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

Jan Youtie
Enterprise Innovation Institute
Georgia Institute of Technology
Atlanta, GA 30332-0640, USA

Philip Shapira
Manchester Institute of Innovation Research, Manchester Business School
University of Manchester
Manchester, M13 9PL, UK
and School of Public Policy
Georgia Institute of Technology
Atlanta, GA 30332-0345, USA.

Juan Rogers
School of Public Policy
Georgia Institute of Technology
Atlanta, GA 30332-0345, USA.

***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.**

I. 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 [1,2], 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 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

We gratefully acknowledge support from NSF under award number SBE-0738126. This work also draws on earlier research funded by the European Commission under award number EU-NEST/CREA-511889. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or the European Commission

the relationship between quantity of publications and quality in terms of citations. [3] This argument is popularized in Gladwell's 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.[4] 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 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.[5] 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 evaluation of the Markey Scholars program conducted just this type of matching.[6] 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.[7,8] 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. [9] Unobserved differences are not accounted for in propensity score matching, unlike in the case of randomized experimental designs. Pion and Cordray 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). [10] 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.

II. 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. [11]. 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. [12, 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 has been shown to be an emerging interdisciplinary field [13,14] 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 in the case of nanotechnology and 1970-2006 in the case of human genetics. [15,12]

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 [16] 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, a few highly creative researchers have published in journals that are not well covered by the domain definition used in this study. [15] The search strategy 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 another under-covered researcher, this researcher publishes in oncological nursing journals which is a rather specialized field and 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%

