Do elite scientists play a key role in the genesis of transformative research of "sparking type"? An investigation in the science of science

Fangjie Xi[1], Ronald Rousseau[2,3] and Xiaojun Hu[1]*

[1]Medical Information Center, Zhejiang University School of Medicine,
Hangzhou 310058, CHINA

[2] Facultair Onderzoekscentrum ECOOM, KU Leuven,
Naamsestraat 61, Leuven, B-3000, BELGIUM

[3]Faculty of Social Sciences, University of Antwerp (UA),
Middelheimlaan 1, Antwerp, B-2000, BELGIUM

e-mail: fangjiexi@zju.edu.cn; ronald.rousseau@kuleuven.be;
*xjhu@zju.edu.cn (corresponding author);

ORCID: F.Xi: 0000-0003-0408-9754
R.Rousseau: 0000-0002-3252-2538
X.Hu: 0000-0001-8384-0221

ABSTRACT

The purpose of this study is to explore if elite scientists play a key role in the genesis of transformative research. As there exist different types of transformative research, this paper focuses on one type of work, i.e. under-cited influential work,referred to as "sparking" articles. A comparative study between the h-indices of authors citing Nobel Prize-winning papers of sparking type and those of authors citing ordinary ones is conducted, focusing on the first author and the corresponding author of each paper. The results show that the citers of the Top 1% or Top 10% Citations Sets in the sparking group have much higher h-indices than those in the ordinary group. These findings imply that elite scientists, operationalized as those with a high h-index in the corresponding fields, are more sensitive to sparking work and, as such play a pivotal role in the genesis of transformative research. This investigation provides new insight into the study of detecting transformative research, and hence, contributes to the science of science.

Keywords: Transformative research; Sparking articles; Elite scientists; H-indices; Nobel Prize-winning articles; Science of science.

INTRODUCTION

Transformative research (Trevors et al. 2012), also referred to as breakthrough research (Min, Bu and Sun 2021), is of the highest importance to the advancement of science and society. The concept of transformative research is defined in many ways but is generally considered to be associated with creativity and long-lasting impact (Trevors et al. 2012; Huang, Hsu and Lerman 2013; Prabhakaran, Lathabai and Changat 2015; Winnink, Tijssen and van Raan 2019; Du et al. 2020). The results of transformative research may revolutionize scientific inquiry, expand understanding of the world, and even have the potential to create or overturn fundamental scientific paradigms (Trevors et al. 2012; Huang, Hsu and Lerman 2013). Many scientists believe that transformative research may often lead to delayed recognition (Huang, Hsu and Lerman 2013; Min et al. 2021) because of a widespread bias against novelty and the competition for attention in the scientific community (Chai and Menon 2019). Yet, a report by the National Academies of Sciences, Engineering, and Medicine of the USA showed that transformative innovations can also arise from older and long-ignored ideas (National Academies of Sciences, Engineering, & Medicine 2016). Therefore, recognizing transformative research, especially at the early stage, is still a big challenge (Du et al. 2020).

In recent years, many scholars have explored transformative research, and this from different perspectives. Numerous papers focus on traditional bibliometric indicators and a variety of other quantitative metrics derived from them. For example, Cole (1970) and Marx (2014) stated that citations could be used to measure delayed recognition of scientific discoveries in science. However, identifying influential publications is a multidimensional process that should consider all types of indicators, moving beyond the calculation of simple citation-derived indicators (Ponomarev et al. 2014), e.g., by taking the network into account (Min, Bu and Sun 2021). In addition, Savov, Jatowt and Nielek (2020) proposed a simple, yet novel classification-based method, which may be used to complement traditional citation analysis, while Huang, Hsu and Lerman (2013) focused on the structure of the citation cascade to identify transformative research. Although the identification of transformative research can be explored from different angles, the focus of this article is on exploring if elite scientists play a key role to the genesis of transformative research.

The Nobel Prize, awarded for truly innovative research, is considered the highest honor that any scientist can achieve (at least in those fields for which a Nobel Prize exists). When a scientist is awarded with a Nobel, he or she is selected to join one of the most elite groups in the world. For this reason, Nobel Prize-winning works can be used as proxies for transformative research (Winnink, Tijssen and van Raan 2019; Min et al. 2021). Admittedly, Nobel Prize-winning publications come in many types, among which the sparking type and the igniting type (Hu and Rousseau 2016; 2017; 2019) can be distinguished.

Which factors contribute to the sparking phenomenon? In earlier studies, it was found that authoritative or elite citers could be involved (Hu et al. 2018; Hu and Rousseau 2019). Inspired by these two studies, the following hypothesis is stated: *Elite scientists are more*

prone to be involved in citing sparking work than ordinary scientists. Stated otherwise, elite scientists play a key role in the genesis of transformative research. This hypothesis is tested by operationalizing the notion of "elite scientists" as scientists with a high h-index (within a fixed field) and comparing the h-indices of authors citing Nobel Prize-winning papers with those of authors citing a comparison group of non-Nobel Prize-winning papers to explore the possibility of identification of a certain type of transformative research, namely sparking papers leading to a Nobel Prize. Therefore, the purpose of this study is to explore if elite scientists play a key role in the genesis of transformative research, by comparing the h-indices of authors citing Nobel Prize-winning papers with those of authors citing other, say ordinary papers.

Sparking Indices

In previous studies (Hu and Rousseau 2016; 2017), the concept of under-cited influential work, that is, articles behaving as sparks in the scientific landscape has been introduced. Such articles do not receive large numbers of direct citations and need subsequent publications to realize their full potential with respect to the topic they deal with.

