March 1, 2013

Citation time window choice for research impact evaluation

Jian Wang, *iFQ*

Citation time window choice for research impact evaluation

Jian Wang. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, 94(3), 851-872. doi: 10.1007/s11192-012-0775-9

Jian Wang

Institute for Research Information and Quality Assurance (iFQ), Schuetzenstrasse 6a, 10117 Berlin, Germany

jianwang@gatech.edu; wang@forschungsinfo.de

Abstract: This paper aims to inform choice of citation time window for research evaluation, by answering three questions: (1) How accurate is it to use citation counts in short time windows to approximate total citations? (2) How does citation ageing vary by research fields, document types, publication months, and total citations? (3) Can field normalization improve the accuracy of using short citation time windows? We investigate the 31-year life time non-self-citation processes of all Thomson Reuters Web of Science journal papers published in 1980. The correlation between non-self-citation counts in each time window and total non-self-citations in all 31 years is calculated, and it is lower for more highly cited papers than less highly cited ones. There are significant differences in citation ageing between different research fields, document types, total citation counts, and publication months. However, the within group differences are more striking; many papers in the slowest ageing field may still age faster than many papers in the fastest ageing field. Furthermore, field normalization cannot improve the accuracy of using short citation time windows. Implications and recommendations for choosing adequate citation time windows are discussed.

Keywords: *Citation time window; Citation ageing; Research evaluation; Field normalization*

Introduction

Citation counts have been widely used to indicate research impact, or even research quality. Although the validity of such indicators is still in dispute (De Bellis 2009), citation counts have been increasingly used in real-world research evaluations and funding allocations (Abbott 2009; King 2004). One important decision confronting such practice is the choice of a

time window, that is, citations within how many years after publication should be counted to measure research impact. In research evaluation there is an enduring tension between the needs of funders for timely assessment of funded research and the long time period it takes for research to reveal its full impact. On the one hand, a short time window of one or two years would allow timely monitoring and evaluation (Adams 2005). On the other hand, a short time window is criticized as biased for two primary reasons: First, at the field level, it takes much longer to be recognized and cited in fields such as the social sciences or mathematics than in biomedical research fields. Therefore, a short time window results in an unfair evaluation across different research fields (Glänzel and Schoepflin 1995). Second, the pattern of "obsolescence" (Line 1993), "ageing" (Glänzel and Schoepflin 1995), or "durability" (Costas et al. 2010) varies at the article level. Garfield (1985a, 1985b) found that citation counts for some papers rose to a peak and then steadily declined, while for other papers citation counts continued rising. Aversa (1985) also found two patterns: "delayed rise - slow decline" and "early rise - rapid decline." Therefore, a short time window would discriminate against "delayed rise - slow decline" papers, which often turn out to be more valuable and influential and are also known as "scientific prematurity" (Stent 1972), "delayed recognition" (Garfield 1980), and "sleeping beauties" (Van Raan 2004). "Sleeping beauties" are very rare and therefore may not cause serious problems in research evaluation using short citation time windows (Glänzel et al. 2003; Van Raan 2004). However, even excluding these extreme cases, there are still significant differences in ageing patterns between papers, which may affect evaluation results (Costas et al. 2011; Abramo et al. 2012b). Therefore, it is important to assess systematically how accurately citation counts in short time windows approximate total citation counts.

Several studies have already assessed the accuracy of using short time window citation counts. Adams (2005) analyzed the United Kingdom's papers in six subject categories across the life and physical sciences published in 1993 (8,258 papers in total) and found significant correlations between citation counts in initial (year 1-2) and later years (year 3-10) in all six categories, with the minimum correlation of 0.653 observed in the field of optics and acoustics. Levitt and Thelwall (2008) studied the most highly cited articles in six subjects published in 1970 (54 papers from the Science Citation Index and 33 papers from the Social Sciences Citation Index) and found that four fields out of six have a Spearman correlation over 0.42 between the total citation ranking and the percentage of early citations in the first six years after publication.

Rogers (2010) studied the citation history from 1991 to 2008 of 168,603 papers published from 1991 to 2000 in the field of nanotechnology in Thomson Reuters Web of Science (WoS) and found that it took many years for the top cited papers to establish themselves as top papers and many papers showed a continually increasing citation pattern. These studies came to different conclusions about the accuracy of using short citation time windows because of different data samples or assessment criteria. Therefore, this paper aims to provide a more systematic and comprehensive assessment, by analyzing all WoS journal publications in 1980 and calculating the correlations between total citations in all 31 years and cumulative citation counts in each possible time window, namely from 1 to 30.

Besides the accuracy of using short citation time windows, it is also important to understand factors that affect citation ageing. The first intensively studied factor is research field. Glänzel and Schoepflin (1995) showed that citation ageing in the social sciences and mathematics journals is slower than in medical and chemistry journals. Aksnes (2003a) found 33% of the papers in the physical, chemical and earth sciences were of the "early rise - rapid decline" type, but none in biology or environmental sciences. Abramo et al. (2011) also found significant differences in citation ageing between clusters of disciplines. The second factor is document type. Costas et al. (2010) noticed that "delayed rise" documents were more represented in articles, while "early rise - rapid decline" type were published more often as notes, letters, and editorials. The third factor is the quality of the paper (as indicated by total citation counts). Many studies revealed that highly cited papers had a slower ageing process (Aversa 1985; Levitt and Thelwall 2008; Walters 2011). The relationship between citation ageing and quality of the paper is important not because we should normalize citation counts by the quality but because using short citation time window may disadvantage high quality papers as discussed in the first paragraph. Another factor that has not been investigated is the month of publication. Citation time window is typically on a yearly base, therefore, a paper published in December may be unfairly compared with a paper published in January. This paper aims to uncover the differences in citation ageing in dependence of research field, document type, total citation count, and publication month.