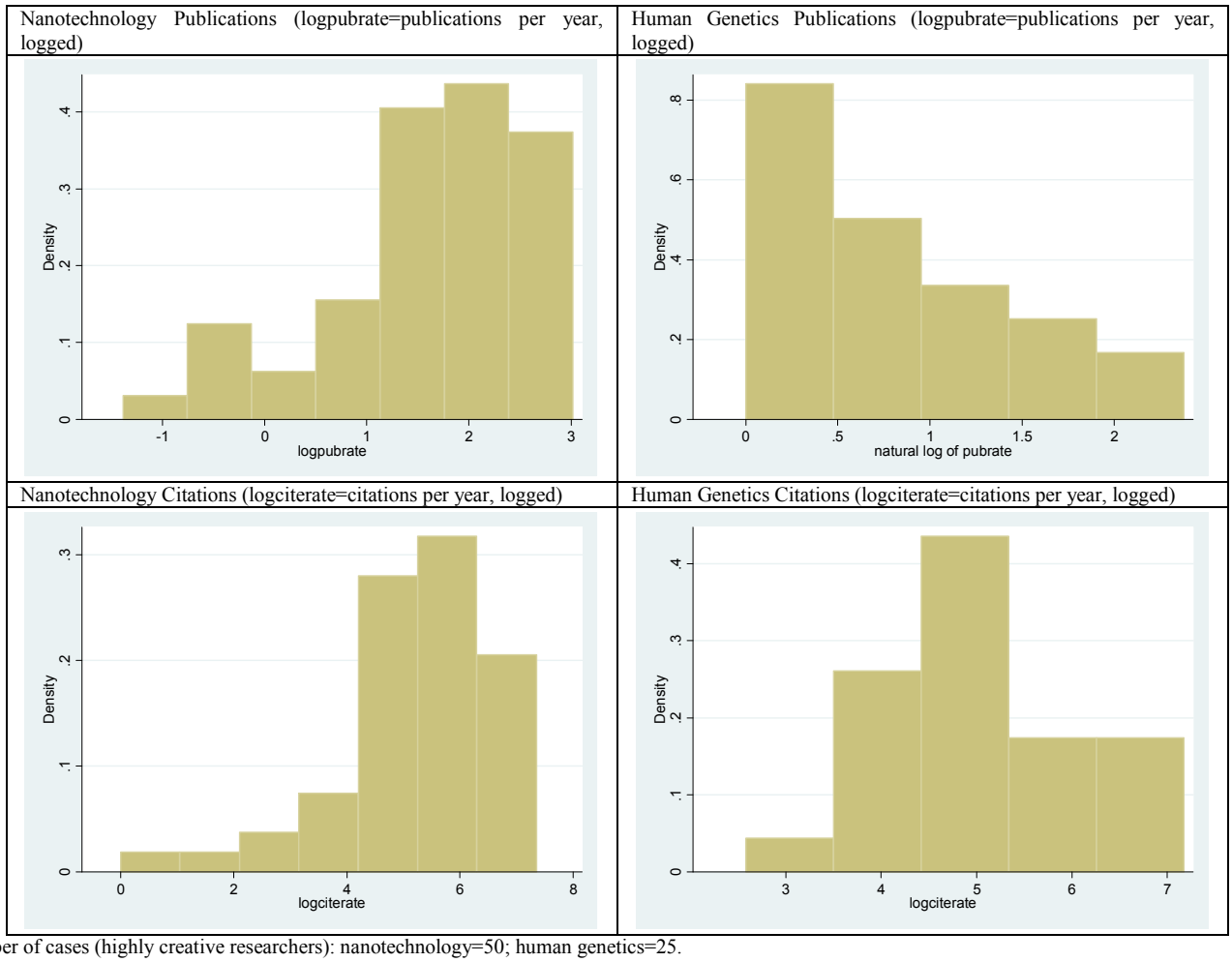
N of cases=51

TABLE 1B
COVERAGE OF HIGHLY CREATIVE RESEARCHER'S FULL SCI PUBLICATION RECORD IN HUMAN GENETICS SUBSET

Highly Creative Researcher (ID)	Human Genetics Dataset	Full WOS-SCI Record	Percent Coverage
225	216	389	55.5%
212	85	155	54.8%
202	30	59	50.8%
224	47	108	43.5%
217	102	251	40.6%
205	101	292	34.6%
219	75	242	31.0%
206	115	376	30.6%
215	42	160	26.3%
222	6	23	26.1%
218	12	52	23.1%
211	35	218	16.1%
216	14	113	12.4%
223	14	162	8.6%
204	26	315	8.3%
214	27	348	7.8%
220	17	261	6.5%
209	19	309	6.1%
213	11	191	5.8%
221	2	40	5.0%
210	7	144	4.9%
203	6	130	4.6%
201	5	266	1.9%
208	2	127	1.6%
207	27	2048	1.3%

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.



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

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

III. 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 [3, 5, 17].

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. [18]. 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. [19,20,21,22,23,24] 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. [25,26]. 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 [3,5].

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. [27] 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[a]	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

[a]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[a]	Inferior of Prob(highly creative researcher)	Number of controls	Number of highly creative researchers	Total
Controls=20,000[b]				
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

[a]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.
[b]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.
* Significant at the 10% level ** Significant at the 5% level *** Significant at the 1% level.