Given an article A, its Top *a%* Citations Set consists of the top *a*% articles citing A (first generation citations of article A) in descending order of the number of citations received. In this article, the number *a* is 1 or 10. Then the terminology of *Sparking Indices* S_1 and S_{10} based on the Top 1% Citations Median ($TOPCM_3$) and the Top 10% Citations Median ($TTPCM_3$) has been introduced (for more details, see Hu and Rousseau (2016; 2017)):

$$S_1(A) = TOPCM_3(A) = \frac{2}{3}\mu_1 + \frac{1}{3}\mu_2$$

Here μ_1 denotes the median of the top 1% citations of article A, μ_2 is the median of medians for the number of citations received by the second citation generation in the top 1% set (details follow). When considering the top 10% sets, instead of the top 1%, a similar formula is used:

$$S_{10}(A) = TTPCM_3(A) = \frac{2}{3}\lambda_1 + \frac{1}{3}\lambda_2$$

where λ_1 and λ_2 are citation medians calculated from the top 10% sets instead of the top 1% sets.

MATERIALS AND METHODS

Top *a*% Citations Set

It is well-known that the distribution of the number of citations is highly skewed (Seglen 1992; Rousseau 2014). For this reason, scholars have proposed to use percentiles instead of averages, such as the top 10% of the most cited papers (Bornmann 2013) and the top 1% of the most highly cited articles (National Science Board 2012) to evaluate the quality of research. The same approach has proved effective in previous research (Hu and Rousseau 2016; 2017; Xi, Rousseau and Hu 2021).

Here, as in previous work, to ensure that the number of articles in the Top *a*% Citations Set is reasonably high, the Top 1% Citations Set (TOPCS) is used if a publication is cited more than 200 times, otherwise, its Top 10% Citations Set (TTPCS) will be used. Concretely, the following steps are followed:

Step 1: Collect all articles citing article A. This set is denoted as CIT(A).

Step 2: Rank all articles in the set CIT(A) in a descending way, according to the number of received citations.

Step 3: Choose different thresholds according to the total number of citations of article A. Depending on the case step 4a or step 4b is taken.

Step 4a: Take the top 1% of the publications (rounded up to the nearest integer, i.e., using the ceiling function) from CIT(A). This top 1% list is the TOPCS of article A. Obviously, this set is a subset of CIT(A).

Step 4b: Take the top 10% of the publications (again rounded up to the nearest integer) from CIT(A). This top 10% list is denoted as the TTPCS of article A. Also, this set is a subset of CIT(A) (for more details, see Hu and Rousseau (2016; 2017)). Recall that the ceiling function applied to a real number x returns the smallest integer greater than or equal to it. If, for example, an article has received 277 citations then 1% is 2.77. Rounding up to the closest integer leads to the number 3.

The *Flow Ratio* between Citation Generations

In informetrics, diffusion can be described as a movement through cognitive space (Liu and Rousseau 2010). In this process, citations represent the impact or influence of one paper on another and may therefore be used as an indicator of the knowledge-transfer process for a group of publications (Lewison, Rippon and Wooding 2005; Petersen et al. 2014). Here, the focus lies on how knowledge included in an article is diffused not only by itself but also by the citations received afterward. As such, the concept of *Flow Ratio* (*FR*) of a given article is proposed to measure observed diffusion between subsequent citation generations.

The target article A is its own zeroth-citation generation. Publications citing article A form its first-generation citations. Similarly, publications citing first-generation citations are second-generation citations. Taking possibly overlapping citations into account, generations are regarded as multisets, in which an element may appear several times in different citation generations (Hu, Rousseau and Chen 2011; Rousseau, Egghe and Guns 2018) . Now A's *n*th-generation citations multiset is denoted as H_n. Figure 1 illustrates article A's citation network and its generational structure (as a multiset), showing Article b cites article A and article c; Article c cites article A; Article d cites article A. Five articles, denoted as o_1, o_2, o_3, o_4, o_5, cite article b; and finally, article o_5 also cites article d.

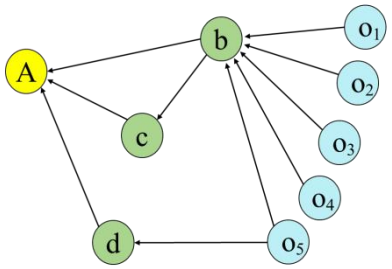


Figure 1: Citation Generations of Article A

Then A's citation-generation multisets are:
H_0= {A}
H_1= {b, c, d}
H_2= {b, o_1, o_2, o_3, o_4, o_5, o_5}

The *Flow Ratio* of article A's *n*th-citation generation ($n \geq 0$), denoted as FR_n, is defined as follows:

$$FR_n = \#(H_{n+1}) / \#(H_n)$$

For the example the following results are obtained: FR_0(A) = #(H_1) / #(H_0) =3/1 = 3 and FR_1(A) = #(H_2) / #(H_1) = 7/3 ≈ 2.33.
Note that *Flow Ratio*s are ratios and hence are independent of the absolute numbers of citations involved. For that reason, one can compare *Flow Ratio*s of highly cited scientists with those of lesser cited ones.

The h-index

To test the hypothesis proposed, the *h*-indices of scientists in the Top *a%* Citations Set for sparking articles are determined and are compared to these of scientists citing 'ordinary' articles (defined further on). The h-index, proposed by Hirsch (2005), took bibliometrics by storm and became one of the most popular indicators (Egghe and Rousseau 2021). Although there are many limitations, the h-index can be considered a mathematically simple and broadly acceptable index to identify elite scientists in the same field (Rousseau, Egghe and Guns 2018). Indeed, comparing h-indices within the same field removes one of the problems with the h-index. Moreover, the fact that the h-index favors older scientists is an advantage here as being an elite scientist implies that this person is experienced (and hence not young anymore).

Name Disambiguation

Scientific researchers often use author names in queries for retrieving scientific literature. However, due to the ambiguity of author names, the accuracy of the results can be low (Liu et al. 2014). This is a major unsolved problem for the information and computer systems and a major roadblock to scientometric research at the individual level (Li et al. 2019). The solution to this problem is to provide each scientist with a persistent and unique digital identifier such as the ORCID (Open Researcher and Contributor ID) and include this

identifier in bibliographic databases. However, to obtain an ORCID, researchers must register their ID, and many researchers have not yet done so. Fortunately, Microsoft Academic (MA) has solved this problem, at least to a large extent, by combining two main sources of knowledge. The first source is Microsoft Academic Graph (MAG). This is a heterogeneous entity graph containing scientific publication records, citation relationships between these publications, as well as authors, institutions, journals, conferences, and research fields (Sinha et al. 2015), in which the authors' profiles have been processed through a disambiguation algorithm.

