovation performance of the United States and the European Union. Paper prepared for the TrendChart Policy Workshop.Brussels, Belgium

Garfield, E, 1973, "Citation Analysis as a Tool in Journal Evaluation," Science, New Series 178, pp. 471-479

Gladwell, M., 2008. Outliers: The Story of Success. New York: Little Brown.

Glanzel W, Thijs B, and Schlemmer B., 2004, A bibliometric approach to the role of author self-citations in scientific communication, Scientometrics 59 (1), pp. 63-77

Heinze, T., Shapira P., 2006. Research Creativity: Analyses of Unconventional and Path-breaking Solutions in Science, paper presented before SPRU 40th anniversary conference, Brighton, UK, 11-13 September 2006.

Heinze T., Bauer G., 2007, Characterizing Creative Scientists in Nano S&T: Productivity, Multidisciplinarity, and Network Brokerage in a Longitudinal Perspective, Scientometrics, 70, pp. 811-830.

Heinze, T., Shapira, P., Rogers, J., Senker, J., 2007. Creativity Capabilities and the Promotion of Highly Innovative Research in Europe and the United States: Final Report. Twente Netherlands: University of Twente.

Heinze, T., Shapira, P., Rogers, J., Senker, J., 2009. Organizational and institutional influences on creativity in scientific research. Research Policy 38 (4), 610-623.

Katz, S., Rogers, J., Hicks, D., (forthcoming). Citation distributions to scientific papers: 1996-2007. Working paper.

Kostoff, R., 2002, Citation analysis of research performer quality, Scientometrics, 53 (1), pp. 49-71

Lee, W., 2006, Propensity Score Matching and Variations on the Balancing Test. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=936782#.

Merton, R.K. 1973. The Sociology of Science: Theoretical and Empirical Investigations. Chicago: Univeristy of Chicago Press.

Narin F. and Hamilton, K., 1996, Bibliometric performance measures, Scientometrics, 36 (3), pp. 293-310

National Research Council, 2006, Evaluation of the Markey Scholars Program. Washington DC: The National Academies Press.

Newman, M., 2005, Power Laws, Pareto Distributions, and Zipf's Law, Contemporary Physics, 46 (5), pp. 323 – 351.

Pion G. and Cordray, D., 2008. The Burroughs Wellcome Career Awards in the Biomedical Sciences: Challenges to and Prospects for Estimating the Causal Effects of Career Development Programs. Evaluation & the Health Professions 31 (4), 335-369.

Porter, A.L., Youtie, J., Shapira, P., and Schoeneck, D.J., 2008. *Journal of Nanoparticle Research*, **10** (5), 715-728.

Porter, A.L., and Youtie, J., 2009, How interdisciplinary is nanotechnology?, Journal of Nanoparticle Research, **11** (5), 1023-1041.

Porter, A.L., Roessner, J.D., and Heberger, A.E. (2009) How Interdisciplinary is a Given Body of Research? Research Evaluation, forthcoming.

Rafols, I. and Meyer, M. Diversity and Network Coherence as indicators of interdisciplinarity: case studies in bionanoscience, *Scientometrics* (Forthcoming)

Rosenbaum, P. and Rubin, D. (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," Journal of the American Statistical Association, 79, pp. 516-524.

Rubin, D. (1997). "Estimating Causal Effects from Large Data Sets using Propensity Scores." Annals of Internal Medicine, 127, pp. 757-763.

Simonton, D, 1999, Origins of Genius: Darwinian Perspectives on Creativity, New York: Oxford University Press.

Simonton, D, 2004, Creativity in Science: Chance, Logic, Genius, and Zeitgeist, Cambridge: Cambridge University Press.

Stokes, D, 1997, Pasteur's Quadrant: Basic Science and Technological Innovation. Washington DC: Brookings.

Appendix: Search Strategy from Publications in WOS-SCI

Human Genetics: 1970-2006.