Research field differences in citation behavior have drawn a lot of attention from not only citation ageing studies but also more general research on citation-based indicators. Many field normalization approaches have been developed to make citations more compatible across

research fields (Leydesdorff and Opthof 2010; Radicchi et al. 2008; Schubert and Braun 1996). Therefore, the final research question of this paper is whether field normalization can improve the accuracy of using short citation time windows. Two major sources of field variation in citation behavior are the ageing differences as discussed before and the size differences, that is, some research fields have less citing papers or shorter reference lists to give out citations (Moed et al. 1985). However, field normalization methods in literature pay attention exclusively to the size but not the ageing differences. This would not be a problem if the citation ageing is homogeneous within the same research field, that is, papers in the same field have similar ageing patterns. However, this might not be the case. Leydesdorff (2008) warned that the assumption that citation pattern is homogeneous within field is invalid. In an analysis at the journal level, Moed et al. (1998) found that citation ageing characteristics were primarily specific to the individual journal rather than to the subfield. Levitt and Thelwall (2008) noticed significant ageing differences between articles within the same field. Radicchi and Castellano (2011) also found that citation patterns were different between subfields within the same research field. Therefore, this paper aims to investigate the ageing differences within the same field and whether field normalizations could improve the accuracy of using short citation time windows.

In sum, this paper addresses the following three research questions:

1. How accurate is it to use citation counts in short time windows to approximate long term citation counts (i.e. 31 years)?
2. How does citation ageing differ by research fields, document types, total citation counts, and publication months?
3. How does citation ageing differ within the same research field, and can field normalization improve the accuracy of using short citation time windows?

Data

Data are from a bibliometrics database developed and maintained by the Competence Center for Bibliometrics for the German Science System (KB) and derived under license from the Thomson Reuters Web of Science (WoS).

Dataset 1: To evaluate the general accuracy of short citation time windows, all journal papers published in the year 1980 in WoS are used for analyses, that is, 746,460 papers in total. Non-self-citations received by each paper are counted for each year from 1980 to 2010.

Although it is still in debate, many scholars suggest that self-citations (i.e. citations by authors themselves) hardly reflect research impact in the scientific community, and therefore non-self-citations should be used to measure research impact (Porter 1977; Glänzel et al. 2006; Aksnes 2003b). Throughout this paper, non-self-citations are analyzed.

Dataset 2: For evaluating research field, document type, total citation, and publication month differences in citation ageing, several restrictions are imposed on the data. First, the question in focus is the ageing of citation, so we exclude papers that are never cited by others in all 31 years. Second, we keep journals with at least four issues in 1980 to allow a reliable comparison of publication months. Third, we keep only the six most frequent document types for comparison: article, note, meeting abstract, letter, review, and editorial material. 358,100 papers are available for analyses. Dataset 2 is a subset of dataset 1.

It is more appropriate to use dataset 1 to give a general picture about how accurate it is to use short citation time windows, while restrictions imposed in dataset 2 are needed to investigate citation ageing and make reliable comparisons. Therefore, we firstly use dataset 1 to give a general picture, and then switch to dataset 2 for detailed analyses, and finally switch back to dataset 1 to inform real-world research evaluations and future studies. Which dataset is used is noted in figure captions and table titles.

Month Coding: Publication month information is available in WoS for recent publications but not papers published in 1980, so we have to infer the publication month from volume and issue number. For journals using month as issue number, the issue months are transferred into numeric value 1 to 12 for January to December correspondingly. For journals numbering volumes and issues continuously, we firstly sort the volume and issue number from the earliest to the latest to get the rank $\boldsymbol{R_i}$ for each issue, and then estimate the month as $\boldsymbol{12*R_i/R_{MAX}}$, where $\boldsymbol{R_{MAX}}$ is the largest number of ranking and also the total number of issues in the year. For journals with missing issue numbers, $\boldsymbol{R_{MAX}}$ value is obtained after adding in these missing issues. In addition, month of the combined issue takes the middle value, that is, publication month of issue SEP-OCT (or 9-10) is coded as 9.5. For the remaining irregular cases (i.e. journals using letters or a combination of letters and numbers as issue numbers), publication month information is decided on a case-by-case basis.

Field Classification: The United States National Science Foundation (NSF) journal field classification scheme developed by the Patent Board is used for classifying journals into research

fields. It is a two-level system classifying journals into one unique research field and subfield. However, we keep journals with WoS subject category 'multidisciplinary sciences' as 'multidisciplinary sciences.' Furthermore, the NSF scheme does not cover the arts and leaves some social sciences and humanities journals as 'unassigned,' so we manually code the remaining journals (which are not classified by NSF scheme or classified as 'unassigned'). Most of them are about literature and arts.

Total Citation Tier: Papers are categorized into four tiers by their total citations in all 31 years: Tier 1 to 4 correspond to the 1st to the 4th quarter of top cited papers correspondingly.

Data are stored in the Oracle SQL developer, we write SQL queries to extract citation history and other relevant information for each publication, and then the data are delivered to R for statistical analysis. R is a free and open-source software environment for statistical computing and graphics and is available at: http://www.r-project.org/.

Results

Time window accuracy

The correlation between the cumulative non-self-citation counts in each time window (from 1 to 30 years) and total non-self-citations (in all 31 years) is calculated in three approaches: Pearson correlation of citation counts on the original scale, Pearson correlation of natural logarithm transformed citation counts (i.e. ***ln(citation count +1)***), and Spearman rank correlation. Results from all three approaches are reported here to allow comparison with previous findings in literature using different approaches. Given that citation counts distribution is far away from the normal distribution, the nonparametric Spearman correlation gives most reliable results. The Pearson correlation of the natural logarithm transferred citation counts gives similar results to the Spearman correlation (Fig. 1 Plot a).