The second source is the data mined from authors' websites and online curriculum vitae (CV). MA determines whether authors with identical names are the same person or not by comparing the list of papers found online with the data in MAG[1] (Figure 2). This procedure implies that the reader may be confident (but of course not absolutely certain) that when MA attributes a set of papers to an author, they were really written by that author.

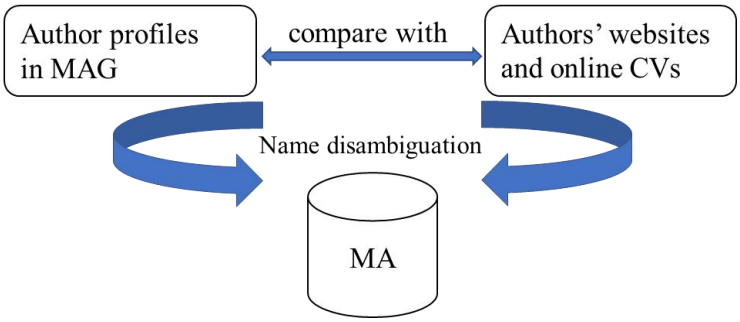


Figure 2: Two Main Sources to Address the Problem of Author Name Disambiguation

Data Collection

As indicated earlier, the purpose of this study is to explore if elite scientists play a key role in the evolution of transformative research, by comparing the h-indices of authors citing Nobel Prize-winning papers with those of authors citing other, say ordinary papers. In particular, two fields were chosen to illustrate this i.e. Physiology or Medicine, and Physics. This is because in a previous study (Xi et al. 2021), it was found that 78.57 percent and 68.75 percent of the 2020 Nobel Prize-winning publications in Physiology or Medicine and in Physics respectively were of the sparking type. Specifically, the data collection process, comprising five steps is illustrated in Figure 3.

Step 1: Searching for Nobel Prize-winning papers

The official website of the Nobel Prize provides information related to the Nobel Prize, such as the laureates' names, the reason(s) for obtaining the Prize, and the relevant award-winning articles. A total of 154 Prize-winning articles mentioned by the Nobel Prize Committee from 2016 to 2020 in the fields of Physiology or Medicine and Physics (https://www.nobelprize.org/) were collected.

[1] https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-uses-knowledge-address-problem-conflation-disambiguation

Step 2: Scanning sparking publications

The records of 154 Nobel Prize-winning papers were retrieved through the Web of Science (WoS) and their Sparking Indices were calculated. Ultimately, 95 articles (comprising 60 Physiology or Medicine and 35 Physics) were included in the sparking group, denoted as SA1, SA2, …, SA95.

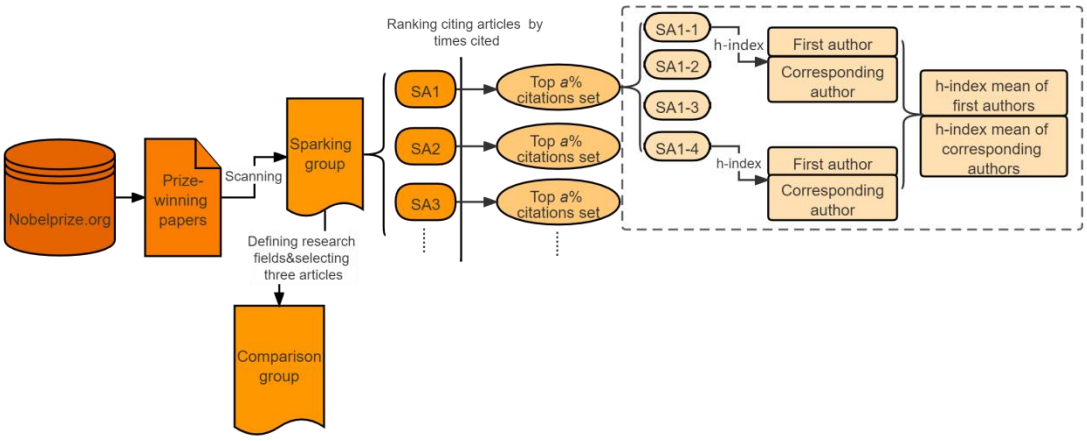


Figure 3: The Data Collection Procedure

Step 3: For each sparking article, the Top *a%* citation set (Top 1% or Top 10%) was determined.

For articles included in the sparking group, their first and second-generation citation counts were calculated separately. Finally, 644 citing articles, denoted as SA1-1, SA1-2, etc., are extracted from 36,941 first-generation citations and are listed in the Top *a*% Citations sets.

Step 4: Calculating h-indices of the first authors and the corresponding authors for citing articles included in each Top *a%* Citations set.

For 644 citing articles included in the Top *a%* Citations sets, MA was used to obtain the h-indices of the first and the corresponding authors (sometimes this is the same person). Next, the average of the h-indices of the first and of corresponding authors for each Top *a*% Citations set was calculated.

Step 5: Setting up the comparison group corresponding to each sparking fundamental work.

For articles included in the sparking group, the WoS subject category was chosen to define their fields. In the case the journal in which an article has been published belongs to more than one WoS category, the average of the relevant data for these categories was obtained. If an article was published in a journal belonging to the category of *Multidisciplinary Sciences*, it is considered to belong to the WoS category to which the most citing articles belong (Hu and Rousseau 2016; 2017). Each article included in the sparking group was compared with articles (publications of article type) in the same field and with the same publication year; these articles are ranked in decreasing order of received citations. From this ranking, three middle articles were selected to be included in the comparison group, that is, the median article and the two articles surrounding it were taken.