SEARCH TERM
1. ts= (inborn errors or connective tissue disorders or Duchenne muscular dystrophy or congenital anomalies or prenatal diagnosis or neural tube defects or folic acid and neural tube defects or mitochondrial disorders or Waardenburg* Syndrome or congenital anomalies or thalassemia or beta-thalassemia or congenital adrenal-hyperplasia or Gaucher disease or lysosomal storage disease)
2. ts= (gene cloning or positional cloning or structural genes or gene positioning or linkage disequilibrium or Chromosome or genomic imprinting or mitochondrial inheritance or heteroplasmy or DNA polymorphisms or developmental genetics or population genetics or artifical chromosomes or genomics or proteomics or DNA profiling)
3. ts= (autosomal recessive transmission or somatic mosaicism or collagen genes or human genome or cystic fibrosis or linkage analysis or linked genetic transmission or chromosomal abnormalities or chromosomal mosaicism or Down* syndrome or multifactorial inheritance or genetics of common disease or hypertension and genetics or maternal genetics transmission or cancer genetics or retinoblastoma or proto-oncogenes or suppresor genes or human dysmorphology or renal poliquistosis or genomics library or fragile X syndrome or pathogenic mutations)
4. #3 OR #2 OR #1
5. SO=(ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY) OR SO=(ACTA CRYSTALLOGRAPHICA SECTION E STRUCTURE REPORTS ONLINE) OR SO=(BIOCHEMICAL "AND" BIOPHYSICAL RESEARCH COMMUNICATIONS) OR SO=(BIOCHIMICA ET BIOPHYSICA ACTA BIOENERGETICS) OR SO=(BIOMETRICS) OR SO=(BLOOD) OR SO=(BONE MARROW TRANSPLANTATION) OR SO=(CELL) OR SO=(CHEMICAL PHYSICS) OR SO=(CIRCULATION) OR SO=(CURRENT BIOLOGY) OR SO=(CYTOGENETICS "AND" CELL GENETICS)
6. SO=(DEVELOPMENTAL BIOLOGY) OR SO=(DEVELOPMENTAL CELL) OR SO=(DEVELOPMENTAL CELL) OR SO=(DEVELOPMENT) OR SO=(DIABETES) OR SO=(EMBO JOURNAL) OR SO=(EMBO REPORTS) OR SO=(FASEB JOURNAL) OR SO=(FLEISCHWIRTSCHAFT) OR SO=(FREE RADICAL BIOLOGY "AND" MEDICINE) OR SO=(JOURNAL OF APPLIED PHYSICS) OR SO=(JOURNAL OF BACTERIOLOGY) OR SO=(JOURNAL OF BIOLOGICAL CHEMISTRY) OR SO=(JOURNAL OF CELL BIOLOGY) OR SO=(JOURNAL OF CELLULAR BIOCHEMISTRY) OR SO=(JOURNAL OF CHEMICAL PHYSICS) OR SO=(JOURNAL OF CLINICAL INVESTIGATION) OR SO=(JOURNAL OF GENERAL PHYSIOLOGY) OR SO=(ADVANCES IN LIPID RESEARCH) OR SO=(JOURNAL OF MOLECULAR BIOLOGY) OR SO=(JOURNAL OF PHYSICS "AND" CHEMISTRY OF SOLIDS) OR SO=(JOURNAL OF PHYSIOLOGY LONDON) OR SO=(JAMA JOURNAL OF THE AMERICAN MEDICAL

ASSOCIATION)
7. SO=(MECHANISMS OF DEVELOPMENT) OR SO=(MEDICINE) OR SO=(MOLECULAR BIOLOGY OF THE CELL) OR SO=(NEW ENGLAND JOURNAL OF MEDICINE) OR SO=(NUCLEIC ACIDS RESEARCH) OR SO=(PFLUGERS ARCHIV EUROPEAN JOURNAL OF PHYSIOLOGY) OR SO=(PROCEEDINGS OF THE BIOLOGICAL SOCIETY OF WASHINGTON) OR SO=(PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA) OR SO=(REPRODUCTIVE BIOMEDICINE ONLINE) OR SO=(TETRAHEDRON) OR SO=(BIOPHYSICAL JOURNAL)
8. #7 OR #6 OR #5
9. #8 AND #4
10. Restrict 9 to Articles Only

Source: Based on Heinze et al., 2007.

Nanotechnology: 1990-2006
Phase 1

Search	**Terms**	**RESULT: SCI 2005 as of 4/22/06**
1. Nano*	nano*	39101
2. Quantum	(quantum dot* OR quantum well* OR quantum wire*) NOT nano*	3633
3. Self-Assembly	(((SELF ASSEMBL*) or (SELF ORGANIZ*) or (DIRECTED ASSEMBL*)) AND MolEnv-I) NOT nano*	3532
4. Terms to include as Nano without other delimiters	((molecul* motor*) or (molecul* ruler*) or (molecul* wir*) or (molecul* devic*) or (molecular engineering) or (molecular electronic*) or (single molecul*) or (fullerene*) or (coulomb blockad*) or (bionano*) or (langmuir-blodgett) or (Coulomb-staircase*) or (PDMS stamp*)) NOT nano*	3550
5. Microscopy - terms to include but limit to the molecular environment	((TEM or STM or EDX or AFM or HRTEM or SEM or EELS) or (atom* force microscop*) or (tunnel* microscop*) or (scanning probe microscop*) or (transmission electron microscop*) or (scanning electron microscop*) or (energy dispersive X-ray) or (X-ray photoelectron*) or (electron energy loss spectroscop*)) AND MolEnv-I) NOT nano*	11665
6. Nano-pertinent; Limit to the Molecular Environment - More Inclusively	(pebbles OR NEMS OR Quasicrystal* OR (quasi-crystal*)) AND MolEnv-I) NOT nano*	128