The Spearman correlation between total citations and cumulative citation counts in the first year, three years, five years, and ten years are 0.266, 0.754, 0.871, and 0.948 respectively. The Spearman correlation increases rapidly in the first several years and then slowly until eventually reaching one. However, the correlation may be overoptimistic because about half of the papers are never cited in the whole history of 31 years and therefore stay in the low rank in

all years. Therefore, we expect the correlation for highly cited papers to be lower than for the whole population. To verify this proposition, we remove reiteratively half of the papers that are less cited and then calculate the correlation for the remaining papers. In the first step, about half of the papers are never cited, so we remove them and calculate correlation for papers cited at least once. In the second step, about half of the papers in the remaining dataset are cited no more than six times, so we remove them and calculate the correlation for the papers cited more than six times. In the last step, we keep only papers cited more than 36 times. Fig. 1 Plot b shows that correlations for highly cited papers are lower than for the whole population. The correlation between five-year citation counts and total citation counts is 0.87 for all papers, 0.77 for papers cited at least once, 0.66 for papers cited more than six times, 0.57 for papers cited more than 18 times, and 0.50 for papers cited more than 36 times.

In addition to citation counts, number or share of papers in the top ***z***% (e.g. 10%) of highly cited papers is another commonly used indicator for evaluating research impact of individuals, institutions, and countries. Therefore, we further identify the top 10% of papers in each time window and count how many of them remain in this elite (i.e. top 10%) group in year 31. As shown in Fig. 1 Plot c, elite papers are not identifiable in the first several years. All papers are in the "top 10%" in the first year because there are a large number of papers with few citations and many ties in the citation count rankings. In the first year or two, elite papers have not gotten enough time to distinguish themselves. Starting in the fourth year, a distinct top 10% group can be identified. However, it is unstable over time. Only 68.3% of the elite papers in year four and five will remain as elite through the final year, that is, when we use a five-year citation time window to evaluate research units by their number of elite papers, more than 30% of the elite papers will turn out to be not elite in the end. The situation is even worse with a three-year window, which is another commonly used time window in bibliometrics studies. More than 40% papers identified as elites by the third year will not be elite in the final year. The percentage of 'final' elite papers, namely elite papers in the final year, increases to 82% in year 10, 92% in year 20, and eventually 100% in year 29.

