STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Why ESI is unreliable in selecting highly cited papers?[1]

HU Zhigang[*], TIAN Wencan[*], XU Shenmeng[**], WANG Xianwen[*], LI Jiang[***] and ZHANG Chunbo[*]

[*]*huzhigang@dlut.edu.cn, tianwen@mail.dlut.edu.cn, xianwenwang@dlut.edu.cn, zhangcb@mail.dlut.edu.cn*
WISE Lab, Dalian University of Technology, Dalian, 116024 (China)

[**]*shenmeng@email.unc.edu*
School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, 27517 (US)

[***]*li-jiang@zju.edu.cn*
School of Information Management, Nanjing University, Nanjing, 210023 (China)

Introduction

InCites Essential Science Indicators (ESI), a citation-based research analytic tool developed by Clarivate Analytics for identifying top-performing research, is becoming increasingly used in evaluating the impact of countries (Csajbók, Berhidi, Vasas, & Schubert, 2007; Fu, Chuang, Wang, & Ho, 2011), institutes (Chuang, Wang, & Ho, 2011; Ma, Ni, & Qiu, 2008) and scientists (Harzing, 2015). Highly Cited Papers (HCPs) and Hot Papers (HPs), representing the most influential research articles in one of 22 research fields annually, are two fundamental components of ESI. Specifically, HCPs represent the papers that received enough citations to be placed in the top 1% of papers in its academic field and publication year; while HPs represent the papers published in the past two years and received enough citations in last two months to be included in the top 0.1% of papers in its academic field. As reliable datasets including the most inspiring, readable and citable papers among tens of thousands of search results, ESI HCPs and HPs are also increasingly valued by Clarivate Analytics. In the new version of Web of Science Core Collection (v5.47) updated in October of 2017, the location of ESI HCP and HP filters were moved up from the bottom in the left pane on the search results webpage.

Although widely used in research evaluation and literature retrieval, ESI's validity and reliability in selecting high-impact papers have been rarely discussed in previous works. To our knowledge, the study of Harzing A. (2015), which reveals how namesake problems can compromise the ESI ranking of authors, is the only related paper.

Problems of ESI indicators

To make fair citation comparisons, ESI normalizes citation counts by the published year and the research fields. In this way, a Highly Cited Paper needs only to surpass its counterparts with the same citation time window to rank the highest (top 1%) in the research field. However, ESI indicators are not as fair as they appear to be.

[1] This work was supported by the Natural Science Foundation of China (NSFC) under Grant Nos. 71503031 and 71673038.

Firstly, it is partial to compare citation counts of papers published in different months of the same year. The papers published in January have an almost one year longer citation time window than those published in December. For example, as of today (April, 2018), papers published in January 2017 have a 15-month citation time window; while those published in December 2017 only have 4 months to accumulate their citations so far. However, papers published in all the months of 2017 will be compared together according to the definition of ESI. In this research, we will examine to what extent a paper's published month affects its probability for being selected as an ESI top paper.

Second, ESI considers a paper's issued date (or print date) as the beginning of its citation time window, which is, unfortunately, not exactly true for many papers. The online date marks the time when articles are firstly made publicly available on the publishers' websites. Examples include the "Online First" of Springer, the "Early View" of Wiley-Blackwell, and the "Advance Online Publication" of Natural Publishing Group, and so on. This means that a paper can be available and citable earlier than its issued date. A paper could benefit a lot from the time difference between its online date (the actual beginning of citation window) and its print date (the beginning of citation window considered by ESI), because being online first gives the paper extra time to accumulate citations than those issued immediately. In this paper, the lag between a paper's online date and issued date is called online-to-print delay. Online-to-print delay can lead to unfair competitions in HCP selections by providing a jump start for those papers with longer online-to-print delays. For example, if a paper was issued in 2017 but was online in 2015, it would have a longer citation time window than those that were printed and online both in 2017. This paper thus has a higher chance to be selected as a HCP. Previous works have revealed the increasing online-to-print delays in the past years and how it inflates Journal Impact Factor (Amaral & Tort, 2012; Yu, Wang, & Yu, 2005). We aim to explore whether this is also the case for ESI HCPs in this study.

Data

ESI updates its dataset of Highly Cited Papers and Hot Papers every two months. In this study, we use the dataset of HCPs as of April 2017 to examine to what extent a paper's published month and online-to-print delay will affect its HCPs eligibility. The dataset contains 135,509 HCPs published from 2007 to 2017. All these papers' bibliographic information, including their title, authors, journal, volume and page, published date, DOI, etc., were exported for our analysis.