7. Nano-pertinent; limit to the Molecular Environment - More Restrictive	(biosensor* or (sol gel* or solgel*) or dendrimer* or soft lithograph* or molecular simul* or quantum effect* or molecular sieve* or mesoporous material*) AND (MolEnv-R)) NOT nano*	2104
	1 or 2 or 3 or 4 or 5 or 6 or 7	**61173**
8. Additional Items in Nano Journals	fullerene* or ieee transactions on nano* or journal of nano* or nano* or materials science & engineering C - biomimetic and supramolecular systems (in JOURNAL title field) NOT nano*	506
MolEnv-I (inclusive)	(monolayer* or (mono-layer*) or film* or quantum* or multilayer* or (multi-layer*) or array* or molecul* or polymer* or (co-polymer*) or copolymer* or mater* or biolog* or supramolecul*)	>100000
Or		
MolEnv-R (more restrictive	(monolayer* or (mono-layer*) or film* or quantum* or multilayer* or (multi-layer*) or array*)	78390
Total	**1 or 2 or 3 or 4 or 5 or 6 or 7 or 8**	**61479**

PHASE 2. Exclusions from Nano*

Terms excluded from Search #1 (Nano*) – these records are deleted from dataset.

Exclusion Terms	
Records containing these terms should be removed from "Nano*" dataset	**Exclude any nano* records containing only one of these terms and no other nano terms**
plankton*	nanometer*
n*plankton	nanosecond*
m*plankton	nanomolar*
b*plankton	nanogram*
p*plankton	nanoliter*
z*plankton	nano-second
nanoFlagel*	nano-meter
nanoAlga*	nano-molar
nanoProtist*	nano-gram
Nanofauna*	nano-liter
Nano*aryote*	
Nanoheterotroph*	
Nanophtalm*	
Nanomeli*	
Nanophyto*	
Nanobacteri*	
nano2*, nano3*, nanos_, nanog_, nanor_, nanor_, nanoa_, nanoa_, nano-, nanog-, nanoa-, nanor-	

Source: Porter et al., 2008.

MOD: Measurement and Analysis of Highly Creative Research in the US and Europe (CREA II) NSF Award 0738126

Note on Methods for Developing a Matching Control Group for Highly Creative Researchers

Philip Shapira*, Jan Youtie*, Juan Rogers*, Thomas Heinze**, Reynold Galope*
* Georgia Institute of Technology; ** University of Bamberg

September 30, 2008

This note details the methods explored and adopted for identifying a matching control group for highly creative researchers (HCRs) in the US and Europe identified in the CREA I study (51 in nanotechnology; 25 in human genetics).[1] In the CREA II research, the organizational and career profiles of these HCRs will be compared and contrasted with those of the matches.

1. THE DATASET

The Human Genetics dataset was downloaded from the Web of Science in early March 2008 using the same search strings used in CREA I[2] and (b) the Nanotechnology dataset was obtained from the existing Nanotech database of STIP. Originally, both datasets have a timeline between 1990 and 2006.

a. The Human Genetics dataset has 88,207 raw publication records of 193,147 authors in 159 journals.
b. The Nanotechnology dataset has 406,967 raw publication records of 350,943 authors in 4,993 journals.

The HG dataset was expanded in June 2008 by extending the timeline to 1970 and including the top 5 journals of HG HCRs that were not part of the original WOS search.

2. DEFINITION OF THE MATCHED SAMPLE

We are treating CREA II as an observational study, i.e. "an empirical investigation of treatments, policies, or exposures and the effect they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects (Rosenbaum, 2002)."

The primary methodological challenge of observational studies is making the treatment and comparison groups homogenous except in the key variable of interest. Matching provides a way of controlling for observable heterogeneity by finding "look-alikes" of the members of the treatment group (Essama-Nssah, 2006). In the case of CREA II, we need to have at least 76 non-HCRs who have the same observable characteristics as that of the 76 HCRs previously defined by CREA I. Achieving homogeneity of observable characteristics between the HCR group and the non-HCR

[1] See: Heinze, T., Shapira, P., Senker, J., and Kuhlmann, S., "Identifying Creative Research Accomplishments: Methodology and Results for Nanotechnology and Human Genetics," Scientometrics, Vol. 70, No. 1, 2007, pp. 125-152
[2] The search strings are on page 106-108 of the CREA 1 Final Report.

group would ensure that the coefficients of the treatment effects model (which will be fitted later on) are unbiased estimates of the impacts of meso-level factors on scientific creativity.