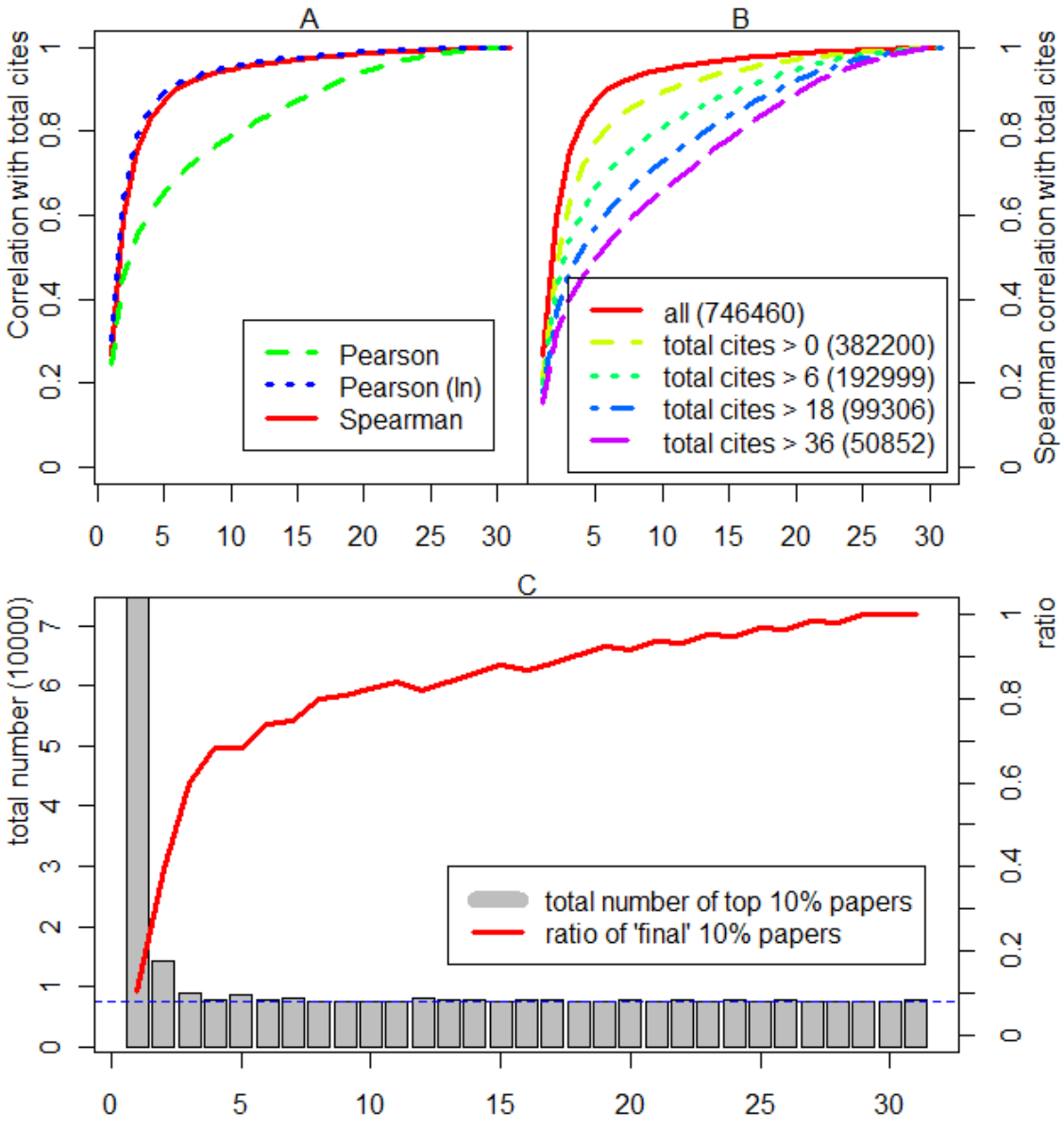

Fig. 1 Time window accuracy evaluation: Based on dataset 1. X-axes are year after publication (i.e. year 1 to 31 correspond to 1980 to 2010 respectively). Plot a reports three correlations between cumulative non-self-citation counts in each year and total non-self-citation counts in year 31 for all papers (i.e. 746,460 papers). Plot b reports Spearman correlations for different sets of papers (e.g. 382,200 papers with at least one total non-self-citation). "Total number of top 10% papers" in Plot c is the number of papers with citation counts above the 10th percentile. Top 10% papers are not identifiable in the first two years because of too many ties in the citation count rankings. "The ratio of 'final' 10% paper" in Plot c is the fraction of top 10% papers identified in year ***x*** that are actual top10% paper in the final year

Between group ageing differences

To investigate ageing patterns, we calculate the ratio between cumulative non-self-citation counts in each time window and total non-self-citations and then analyze the ratio trend over time. Fig. 2 plots the median ratio for each group, that is, one point on one line indicates the median of ratios between cumulative citation counts in the given year and total citation counts, for all papers in the given group (e.g. meeting abstracts, tier 4, or multidisciplinary sciences). The "early rise - rapid decline" papers will have a very steep increase in a short time period and then stay at the 100% level, while delayed documents will have a slower growth.

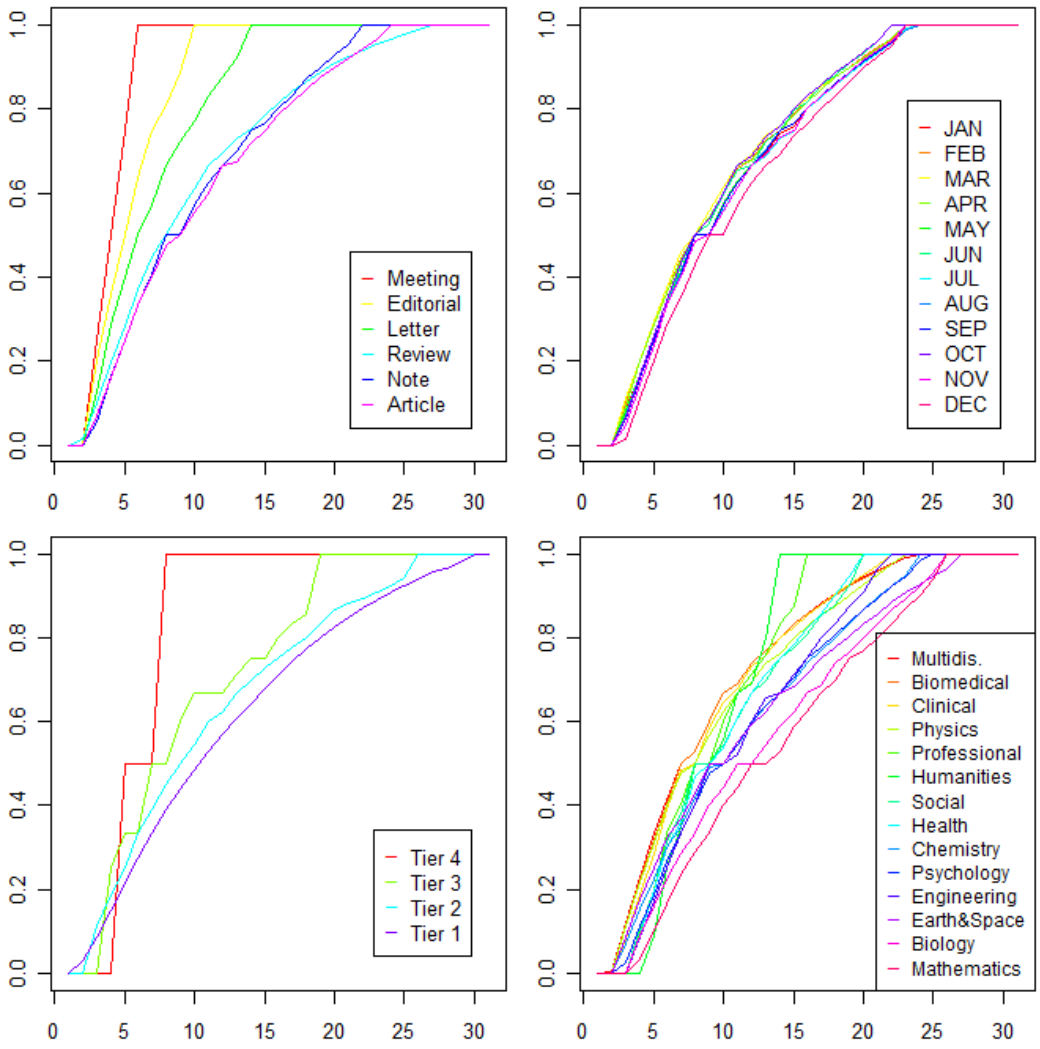

Fig. 2 Citation ageing comparison: Based on dataset 2. X-axes are year. Y-axes are the median ratio between cumulative non-self-citation counts in year ***x*** and total non-self-citation counts