RESULTS

The FR_1s of the Sparking Group and the Comparison Group in Physiology or Medicine and Physics

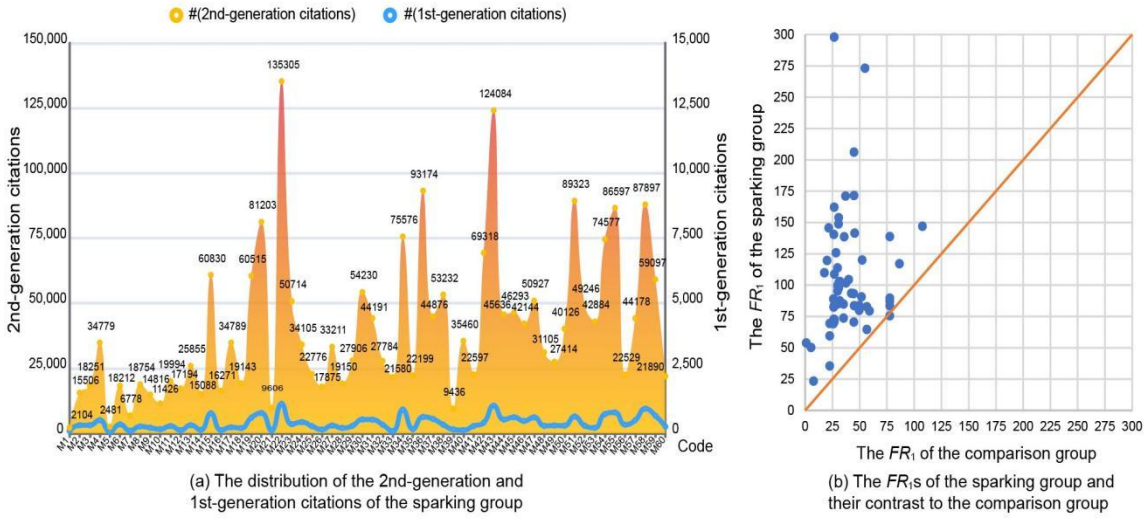
(a) Physiology or Medicine

Figure 4 shows the distribution of the first and second-generation citations of the sparking group and their FR_1s in contrast to the comparison group in the field of Physiology or Medicine. Note that, in Figure 4(a), the scale of the y-axis on the left-hand side is ten times that of the y-axis on the right-hand side.

One may say that Figure 4(a) has no special features and that the numbers of the second-generation citations are just much greater than those of the first-generation citations. However, when calculating the FR_1s of articles in the sparking group and articles in the comparison group respectively, a remarkable difference has been observed (see Figure 4(b) and Table 1).

In Figure 4(b), the horizontal axis and the vertical axis have the same scale. The origin (0,0) and the point (300,300) are connected by a straight line (the solid line) which is the isoline of the FR_1s of articles in the two groups. If a point is on the isoline, the FR_1s of articles in the two groups are equal.

Figure 4(b) shows that 59 points are above (to the left of) the isoline and only one is below, which means that the FR_1s of articles in the sparking group are generally higher than those of articles in the comparison group. Specifically, the former's FR_1s are concentrated between 50 and 150, while the latter's FR_1s are between 25 and 50, which shows that there are large differences in FR_1s between the two groups; the maximum value even reaches 271.



(a) The distribution of the 2nd-generation and 1st-generation citations of the sparking group

(b) The FR_1s of the sparking group and their contrast to the comparison group

Figure 4: Physiology or Medicine: The Distribution of the First and Second Generation Citations of the Sparking Group and their FR_1s Contrast to the Comparison Group

Further, a paired samples t-test was performed (data is roughly normally distributed). As shown in Table 1, in Physiology or Medicine, the difference in mean FR_1s between the two groups is 67.06 (95% CI, 54.27–79.85) which is statistically significant ($p < 0.01$). The articles in the sparking group have higher FR_1s than articles in the comparison group.

Table 1: Results from a Paired Samples T-Test for the FR_1s in Physiology or Medicine

	Paired Differences							
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Sparking group-Comparison group	67.06	49.53	6.39	54.27	79.85	10.49	59.00	0.000

(b) Physics

For Physics, with fewer data than for Physiology or Medicine, (35 versus 60 articles) similar characteristics occur (Figure 5). As shown in Figure 5(b), all points are above the isoline, that is, all FR_1s of the sparking group are higher than those of the comparison group. In particular, the FR_1s of the former are concentrated between 25 and 100, ranging from 12.98 to 279.41. For the latter, in contrast, they are concentrated between 0 and 50, ranging from 1.67 to 56.77.

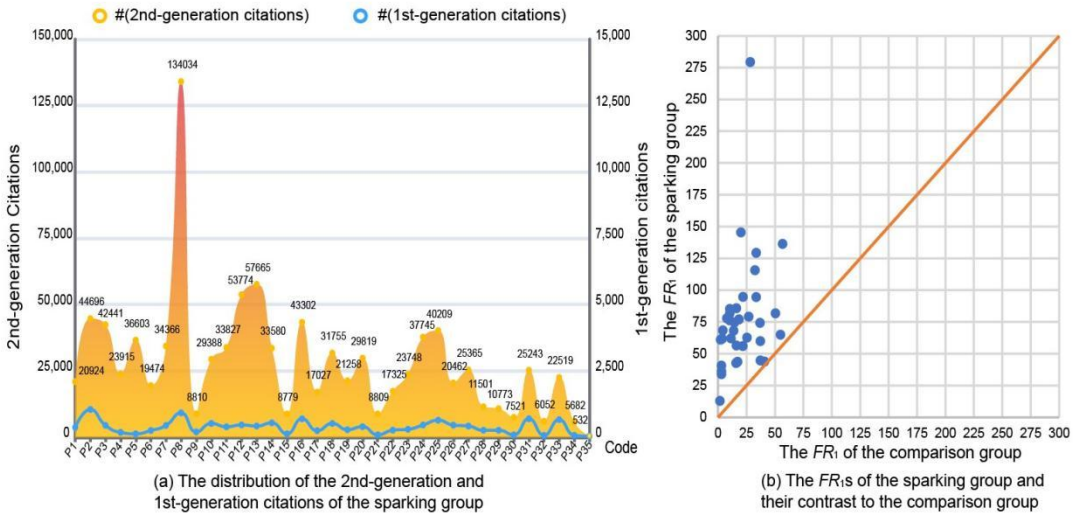


Figure 5: Physics: The Distribution of the First and Second Generation Citations of the Sparking Group and their FR_1s Contrast to the Comparison Group

The paired samples t-test (Table 2) shows that there are statistically significant differences ($p < 0.01$) in the FR_1s of the two groups and the articles in the sparking group have higher FR_1s than articles in the comparison one.