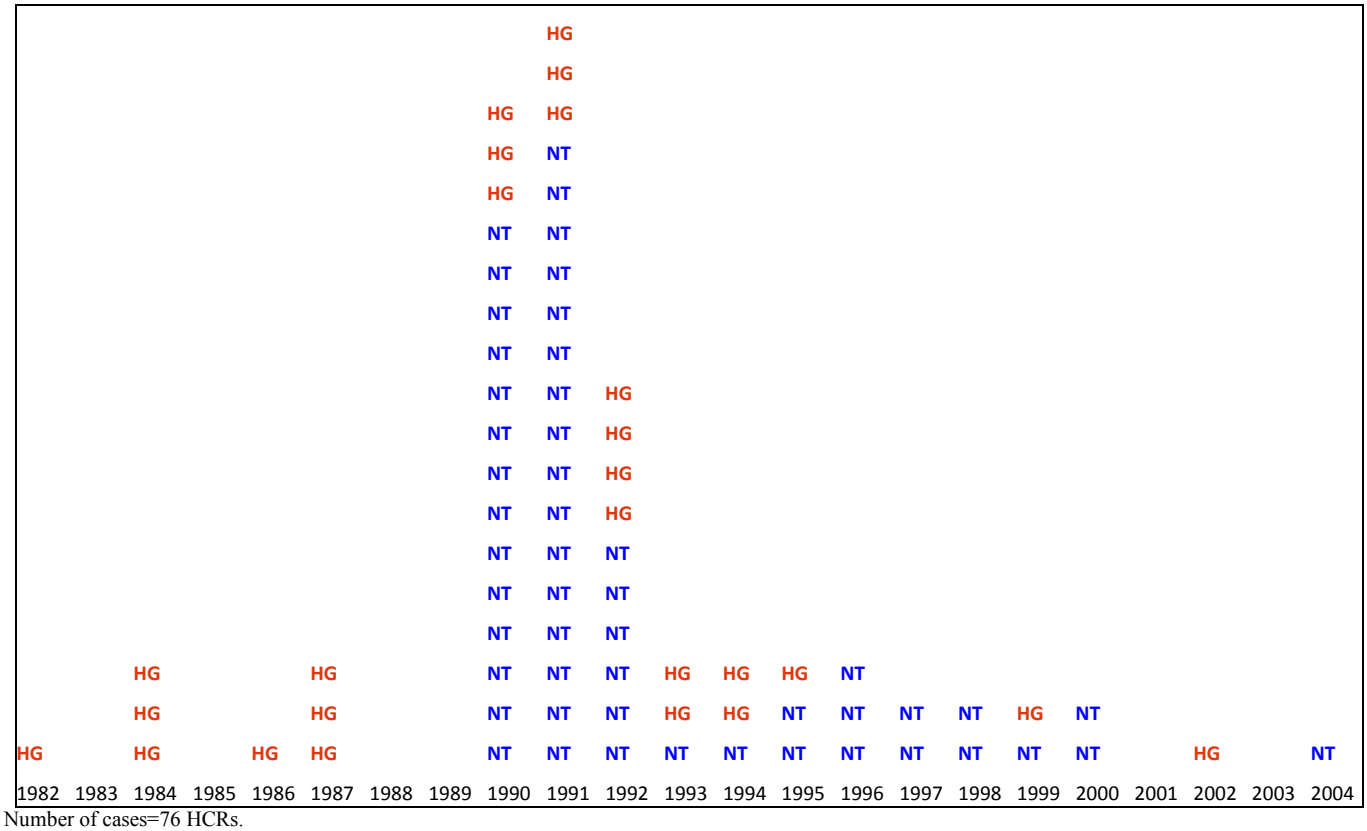
Thus we move to the second approach, which is oriented around early career patterns. It is conjectured that highly creative and comparison researchers may have similar early career research patterns in the timing, quantity, and subject area of their initial publications. Later on they may diverge because 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 such as the launch of Sputnik as well as how the timing within a scientific career is relevant for understanding creative events [28,29,1]. 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. [16] 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 [13,30]. The third category underscores the previously mentioned link between creativity and productivity [3]. 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. [13].