Subsequently we focus on two aspects of citation ageing: "maturing" and "decline" (Glänzel and Schoepflin 1995). We further extract the starting and ending year of citations for each group and produce Fig. 3. Take 'review' as an example, its coordinate is (2, 27), meaning that among all cited review papers, more than half are cited from year 2 to 27, in other words, less than half are cited before year 2 or after year 27. For all cited papers in all groups, the coordinate is (3, 26), so we take (3, 26) as the center and divide the coordinate system into four quadrants. With a coordinate left to the center indicates that this group starts to be cited relatively earlier, and with a coordinate below the center indicates that this group stops being cited earlier. In terms of document type, reviews start to be cited earliest, in the second year, while all other types started to be cited in the third year. Citations of reviews also last longest, while citations of meeting abstracts last shortest. More highly cited papers start to be cited earlier and stop being cited later. Citations of the most highly cited papers (tier 1) start in year 2 and end in year 30, while citations of the least cited papers (tier 4) start in year 5 and end in year 8. This is also in line with previous findings that the accuracy of using short time windows is lower for highly cited papers, because they have longer citation life and therefore require longer time period to reveal their full impacts. There is not much difference between publication months, the coordinates for all months locate at (3, 22), (3, 23), or (3, 24), and therefore they are not plotted to reduce the crowdedness.

Regarding the research field, we confirm previous findings that citations of papers in the biomedical fields rise very quickly while in the humanities it takes a longer time to get recognized and cited. However, we observe another interesting phenomenon: Citations of papers in biomedical and clinical medicine fields not only rise very quickly but also last very long. Taking the biomedical fields as an example, more than half of the cited papers are cited between year 3 and 22. Citations of the humanities papers start rising the latest and terminate the earliest. Citations of papers in mathematics and biology start rising very late and last very long. Multidisciplinary science papers' citations rise earliest (in year 2) and earth and space papers' citations end latest (in year 26). Another point worth noting is the field of biology: Biology in general is a very heterogeneous field; therefore, the NSF field classification scheme adapted in this paper has two fields: "biology" and "biomedical science." The former includes subfields such as agricultural & food sciences, botany, ecology, and zoology, while the latter includes

subfields such as anatomy & morphology, biochemistry & molecular biology, biophysics, and genetics & heredity.