Table 2: Results from a Paired Samples T-Test for the FR_1s in Physics

	Paired Differences							
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Sparking group - Comparison group	56.21	43.00	7.27	41.44	70.98	7.73	34.00	0.000

On the whole, almost all FR_1s of sparking fundamental work among Nobel Prize-winning articles in Physiology or Medicine and Physics from 2016 to 2020 are higher than those of the other papers (published in the same field and the same year). What is the possible mechanism behind this phenomenon or what is a factor playing a main role in driving such great citation diffusion of the first generation of sparking work? Next, a closer examination of the Top *a%* Citations Sets of the two groups is performed.

The h-indices of the Corresponding Authors and First Authors of the Top *a%* Citations Sets for the Two Groups

Based on the publication data collected, the h-indices of the corresponding authors and first authors (occasionally this can be the same person, then this person is used twice, once for each group) are obtained for citing articles included in the Top *a%* Citations Sets in Physiology or Medicine and Physics, and compared between the sparking group and the comparison group (see Figures 6 and 7).

(a) Physiology or Medicine

As shown in Figure 6, the orange bars represent authors' h-indices for citing articles included in the Top *a*% Citations Sets of the sparking group and the blue bars are those of the comparison group. Obviously, for the corresponding authors, the h-indices of the Top *a*% Citations Sets of the sparking group are concentrated between 60 and 140, ranging from 24 to 188, while for the comparison group, they are mostly between 30 and 70, ranging from 8 to 78. It is found that 57 h-indices of the Set of Top *a%* Citations Sets of the sparking group are higher than those of the comparison group, which is 95 percent of the total.

Similarly, for the first authors, most h-indices of Top *a*% Citations Sets of the sparking group are between 20 and 80, and those of the Top *a*% Citations Sets of the comparison group are mainly between 10 and 40. Further, there are 53 (88.3%) h-indices of the Top *a*% Citations Sets of the sparking group higher than those of the comparison one. The difference in mean h-indices between the two groups is 30.78. These results indicate that, as expected, scientists citing the sparking papers have a higher h-index than scientists citing the comparison group.

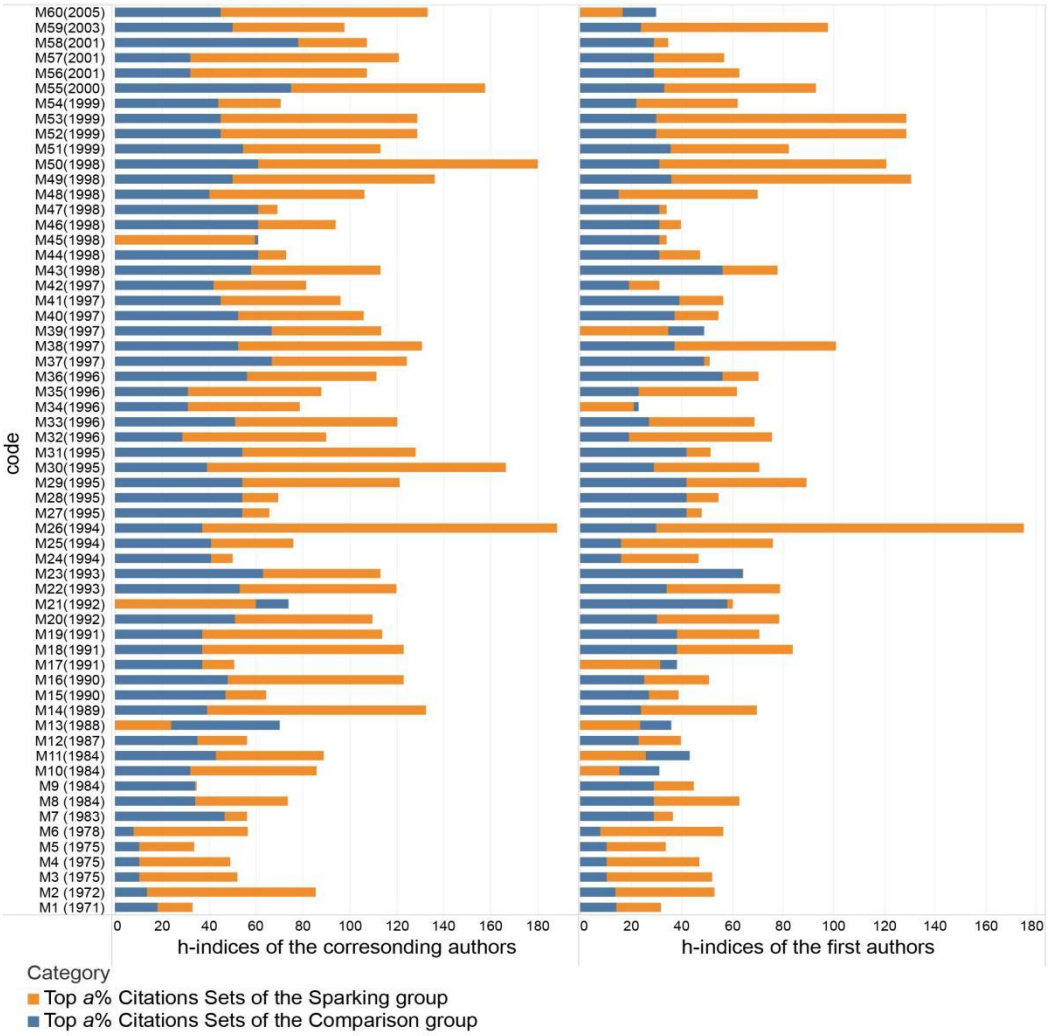


Figure 6: The H-indices of the Corresponding Authors and First Authors of the Top *a*% Citations Sets in Physiology or Medicine