ESI only records a paper's issued date. To access a HCP's online date, we employed the Crossref API, which caches bibliographic metadata of scholarly works and provides free services to access them. After sending a request with a HCP's Digital Object Identifiers (DOI) to Crossref server, we get a JSON-formatted response which records the HCP's online dates and other related information. The HCPs' online dates were then extracted from the JSON files to compute the online-to-print delay.

Results

ESI Highly Cited Papers are determined field by field and year by year. A paper is selected as HCP only if its citation count exceeds the threshold of the corresponding research field and published year. Table 1 lists the distribution of the 135,509 HCPs in terms of the research fields and published years. Among all the 22 research fields, "Clinical Medicine", "Chemistry", "Engineering", and "Physics" are the four most prolific research fields, which account for nearly half of the total HCPs in total. "Multidisciplinary" is the smallest research field with only 183 HCPs.

Table 1 The distribution of ESI Highly Cited Papers as of April, 2017

Fields*	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017†	ALL
CLIN	1,969	2,094	2,190	2,278	2,367	2,565	2,666	2,745	2,866	2,774	649	25,163
CHEM	1,293	1,332	1,384	1,426	1,554	1,548	1,649	1,711	1,787	1,818	450	15,952
ENGI	782	838	931	946	1,028	1,096	1,191	1,245	1,249	1,405		10,711
PHYS	981	1,003	1,007	996	1,030	1,055	1,087	1,111	1,069	1,137	196	10,672
SOCI	541	648	710	763	808	856	862	898	891	921	247	8,145
MATE	503	554	586	585	647	680	759	839	914	932	223	7,222
PLAN	566	598	614	625	672	698	701	724	780	784	169	6,931
BIOL	570	584	589	627	651	693	715	730	755	781	199	6,894
NEUR	384	406	425	436	460	487	511	509	528	514	223	4,883
MOLE	333	349	356	375	394	433	444	467	501	504	135	4,291
ENVI	276	304	326	336	373	400	448	464	506	612	150	4,195
GEOS	314	329	339	354	371	404	441	459	465	478	154	4,108
MATH	303	340	347	365	368	415	413	461	386	393		3,791
PSYC	269	294	313	325	344	371	401	416	428	464	139	3,764
AGRI	288	316	338	358	373	381	401	422	437	396		3,710
PHAR	274	306	307	330	347	369	373	374	397	405	54	3,536
COMP	225	244	258	269	294	324	350	377	402	392	87	3,222
ECON	178	215	225	231	239	254	270	259	276	238	70	2,455
IMMU	189	202	209	218	228	244	257	264	259	246	90	2,406
MICR	147	153	158	167	192	212	202	202	211	219	32	1,895
SPAC	118	122	128	127	134	140	141	143	147	140	41	1,381
MULT	11	11	12	12	16	19	24	25	25	27		182
ALL	10,514	11,242	11,752	12,149	12,890	13,644	14,306	14,845	15,279	15,580	3,308	135,509

*__Abbr. of Research Fields__: AGRI=Agricultural Sciences; BIOL=Biology & Biochemistry; CHEM=Chemistry; CLIN=Clinical Medicine; COMP=Computer Science; ECON=Economics & Business; ENGI=Engineering; ENVI=Environment/Ecology; GEOS=Geosciences; IMMU=Immunology; MATE=Materials Science; MATH=Mathematics; MICR=Microbiology; MOLE=Molecular Biology & Genetics; MULT=Multidisciplinary; NEUR=Neuroscience & Behavior; PHAR=Pharmacology & Toxicology; PHYS=Physics; PLAN=Plant & Animal Science; PSYC=Psychiatry/Psychology; SOCI=Social Sciences, General; SPAC=Space Science.

† **It is not the full year, only up to April, 2017.**

HCPs are most likely to be published in the early months of the year

For each Highly Cited Papers, the published month is recorded in the original ESI dataset. Figure 1 shows the distribution of the HCPs' published months of each year between 2007 and April of 2017. A much larger number of HCPs are published in January than in December, especially in recent years. For instance, among all the 14,554 HCPs of 2016, 3,870 HCPs (26.6% of the total) were published in January; only 93 of them (0.6%) were published in December. This means that papers published in December are almost 40 times harder to be selected as a HCP than those published in January. In total, only 13.5% of the 2016 HCPs were published in the second half of the year. Ever though HCPs of the early years are less unevenly distributed in every month, there is still a slim advantage in publishing in the beginning of the year. For instance, in the year of 2007, 54.1% of the total HCPs were published in the first half of the year, which is nearly 1/5 more than those published in the second half.