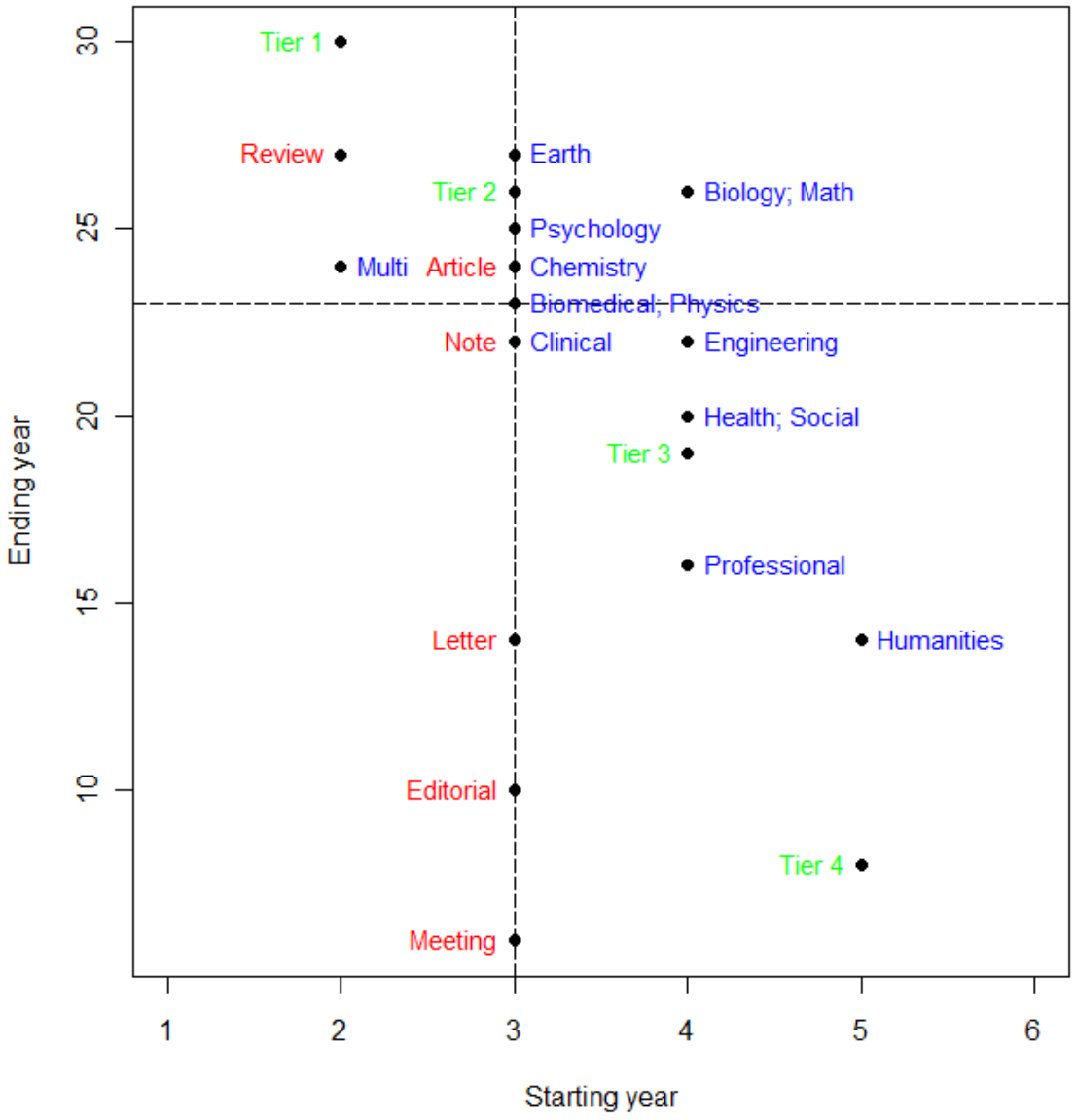

Fig. 3 Citations starting and ending year comparison: Based on dataset 2. Starting and ending year correspond to the first and last year that more than half of the cited papers are cited. Take 'review' as an example, its starting year is 2, meaning among all cited review papers, more than half of them are not cited before year 2, in other words, more than half of them are cited in and after year 2. Its ending year is 27, meaning among all cited review papers, more than half of them are no longer cited after year 27, in other words, more than half of them are cited in and before year 27. The center (3, 26) is the starting and ending year for the whole dataset of 358,100 cited papers

To further simplify the comparison, we construct a single indicator, ***Citation Speed***, to measure how fast in general a paper accumulates its citations:

$$Citation\ Speed = \frac{\sum_{1}^{n-1} C_i / C_n}{n-1}$$

where C_i is the cumulative citation count by year i, and n is the number of years, which is 31 in this paper. Since the cumulative citation ratio is monotonically increasing, fast ageing papers rise early and then stay at the high level, so they will have a high value of ***Citation Speed***. Fig. 4 provides three examples, all of them have 31 citations in total, paper A receives all 31 citations in the first year and therefore gets a citation speed value of 1, paper B receives one citation each year and therefore gets a citation speed value of 0.5, paper C receives all 31 citations in the final year and therefore gets a citation speed value of 0. Hence, papers get a higher citation speed value when they accumulate their citations faster. However, this simplification comes with a price, namely the loss of details about the citation maturation and decline; we cannot distinguish between fast ageing due to early rise and fast ageing because of early decline.

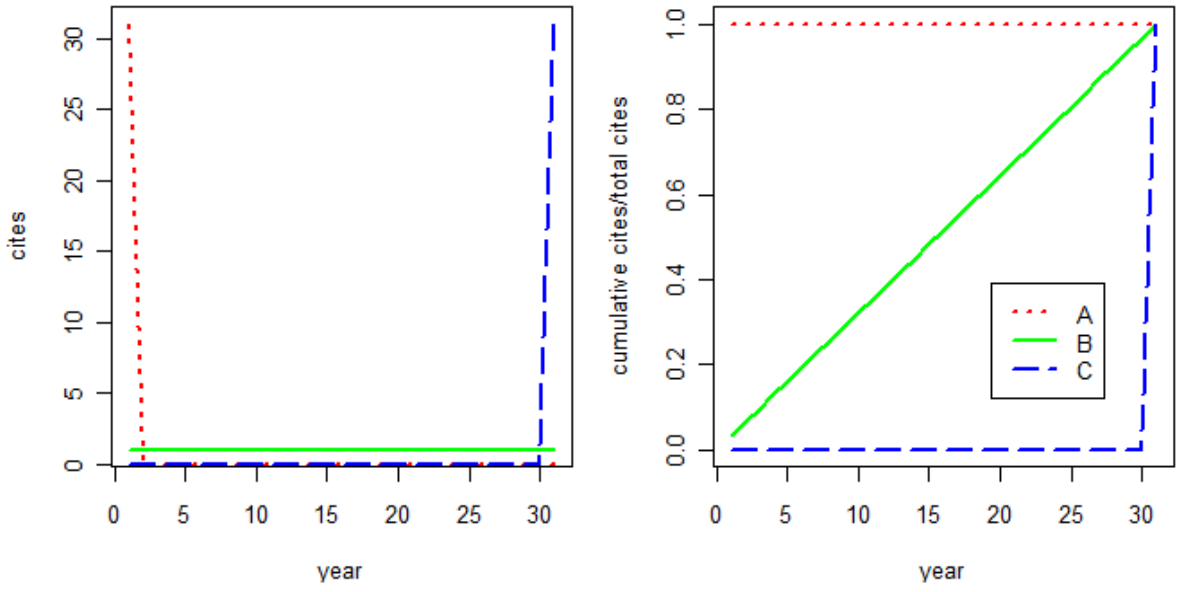

Fig. 4 Citation speed illustration

Fig. 5 shows the distribution of ***Citation Speed*** in each group. It tells similar stories: citation ageing in mathematics is the slowest and in multidisciplinary sciences the fastest, citation ageing of articles the slowest and of meeting abstracts the fastest. ANOVA analysis confirms that there are significant differences in citation aging between different research fields, document types, total citation tiers, and publication months.