To explore whether there is a significant difference in h-index between the Top *a%* Citations Sets of the two groups, again a paired samples t-test (data is roughly normally distributed) is performed. The results are shown in Table 3. For the corresponding authors, the h-index differences between the Top *a*% Citations Sets of the sparking group and those of the comparison group are statistically significant ($p < 0.01$), and the h-indices of the former are higher than those of the latter. The mean of paired differences is even up to 50.57 (95% CI, 41.54–59.60). For the group of first authors, the differences in h-index values are also statistically significant ($p < 0.01$). Those results further confirm the hypothesis that elite scientists are more sensitive to sparking work than ordinary ones.

Table 3: Results from a Paired Samples T-Test for H-indices between the Sparking Group and the Comparison Group in Physiology or Medicine

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
h-index for corresponding authors	Top *a*% Citations Sets of the sparking group-Those of the comparison group	50.57	34.94	4.51	41.54	59.60	11.21	59.00	0.000
h-index for first authors	Top *a*% Citations Sets of the sparking group-Those of the comparison group	30.78	31.98	4.13	22.52	39.04	7.46	59.00	0.000

(b) Physics

The h-indices of the corresponding authors and first authors of the Top *a*% Citations Sets in Physics are shown in Figure 7. All h-indices of the Top *a*% Citations Sets of the sparking group are higher than those of the comparison group, and this for first authors as well as for corresponding authors, without exception.

For the corresponding authors, the h-indices of the Top *a*% Citations Sets of the sparking group range from 34 to 140, and those of the comparison group are from 1 to 73. For the first authors, the h-indices of the Top *a*% Citations Sets of the sparking group are concentrated between 40 and 110, and those of the comparison group are distributed between 0 and 50.

The result of a paired samples t-test in Table 4 shows that the h-index differences between the Top *a*% Citations Sets of the two groups are statistically significant ($p<0.01$), for corresponding authors as well as for first authors, and the Top *a*% Citations Sets of the sparking group have a larger h-index, suggesting that elite scientists are more sensitive to sparking research than ordinary ones.

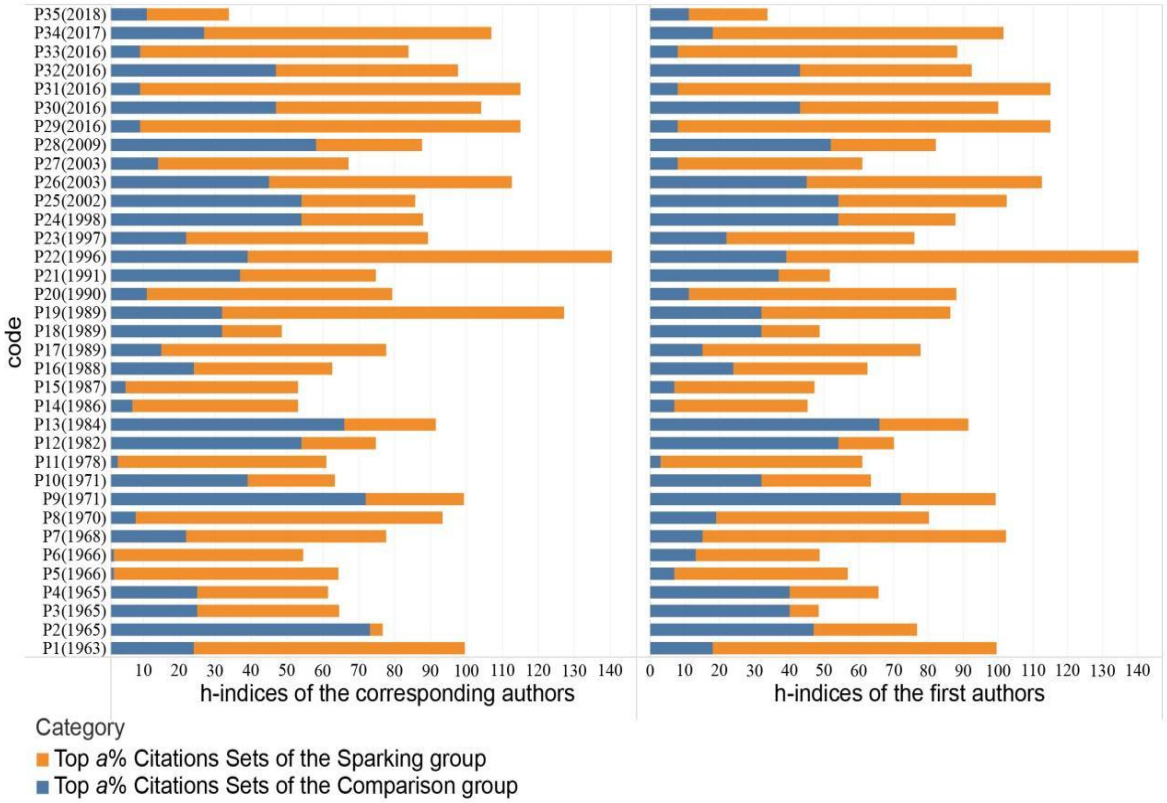


Figure 7: The H-indices of the Corresponding Authors and First Authors of the Top *a*% Citations Sets in Physics

Table 4: Results from a Paired Samples T-Test for H-indices between the Sparking Group and the Comparison Group in Physics