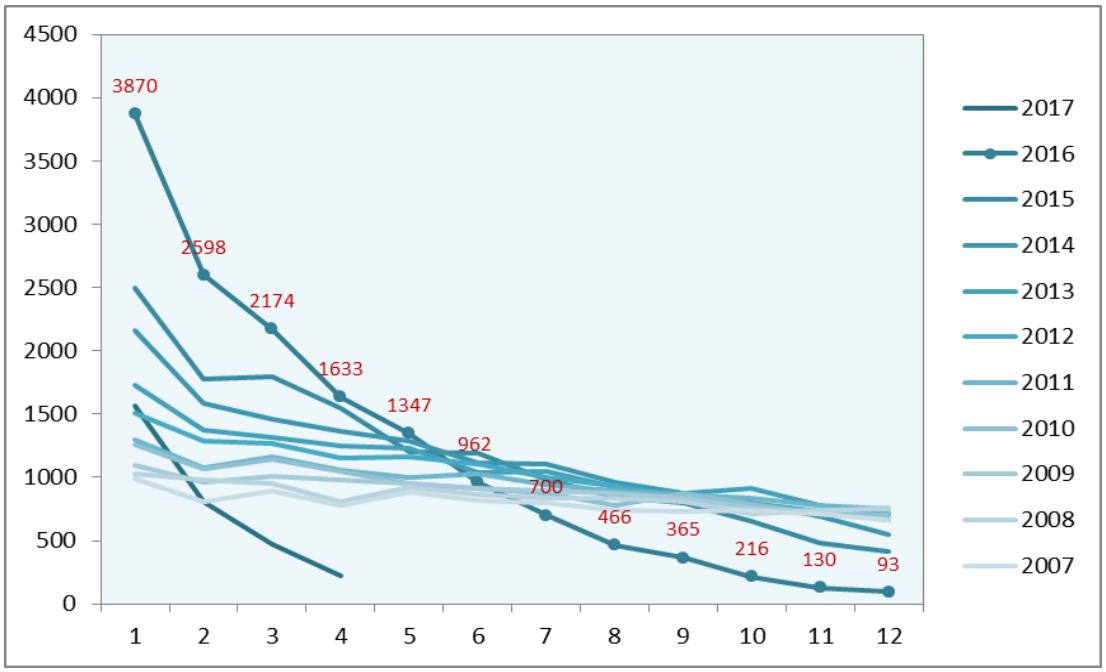

Figure 1 The distribution of the HCPs' published months of each year.

The effect of published months is easy to understand: papers published in earlier months of a year have longer citation time windows; longer citation time windows contribute to the accumulation of higher citations. As a results, determining the most highly cited papers by year does not make the HCP selection as fair as it seems. Unless ESI modifies its time interval of comparison from a year to a month, we see no fair comparison between papers and no rationality in such a rough method of HCP selection.

Online-to-print delays are commonplace in HCPs

What makes the HCP selection mechanism more questionable is the ubiquity of online-to-print delays. The histogram in Figure 2 shows the distribution of the HCPs' online-to-print delays in months. Among all the 135,509 HCPs, only 34,946 articles (25.8% of the total) were published in print in the same months with their online dates. 30,022 (22.2%) HCPs were published in print in the months after their online dates. Other 47.3% were published in print with delays of two months or more. The pie chart in Figure 2 further shows the distribution of the HCPs' online-to-print delays in years. Although 74.1% of the total were published in print in the same years with the online dates, there are still approximately 1/4 (24.2%) of the total that were issued to the years following their online dates.