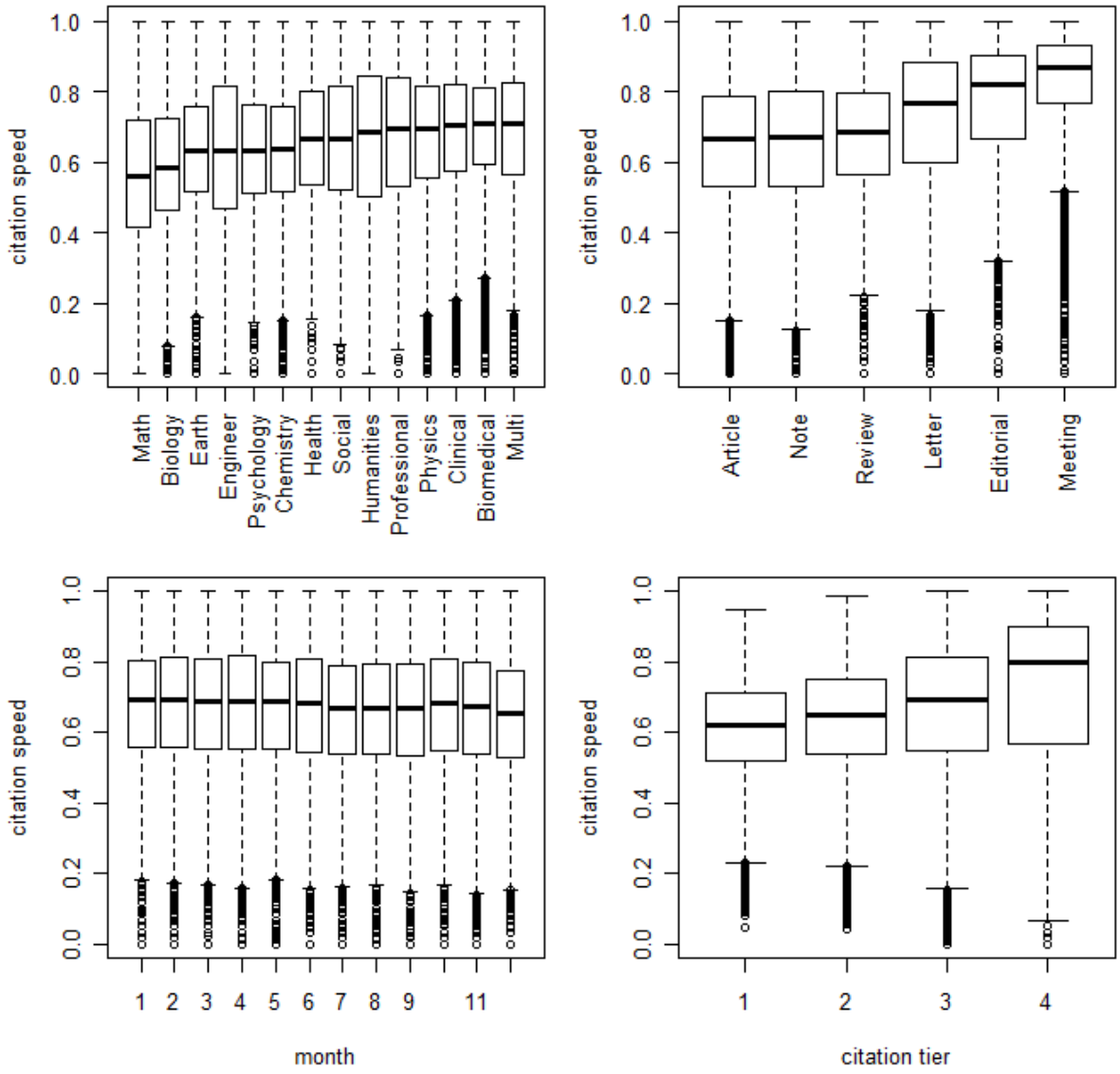

Fig. 5 Citation speed comparison: Based on dataset 2. The bar in the middle of the box is the median (i.e. the second quartile), the upper and lower boundary of the box indicate the third and first quartile respectively, the upper and lower bar outside the box are the theoretical maximum and minimum respectively, and circles are considered as outliers

Within group ageing differences

Compared with the significant difference in group means confirmed by ANOVA, the finding of remarkable differences within the group is more striking and disturbing. Boxplots in Fig. 5 show that ***Citation Speed*** distributes very diversely in each group and overlaps between different groups. For example, multidisciplinary sciences are the fastest ageing field and mathematics the slowest, but citations of many mathematics paper (those in the top part of the box) may still age faster than citations of many multidisciplinary sciences papers (those in the bottom part of the box). Similarly, although citations of articles age slowest and of meeting abstract fastest, there are still a considerable number of articles with citations ageing faster than a number of meeting abstracts. In other words, although the mean of citation speed is significantly different between different groups, this difference is not very powerful to predict citation speed at the individual paper level.

We are particularly interested in testing the field homogeneity assumption, so we further investigate the citation ageing differences between subfields in the same research field. Nine fields are selected and median of cumulative citation ratios for each subfield is plotted in Fig. 6. Compared with the citation ageing assessment at the field level in Fig. 3, citation ageing at the subfield level within the same field is not more homogeneous. There are remarkable differences between different subfields in the same field. Furthermore, although biology is one of the slowest ageing fields and clinical medicine one of the fastest, many subfields in biology may still age faster than many subfields in clinical medicine.

However, this finding may be compromised if our field delineation is not perfect or the subfield is still a too high level for analysis. To address such concerns, we further narrow down our analysis to the journal level. Ten research fields are selected, and two journals are further selected from each field. One journal with very broad interests and general coverage and the other journal with specialized and narrow focuses are selected from each field except for the multidisciplinary sciences (Table 1). Only top journals in each field are selected to control for the effect of journal quality/reputation. If further narrowing down research field can improve citation homogeneity, we would expect that citation ageing of specialized journals is more homogeneous than general journals, because general journals cover more diverse research subjects. However, this hypothesis is not supported by empirical findings. For example, "Neurology" is more specialized than "New England Journal of Medicine," but the citation speed of its papers spreads

wider. Similarly, “Journal of Organic Chemistry” is more specialized than the “Journal of the American Chemistry Society,” but the citation speed of its papers spreads wider and has many more outliners (Fig. 7). Another possible explanation for the fact that citation ageing of specialized journals is not more homogeneous than general journals may be: General journals have much more submitted manuscripts to choose from, and therefore published papers could be more homogeneous to one another, while specialized journals have to publish more heterogeneous papers because of limited choices. However, this possibility can be ruled out by looking at the left plot in Fig. 7. In some cases, papers in general journals have higher total citations, but in some other cases, papers in specialized journals have higher total citations. However, we cannot find evidence that the total citations distribution is more spread for papers published in specialized than general journals.