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
h-index for corresponding authors	Top *a*% Citations Sets of the sparking group-Those of the comparison group	53.22	26.38	4.46	44.15	62.28	11.93	34	0.000
h-index for first authors	Top *a*% Citations Sets of the sparking group-Those of the comparison group	50.71	26.95	4.56	41.45	59.97	11.13	34	0.000

DISCUSSION

The main purpose of this article was to answer the question of whether elite scientists play a key role in the genesis of “sparking type” transformative research. For this, Nobel Prize-winning articles of the sparking type in Physiology or Medicine and Physics from 2016 to 2020 are employed and compared with a comparison group. To illustrate the diffusion power that can distinguish the sparking papers from the comparison group, the concept of *Flow Ratio* (*FR*) is introduced. The results showed that almost all FR_1s of the sparking group are significantly higher than those of the comparison group, in both Physiology or Medicine and Physics (Figures 4 and 5), indicating that this type of work is very special and important to push forward scientific progress. Furthermore, the h-indices of authors included in the Top *a*% Citation Sets of the two groups were analysed and the results showed that most first and corresponding authors citing sparking articles have higher h-indices than those citing the comparison group, and this in Physiology or Medicine as well as in Physics (Figures 6 and 7), suggesting that the sparking type of work is more likely to attract the attention of elite scientists. One possible reason is that elite scientists are more sensitive to new scientific insights making them more likely to identify articles with fundamental value than other scientists (or scientists who have a keen sense of truly fundamental work, are more likely to become elite ones in their research career).

Recall that it is common for the presence of multiple authors in a biomedical publication (Hu, Rousseau and Chen 2010), a fact confirmed in the data set used here (also for Physics). As first and corresponding authors usually play major roles in a scientific contribution, they and their h-indices are the only ones used in this investigation.

The present research has explained the possible formation of the sparking phenomenon. It is widely known that scientometric research often reveals "what" through external indicators (Min et al. 2018; Ponomarev et al. 2014; Prabhakaran, Lathabai and Changat 2015), but rarely explains “why”. In previous studies, the sparking phenomenon has been observed, yet no good explanation was provided for it except for possible ‘‘transitional characteristics’’ in scientific progress (Hu and Rousseau 2016; 2017). However, in this contribution, a possible reason for the sparking phenomenon is revealed, namely, that it involves being cited by elite scientists (who are sensitive to truly fundamental work) or cited by scientists with higher h-indices under a corresponding citation window. In this way sparking work shows a remarkable diffusion power in subsequent citation generations. The results reveal a possible formation mechanism behind the unique citation characteristics of sparking work.

It should be noted that there are various types of transformative research and the sparking type studied here is just one of them. For example, Koshland (2007) proposed the Cha-Cha-Cha Theory (Charge, Challenge and Chance) to categorize Scientific Discoveries. Yet Wuestman, Hoekman and Frenken (2020) pointed out that Koshland’s discovery types are neither exhaustive nor mutually exclusive, and proposed a typology of scientific breakthroughs based on three dimensions, namely, disciplinary occurrence, considerations of use, and citation impact. However, identifying transformative research at an early stage

is still a huge scientific challenge. This is why the focus here is on one type of transformative research, namely sparking Nobel Prize-winning articles, because this type is not always visible and, in the authors' opinion, deserves more attention.

There is no doubt that identifying transformative research is a multidimensional process like that of research itself (Ponomarev et al. 2014; Hu, Luo and Rousseau 2018; Min, Bu and Sun 2021; Min et al. 2021). The findings of this article's research may provide new insight for developing such a multidimensional transformative research indicator and understanding the mechanism that leads to transformative results.

CONCLUSIONS

The findings obtained through this article's research suggest that elite scientists are more likely to have a positive response to transformative research and hence they play a key role in its genesis. Hence, elite scientists' citations (operationalized as citations from scientists with high h-indices) can serve as an early signal for identifying potential transformative research.

Admittedly, as only 95 sparking papers were used, results cannot be considered solid evidence and further research with a larger sample is necessary to confirm the obtained results. This investigation, however, provides new insight into the study of detecting transformative research, and hence it may influence the evaluation of scientific research.

An implication for research evaluation is that short-term citation measures misjudge the value of pioneering and fundamental contributions. Moreover, one should not only focus on direct citations but also take subsequent indirect citations into account.

It is unavoidable that data extraction is affected by the citation database used. Moreover, only two fields were studied in this contribution. An investigation of other fields deserves further exploration in future research. In addition, a combination of the MA database and manual judgment to disambiguate authors' names has been used because at the moment this seems the best name disambiguation method available.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China, Grant Number 71974167. The authors thank Professor Yishan Wu for his insightful comments on this contribution.

REFERENCES

Bornmann, L. 2013. How to analyze percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes, and top-cited papers. *Journal of the American Society for Information Science and Technology,* Vol.64, no.3: 587-595.

Chai, S. and Menon, A. 2019. Breakthrough recognition: Bias against novelty and competition for attention. *Research Policy,* Vol.48, no.3: 733-747.

Cole, S. 1970. Professional standing and the reception of scientific discoveries. *American Journal of Sociology*, Vol.76, no.2, 286-306.

Du, J., Li, P.X., Haunschild, R., Sun, Y.N. and Tang, X.L. 2020. Paper-patent citation linkages as early signs for predicting delayed recognized knowledge: Macro and micro evidence. *Journal of Informetrics,* Vol.14, no.2.

Egghe, L. and Rousseau, R. 2021. The h-index formalism. *Scientometrics,* Vol.126, no.7: 6137-6145.

Hirsch, J.E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America,* Vol.102, no.46: 16569-16572.

Hu, X.J., Hu, X.Y., Zhang, Y.N. and Rousseau, R. 2018. Hibernators, their awakeners and the roles of subsequent elite citers. *Malaysian Journal of Library & Information Science,* Vol.23, no.1: 103-113.