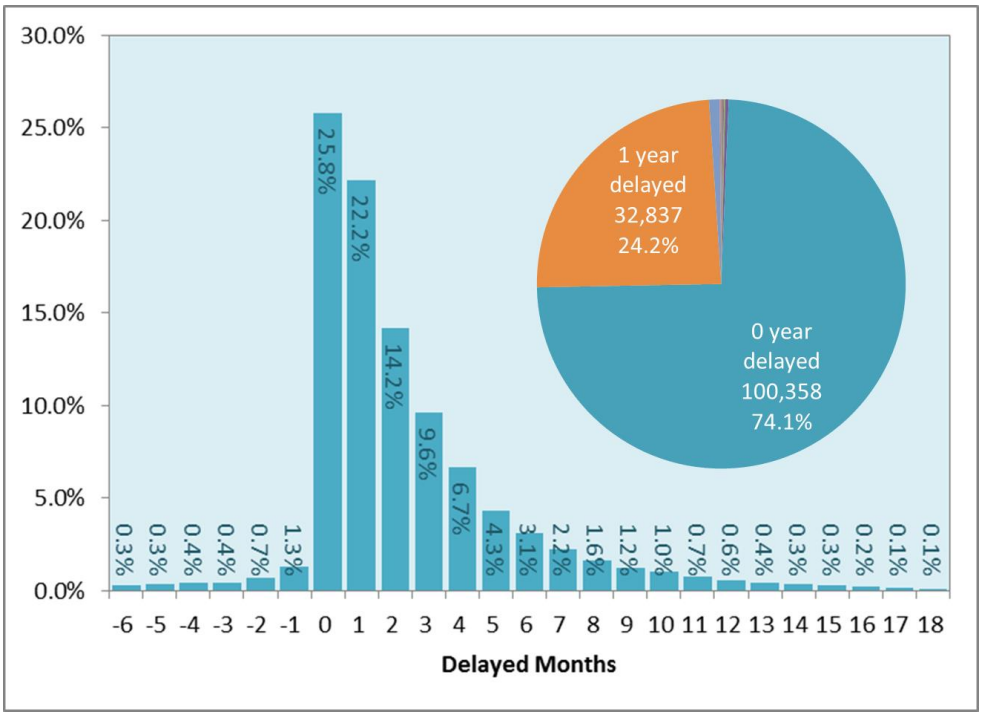

Figure 2 The distribution of HCPs' online-to-print delays

Biases caused by the online-to-print delay

The unconformity of online-to-print delays between journals creates problems in the comparison and ranking of articles, thus resulting in problematic HCP lists. Specifically, papers in a journals with a longer online-to-print delay will have an advantage when competing with those in journals with a relatively shorter delay.

We chose Journal of Management (JOM) as a case to reveal how online-to-print delays affect the HCP selection. JOM is one of the top journals published by Sage Publication Inc. Its recent Journal Impact Factor (JIF) is 7.733, which ranked the 2nd in the research area of "Business", the 1st in "Psychology, Applied", and the 3rd in "Management". There are in total 103 papers in JOM that have been selected as HCPs, ranking the 2nd in the research field of "Economics & Business".

Online-to-print delays are found to be considerably prominent in JOM. Among its 103 HCPs, only two papers are published in print in the same months with their online dates. There are 71 HCPs that were published in the following years of their online dates due to online-to-print delays. In average, there is a delay of 10.6 months.

Figure 3 shows how JOM benefits from this delay. Each red node indicates a HCP's issued date, while its corresponding green node indicates its online date. In the chart, HCP's yearly thresholds of the field of "Economics & Business", which JOM belongs to, were also provided to indicate if a paper is qualified to be a HCP. If a green node is located below the thresholds, it means that this HCP should not have been selected by ESI. As is shown, 47 green nodes are below the thresholds. These HCPs would have been disqualified if judged based on the online dates. In this paper, we call these HCPs "Unqualified HCPs". With approximately half (45.6%) of HCPs as Unqualified HCPs, online-to-print delays in Journal of Management inflate its number of HCPs.