Prior to picking the members of the comparison group, a key challenge is defining the population from which to draw this matched sample of non-HCRs. Picking the matches of the HCRs and defining the population from which to draw these matches are entirely separate (albeit interconnected) activities[3]. Random sampling from the whole population of the NT and HG datasets was immediately ruled out because of the distribution of the entire dataset. Random sampling from a dataset in which almost 60 percent of the authors have only 1 publication may not guarantee good matching quality between the HCR group and the non-HCR group.

Approaches

We explored at least four (4) strategies to define the population from which the comparison group is to be drawn.

a. The first option was to identify the top journals of the HCRs, i.e. the journals where the HCRs have published most frequently. This strategy yielded the following results for human genetics and nanotechnology in the United States:

Human Genetics		**Nanotechnology**	
1. Nature Genetics	73	Abstr. Pap. Am. Chem. Soc	399
2. American Journal of Human Genetics	49	J. Am. Chem. Soc	286
3. Genomics	43	Nano Lett	191
4. Human Molecular Genomics	41	Phys. Rev. B	
5. Genome Research	20	J. Phys. Chem. B	163

b. The second strategy was to find the earliest publications of the HCRs and draw the comparison group from all authors who published in the same journal in the same year when the HCRs started publishing. The obvious limitation of this strategy is the timeline of the two datasets which is 1990-2006. Using this strategy, most of the HCRs were "forced" to have their first publications in HG and NT in the early 1990s[4]. Publications prior to 1990 were effectively ignored. The search results for HG for the United States and Europe are as follows:

Europe HCR	**First Publication**	**US HCR**	**First Publication**
Bickmore, Wendy	Genetics 1991	Collins, F S	Genomics 1990
Fischer, Alain	Human Genetics 1992	Lander, E S	Genetics 1991
Gruss, Peter	Genomics 1991	Sheffield, V C	American Journal of Human Genetics 1992

[3] Picking the final matches for every HCR will be done statistically through PSM using the individual and organizational data from the CVs and publication data from WOS of the HCRs and the non-HCRs in the matched sample.
[4] This is the main reason why the HG dataset was extended to 1980.

Hoeijmakers, Jan	American J. of Human Genetics 1990		American Journal of Human Genetics 1992
Horsthemke, Bernhard	Clinical Genetics 1990	WARREN, S T	Genomics 1990
Jeffreys, Alec J.	Human Heredity 1990	Jaenisch, R	Mammalian Genome 1992
Jentsch, Thomas	Journal of Medical Genetics 1993	Botstein, D	Genes and Development 1991
Larsson, Nils-Göran	American J. of Human Genetics 1993	Venter, J C	American Journal of Human Genetics 1991
Mandel, Jean-Louis	American J. of Human Genetics 1990	Blackburn, E H	Genes and Development 1994
Sulston, John E.	Nature Genetics 1992	Ames, B N	Gene 1998

c. We also explored the idea of matching HCRs and non-HCRs based on propensity scores. In this case, the propensity score is the predicted probability of being categorized as an HCR conditional on a set of covariates. It was decided that the values of these covariates would be generated from the web of science (WOS). The assumption here is that WOS variables like publication count, journal count, citation count, co-author count, etc are significant predictors of the said conditional probability. Initially, WOS data were gathered only for 1,000 authors each for NT and HG. PSM however was not successful because of the lack of matches for the HCRs; there were significant differences between the propensity scores of non-HCRS and HCRs. It was then decided to expand the sample to 20,000 authors. The expectation was that the 20-fold increase in the sample will result to a higher number of potential matches for the HCRs. However, the expectation was not met. The result was the same: most HCRs have high propensity scores that only a few non-HCRs can match. Moreover, some HCRs have low propensity scores which may indicate that using WOS variables as explanatory variables was inadequate for propensity score matching.