Hu, X.J., Luo, J.H. and Rousseau, R. 2018. A warning for Chinese academic evaluation systems: short-term bibliometric measures misjudge the value of pioneering contributions. *Journal of Zhejiang University-Science B*, Vol.19, no.1:1-5.

Hu, X.J. and Rousseau, R. 2016. Scientific influence is not always visible: The phenomenon of under-cited influential publications. *Journal of Informetrics,* Vol.10, no.4: 1079-1091.

Hu, X.J. and Rousseau, R. 2017. Nobel Prize winners 2016: Igniting or sparking foundational publications? *Scientometrics,* Vol.110, no.2: 1053-1063.

Hu, X.J. and Rousseau, R. 2019. Do citation chimeras exist? The case of under-cited influential articles suffering delayed recognition. *Journal of the Association for Information Science and Technology,* Vol.*70*, no.5: 499-508.

Hu, X.J., Rousseau, R. and Chen, J. 2010. In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. *Journal of Information Science,* Vol.36, no.1: 73-85.

Hu, X.J., Rousseau, R. and Chen, J. 2011. On the definition of forward and backward citation generations. *Journal of Informetrics,* Vol.5, no.1: 27-36.

Huang, Y.-H., Hsu, C.-N. and Lerman, K. 2013. Identifying transformative scientific research. Paper presented at the *IEEE 13th International Conference on Data Mining*, December 2013, at Dallas, TX.

Koshland, D.E. 2007. Philosophy of science - The cha-cha-cha theory of scientific discovery. *Science,* Vol.317, no.5839: 761-762.

Lewison, G., Rippon, I. and Wooding, S. 2005. Tracking knowledge diffusion through citations. *Research Evaluation,* Vol.14, no.1: 5-14.

Li, J.C., Yin, Y., Fortunato, S. and Wang, D.S. 2019. A dataset of publication records for Nobel

laureates. *Scientific Data,* Vol.6, no.1: 33.

Liu, W.L., Dogan, R.I., Kim, S., Comeau, D.C., Kim, W., Yeganova, L., Lu, Z.Y. and Wilbur, W.J. 2014. Author name disambiguation for PubMed. *Journal of the Association for Information Science and Technology,* Vol.65, no.4: 765-781.

Liu, Y.X. and Rousseau, R. 2010. Knowledge diffusion through publications and citations: A case study using ESI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology,* Vol.61, no.2: 340-351.

Marx, W. 2014. The Shockley-Queisser paper - A notable example of a scientific sleeping beauty. *Annalen Der Physik,* Vol.526, no.5-6: A41-A45.

Min, C., Bu, Y. and Sun, J.J. 2021. Predicting scientific breakthroughs based on knowledge structure variations. *Technological Forecasting and Social Change,* Vol.164. Available at: https://doi.org/10.1016/j.techfore.2020.120502.

Min, C., Bu, Y., Wu, D., Ding, Y. and Zhang, Y. 2021. Identifying citation patterns of scientific breakthroughs: A perspective of dynamic citation process. *Information Processing & Management,* Vol.58, no.1. Available at: https://doi.org/10.1016/j.ipm.2020.102428.

Min, C., Ding, Y., Li, J., Bu, Y., Pei, L. and Sun, J.J. 2018. Innovation or imitation: The diffusion of citations. *Journal of the Association for Information Science and Technology,* Vol.69, no.10: 1271-82.

National Academies of Sciences, Engineering, & Medicine. 2016. *Fostering transformative research in the geographical sciences*. Washington, DC: The National Academies Press.

National Science Board. 2012. *Science and engineering indicators 2012*. Arlington VA: National Science Foundation.

Petersen, A.M., Fortunato, S., Pan, R.K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H.E. and Pammolli, F. 2014. Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences of the United States of America,* Vol.111, no.43:15316-15321.

Ponomarev, I.V., Williams, D.E., Hackett, C.J., Schnell, J.D. and Haak, L.L. 2014. Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change,* Vol.81: 49-55.

Prabhakaran, T., Lathabai, H.H. and Changat, M. 2015. Detection of paradigm shifts and emerging fields using scientific network: A case study of Information Technology for Engineering. *Technological Forecasting and Social Change,* Vol.91: 124-145.

Rousseau, R. 2014. Skewness for journal citation curves. In E. Noyons (Ed.)., Context counts: Pathways to master big and little data. *Proceedings of the STI conference 2014 Leiden*, 498–501.

Rousseau, R., Egghe, L. and Guns, R. 2018. *Becoming metric-wise. A bibliometric guide for researchers.* Chandos Elsevier: Kidlington.

Savov, P., Jatowt, A. and Nielek, R. 2020. Identifying breakthrough scientific papers. *Information Processing & Management,* Vol.57, no.2: 102168.

Seglen, P. O. 1992. The skewness of science. *Journal of the Association for Information Science & Technology,* Vol.43, no.9: 628–638.

Sinha, A., Shen, Z.H., Song, Y., Ma, H., Eide, D., Hsu, B.J. and Wang, K.S. 2015. An overview of Microsoft Academic Service (MAS) and applications. *Proceedings of the 24th International Conference on World Wide Web*, May 2015 Florence, 243-246.

Trevors, J.T., Pollack, G.H., Saier, M.H. and Masson, L. 2012. Transformative research:

definitions, approaches and consequences. *Theory in Biosciences,* Vol.131, no.2: 117-123.

Winnink, J.J., Tijssen, R.J.W. and van Raan, A.F.J. 2019. Searching for new breakthroughs in science: How effective are computerised detection algorithms? *Technological Forecasting and Social Change,* Vol.146: 673-686.

Wuestman, M., Hoekman, J. and Frenken, K. 2020. A typology of scientific breakthroughs. *Quantitative Science Studies,* Vol.1, no.3: 1203-1222.

Xi, F.J., Rousseau, R. and Hu, X.J. 2021. “Sparking” and “Igniting” Key Publications of 2020 Nobel Prize Laureates. *Journal of Data and Information Science,* Vol.6, no.2: 28-40.