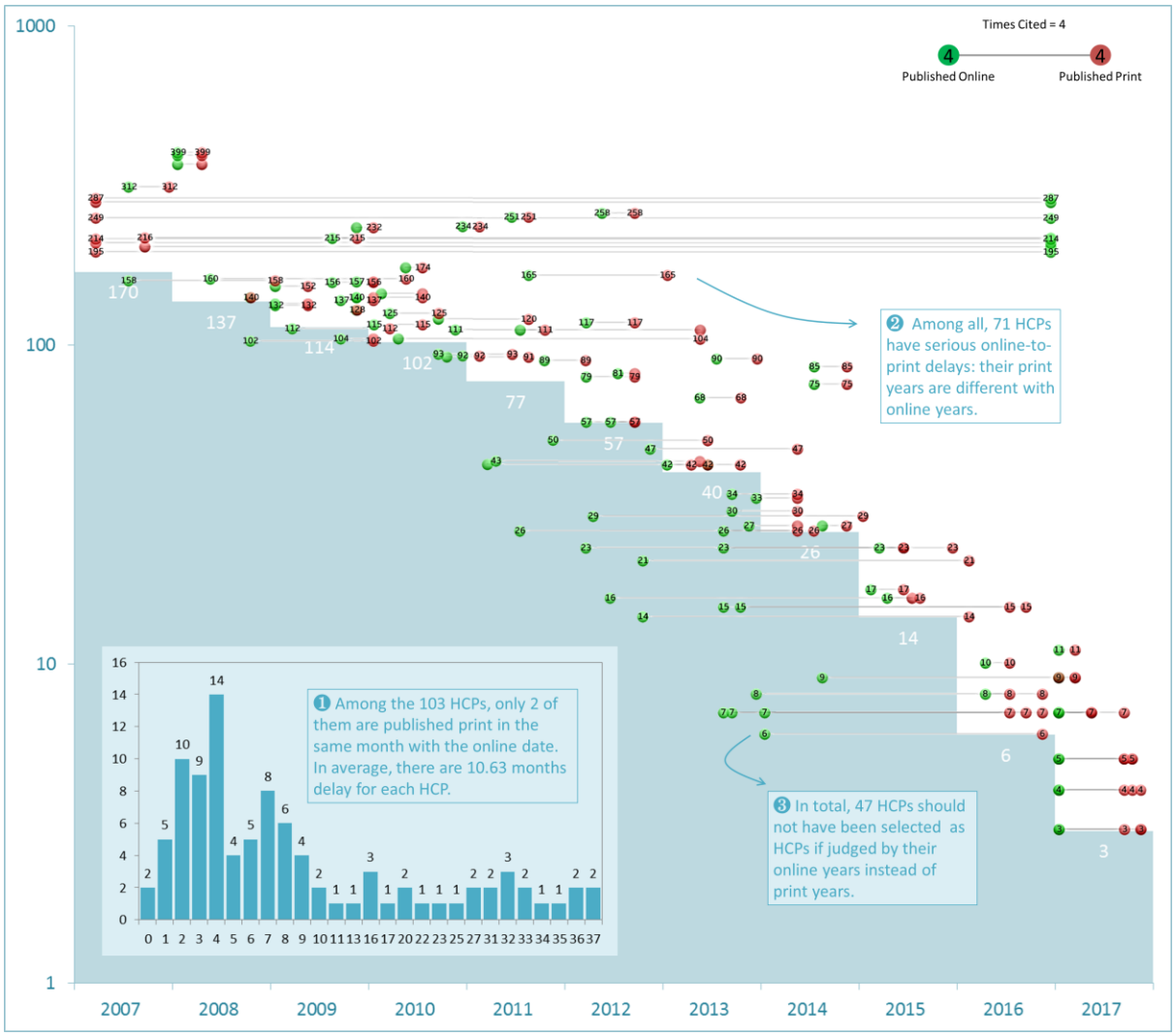

Figure 3 HCP inflation in Journal of Management caused by online-to-print delays

Journal of Management is not the only journal containing considerable Unqualified HCPs. Among all the 135,509 HCPs, 20,848 are Unqualified HCPs, accounting for 15.4% (approximately 1/7) of the total. Table 2 shows the share of Unqualified HCPs in each research field. Compared with the other research fields, "Economics & Business", "Agricultural Sciences" and "Engineering" contain larger percentages of Unqualified HCPs. Approximate 1/4 HCPs in these field are Unqualified HCPs. In other research fields such as "Physics", "Space Science" and "Molecular Biology & Genetics", the numbers of Unqualified HCPs are much lower. This means that the HCP selections in these research fields are more reliable and veritable.

In each research field, we have identified two typical journals with long online-to-print delays and considerable amount of Unqualified HCPs. The results are listed in Table 2. For example, in the research field of "Agricultural Sciences", Food Chemistry, a top-level Open Access journal published by Elsevier, contains 485 HCPs. However, 257 papers (53% of the total) are Unqualified HCPs which should not have been on HCP list if using online date.