d. The fourth approach is not entirely different from the previous approaches. It in fact revisited the first three approaches and redefined the population from which to draw the matched sample on four criteria: (1) first year of publication of the HCRs, (2) subject category of the first publication, (3) publication volume for the first six years, and (4) continental (i.e. US or EU) affiliation. In general, it proceeded as follows:

   i. Determine earliest publication of HCRs in the Human Genetics (HG) and Nanotechnology (NT) datasets. [Note that for the HG dataset, this was extended back to 1980.]
   ii. Obtain year and subject category(ies) of these earliest publications.
   iii. Identify all non-HCRs whose earliest publications fall on the same year and subject category(ies) as that of the earliest publications of the HCRs. For example, Ajayan had his first NT publication in 1991 and in the subject category Applied Physics. All non-HCRs who also had their first publication in 1991 and in Applied Physics are isolated through VantagePoint and Excel.
   iv. Obtain the publication volume for the first six years of HCRs and all non-HCRs identified in step iii., that is, if the earliest year of publication is 1991, get publication count from 1991 to 1996.
   v. Randomly select 20-40 non-HCRs who have the same first year of publication, subject category of the first publication, and publication volume for the first six years as the HCRs and obtain their addresses. Only non-HCRs who are currently working in the same region or continent as the HCRs are included in the matched sample.

3. THE MATCHES

Following the fourth approach defined above, we generated 8-10 initial matches for every HCR. This match sample is thus composed of non-HCRs who have the same earliest year of publication, same subject category, same publication volume at least at their early years of publishing in NT or HG, and the same region as that of the HCRs. While this theory-driven sample would not necessarily guarantee that the observable characteristics (e.g. age, total publication volume in NT or HG, organizational affiliations) of the final matches will be the same as that of the HCRs, it represents an improvement over the random sampling approach. Matching based on the four criteria or variables can potentially include non-HCRs who belong to the same age group or cohort, have the same research interests, have the same publication productivity, and have faced approximately the same cultural and institutional work environments as the HCRs have had. This theory-driven match sample is thus expected to do a better job at achieving homogeneity between the HCR group and the non-HCR group than a random match sample.

The distribution of the match sample in terms of continents and countries are as follows:

Human Genetics

Continent	HCR		NON-HCR MATCH	
	Frequency	Percent	Frequency	Percent
EU	14	56.00	140	56.22
US	11	44.00	109	43.78
Total	25	100.00	249	100.00
Country	**HCR**		**NON-HCR MATCH**	
	Frequency	**Percent**	**Frequency**	**Percent**
Belgium	0	0	7	2.81
Czech Republic	0	0	1	0.40
Denmark	1	4.00	3	1.20
Finland	0	0	5	2.01
France	3	12.00	25	10.04
Germany	4	16.00	18	7.23
Hungary	0	0	1	0.40
Iceland	0	0	1	0.40
Ireland	0	0	1	0.40
Italy	0	0	13	5.22
Netherlands	1	4.00	12	4.82
Norway	0	0	1	0.40
Poland	0	0	2	0.80
Spain	0	0	8	3.21
Sweden	1	4.00	3	1.20
Switzerland	0	0	6	2.41
UK	4	16.00	33	13.25
US	11	44.00	109	43.78
Total	25	100	249	100.00

Nanotechnology

Continent	HCR		NON-HCR MATCH	
	Frequency	Percent	Frequency	Percent
EU	22	43.14	215	46.44
US	29	56.86	248	53.56
Total	51	100.00	463	100.00
Country	**HCR**		**NON-HCR MATCH**	
	Frequency	**Percent**	**Frequency**	**Percent**
Austria	1	1.96	9	1.94
Belgium	0	0	4	0.86
Bulgaria	0	0	1	0.22

Czech Republic	0	0	2	0.43
Denmark	0	0	5	1.08
Finland	0	0	2	0.43
France	4	7.84	27	5.83
Germany	11	21.57	48	10.37
Greece	0	0	1	0.22
Hungary	0	0	4	0.86
Italy	0	0	28	6.05
Netherlands	1	1.96	10	2.16
Norway	0	0	1	0.22
Poland	0	0	2	0.43
Portugal	0	0	1	0.22
Russia	0	0	7	1.51
Spain	0	0	16	3.46
Sweden	0	0	5	1.08
Switzerland	2	3.92	11	2.38
UK	3	5.88	30	6.48
Ukraine	0	0	1	0.22
US	29	56.86	248	53.56
Total	51	100.00	463	100.00

For the complete list of matches as of 30 September 2008, please see the appendix 1 and 2.

4. NEXT STEPS

Please see schedule of the survey and data analysis of this research project below:

Activity	Timeframe
a. Final List of the Matched Sample (approximately 760 researchers/scientists) including their contact addresses	October 11
b. Survey/Gathering of CVs of the Matched Sample	October 20- December 31, 2008
c. CV Coding and Data Analysis	January to March 2009
d. Report Writing	April 2009