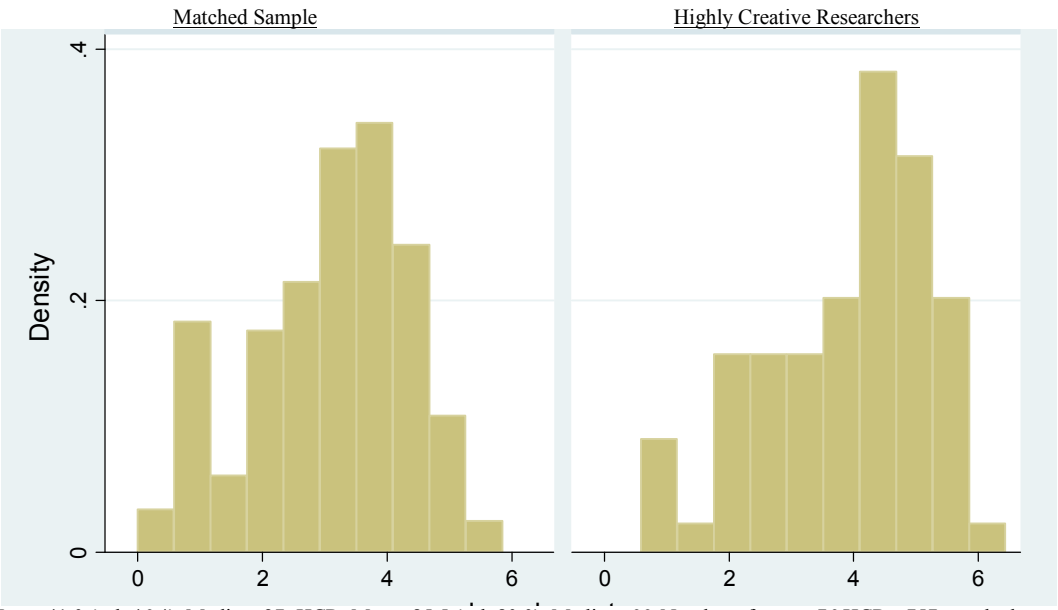
[2] A journal can be associated with more than one subject category. Multiple associations are especially common in the biosciences.



Number of cases=76 HCRs.

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

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 suggest 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.



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

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

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 good 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

* Significant at the 10% level ** Significant at the 5% level *** Significant at the 1% level.

IV. 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.

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.

This lack of distributional similarity among the creative and comparison groups is not helped by the fact 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 especially in 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 makes it is difficult to find a "group" among these creative researchers with which to compare. Indeed, Heinze et al. 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.[11] 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.

ACKNOWLEDGMENT

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.

REFERENCES

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

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

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

[4] M. Gladwell, *Outliers: The Story of Success*. New York: Little Brown, 2008.

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

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

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

[8] D. Rubin, "Estimating Causal Effects from Large Data Sets using Propensity Scores." *Annals of Internal Medicine*, vol. 127, 1997, pp. 757-763.

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

[10] G. Pion and D. Cordray. "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* vol. 31 (4), 2008, pp. 335-369.

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

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

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

[14] I. Rafols, and M. Meyer, "Diversity and Network Coherence as indicators of interdisciplinarity: case studies in bionanoscience," *Scientometrics,* in press.

[15] A.L. Porter, J. Youtie, P. Shapira, and D.J. Schoeneck, "Refining Search Terms for Nanotechnology" *Journal of Nanoparticle Research*, vol. 10 (5), 2008, pp. 715-728.

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

[17] R. S. Burt, Structural holes and good ideas, *American Journal of Sociology*, vol. 110 (2), 2004, pp.349-399.

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

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

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

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

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

[23] R. Kostoff, R., "Citation analysis of research performer quality," *Scientometrics*, vol. 53 (1), 2002, pp. 49-71

[24] F. Narin and K. Hamilton, "Bibliometric performance measures," *Scientometrics*, vol. n36 (3), 1996, pp. 293-310

[25] A. Clauset, C. Shalizi, and M. Newman, Power-law Distributions in Empirical Data. *SIAM Review*, in press.

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

[27] S. Katz, J. Rogers, and D. Hicks., Citation distributions to scientific papers: 1996-2007. Working paper, in press.

[28] P.D. Allison, *Event History Analysis: Regression for Longitudinal Event Data*, Newbury, Park, CA: Sage Publications, 1984.

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

[30] A.L. Porter, J.D. Roessner, and A.E. Heberger, "How Interdisciplinary is a Given Body of Research?" *Research Evaluation*, in press.