Table 2 The share of Unqualified HCPs in each research field

RFs	UHCPs	HCPs	Shares	Typical Journals (UHCPs/HCPs)
ECON	630	2,455	25.66%	J Manage(47/103); J Bus Res(25/37)

AGRI	930	3,710	25.07%	Food Chem(257/485); Food Hydrocolloid(59/115)
ENGI	2,639	10,711	24.64%	Chem Eng J(194/647); Appl Energ(169/535)
COMP	684	3,222	21.23%	Inform Sciences(59/166); Neurocomputing(31/60)
MATH	803	3,791	21.18%	J Math Anal Appl(49/125); J Comput Appl Math(38/83)
PSYC	791	3,764	21.01%	Annu Rev Psychol(54/122); Schizophrenia Bull(50/100)
ENVI	880	4,195	20.98%	Sci Total Envir(82/176); Land Degrad Dev(42/44)
SOCI	1,656	8,145	20.33%	Tob Control(36/81); Tourism Manage(35/53)
PHAR	652	3,536	18.44%	Annu Rev Pharmacol Toxicol(34/85); Pharmacol Ther(31/106)
PLAN	1,225	6,931	17.67%	Plant Cell Environ(47/121); Fish Fish(35/51)
NEUR	862	4,883	17.65%	Biol Psychiat(59/132); Mol Psychiatr(54/98)
GEOS	721	4,108	17.55%	Gondwana Res(49/102); Bull Amer Meteorol Soc(33/92)
BIOL	1,003	6,894	14.55%	Bioresource Technol(58/156); Biol Rev(44/62)
CLIN	3,434	25,163	13.65%	Gut(116/260); Ann Rheum Dis(94/228)
MATE	966	7,222	13.38%	Nano Energy(52/127); J Alloys Compounds(45/57);
IMMU	273	2,406	11.35%	J Allerg Clin Immunol(23/83); Clin Infect Dis(21/116)
CHEM	1,618	15,952	10.14%	Appl Catal B-Environ(120/225); Biosens Bioelectron(57/99)
MICR	192	1,895	10.13%	Isme J(35/116); J Virol(19/84); Parasitol Res(8/24)
MOLE	349	4,291	8.13%	J Tissue Eng Regen Med(22/22); Oncogene(22/42)
SPAC	98	1,381	7.10%	Mon Notic Roy Astron Soc(40/259); Space Sci Rev(10/32)
PHYS	436	10,672	4.09%	Commun Nonlinear Sci Numer Si(13/25); Advanced Sci(12/23)
MULT	5	182	2.75%	Science(3/67); Proc Nat Acad Sci Usa(1/22); Nature(1/61)

Conclusions

According to the definition of ESI Highly Cited Papers, ESI has employed time and field normalizations to make the selection mechanism of HCPs impartial. However, our study has proved that ESI indicators are not as reliable as they seem to be. Both the published month and the online-to-print delay will affect a paper's probability to become a HCP.

Unfortunately, ESI-related indicators risk being abused or misused in research management. In China, for example, the number of HCPs has already been considered as a key reference indictor in university evaluation and resource allocation. Journals are also given the motivation to game ESI to produce more HCPs if they find it effective by simply artificially increasing the online-to-print delays. In this way, a "bad" journal which has long delays might drive out "good" journals that have no or shorter delays.

References

Amaral, O. B., & Tort, A. B. L. (2012). Rising Publication Delays Inflate Journal Impact Factors. *PLOS ONE*, *7*(12), e53374. http://doi.org/10.1371/journal.pone.0053374

Chuang, K., Wang, M., & Ho, Y. (2011). High-impact papers presented in the subject category database of the institute for scientific information. *Scientometrics*, *87*(3), 551–562. http://doi.org/10.1007/s11192-011-0365-2

Csajbók, E., Berhidi, A., Vasas, L., & Schubert, A. (2007). Hirsch-index for countries based on Essential Science Indicators data. *Scientometrics*, *73*(1), 91–117. http://doi.org/10.1007/s11192-007-1859-9

Fu, H., Chuang, K., Wang, M., & Ho, Y. (2011). Characteristics of research in China assessed with Essential Science Indicators. *Scientometrics*, *88*(5), 841–862. http://doi.org/10.1007/s11192-011-0416-8

Harzing, A.-W. (2015). Health warning: might contain multiple personalities--the problem of homonyms in Thomson Reuters Essential Science Indicators. *Scientometrics*, *105*(3), 2259–2270. http://doi.org/10.1007/s11192-015-1699-y

Ma, R., Ni, C., & Qiu, J. (2008). Scientific research competitiveness of world universities in computer science. *Scientometrics*, *76*(2), 245–260. http://doi.org/10.1007/s11192-007-1913-7

Yu, G., Wang, X.-H., & Yu, D.-R. (2005). The influence of publication delays on impact factors. *Scientometrics*, *64*(2), 235–246. http://doi.org/10.1007/s11192-005-0249-4