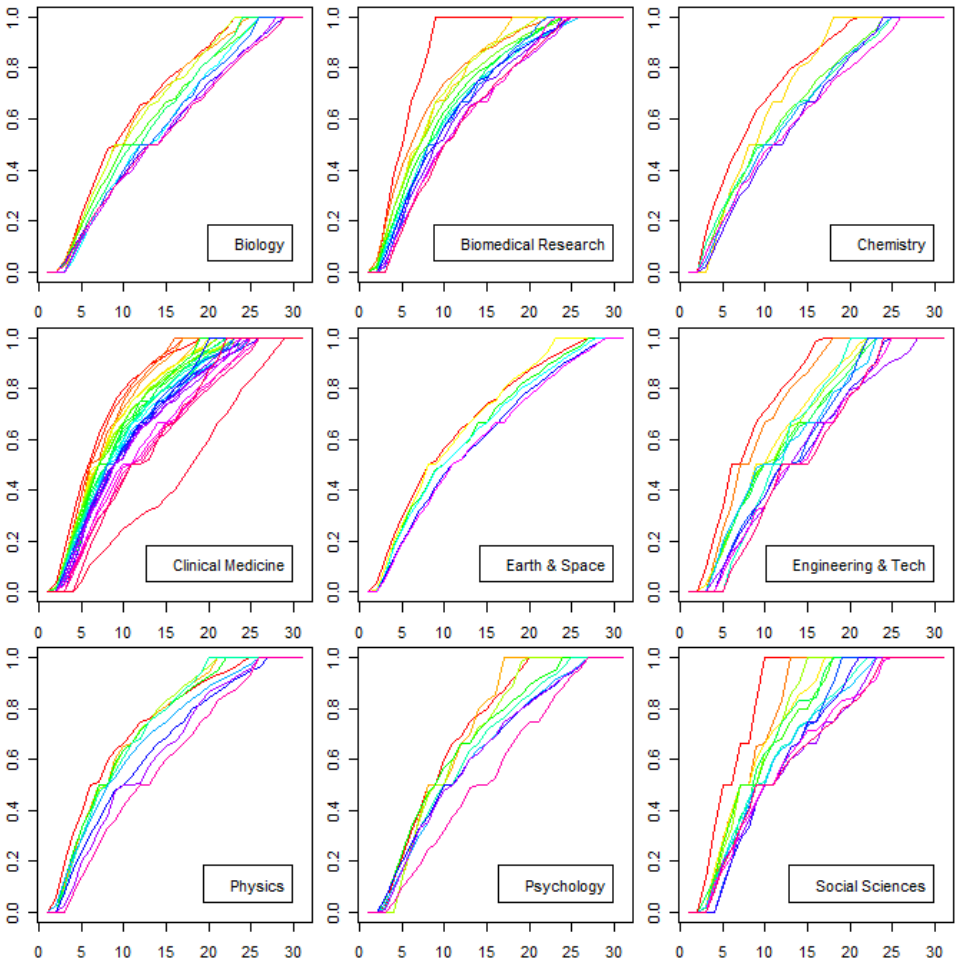

Fig. 6 Within field citation ageing comparison: Based on dataset 2. X-axes are year. Y-axes are the median ratio between cumulative non-self-citation counts in year ***x*** and total non-self-citation counts

Table 1 Selected journals

Field	Coverage	Jounal Title	Abbreviation
Multidisciplinary	General	Science	SCIENCE
Multidisciplinary	General	Nature	NATURE
Mathematics	General	Journal of Mathematical Analysis and Applications	J MATH ANAL APPL
Mathematics	Specialized	Journal of Differential Equations	J DIFFER EQUATIONS
Clinical Medicine	General	New England Journal of Medicine	NEW ENGL J MED
Clinical Medicine	Specialized	Neurology	NEUROLOGY
Physics	General	Physical Review Letters	PHYS REV LETT
Physics	Specialized	Physical Review A	PHYS REV A
Chemistry	General	Journal of the American Chemical Society	J AM CHEM SOC
Chemistry	Specialized	Journal of Organic Chemistry	J ORG CHEM
Engineering & Tech	General	International Journal of Engineering Science	INT J ENG SCI
Engineering & Tech	Specialized	IEEE Transactions on Information Theory	IEEE T INFORM THEORY
Psychology	General	Psychological Bulletin	PSYCHOL BULL
Psychology	Specialized	Journal of the American Academy of Child and Adolescent Psychiatry	J AM ACAD CHILD PSY
Social Sciences, Sociolgy	General	American Sociological Review	AM SOCIOL REV
Social Sciences, Sociolgy	Specialized	Journal of Marriage and the Family	J MARRIAGE FAM
Social Sciences, Economics	General	American Economic Review	AM ECON REV
Social Sciences, Economics	Specialized	Journal of Political Economy	J POLIT ECON
Humanities	General	Critical Inquiry	CRIT INQUIRY
Humanities	Specialized	American Historical Review	AM HIST REV

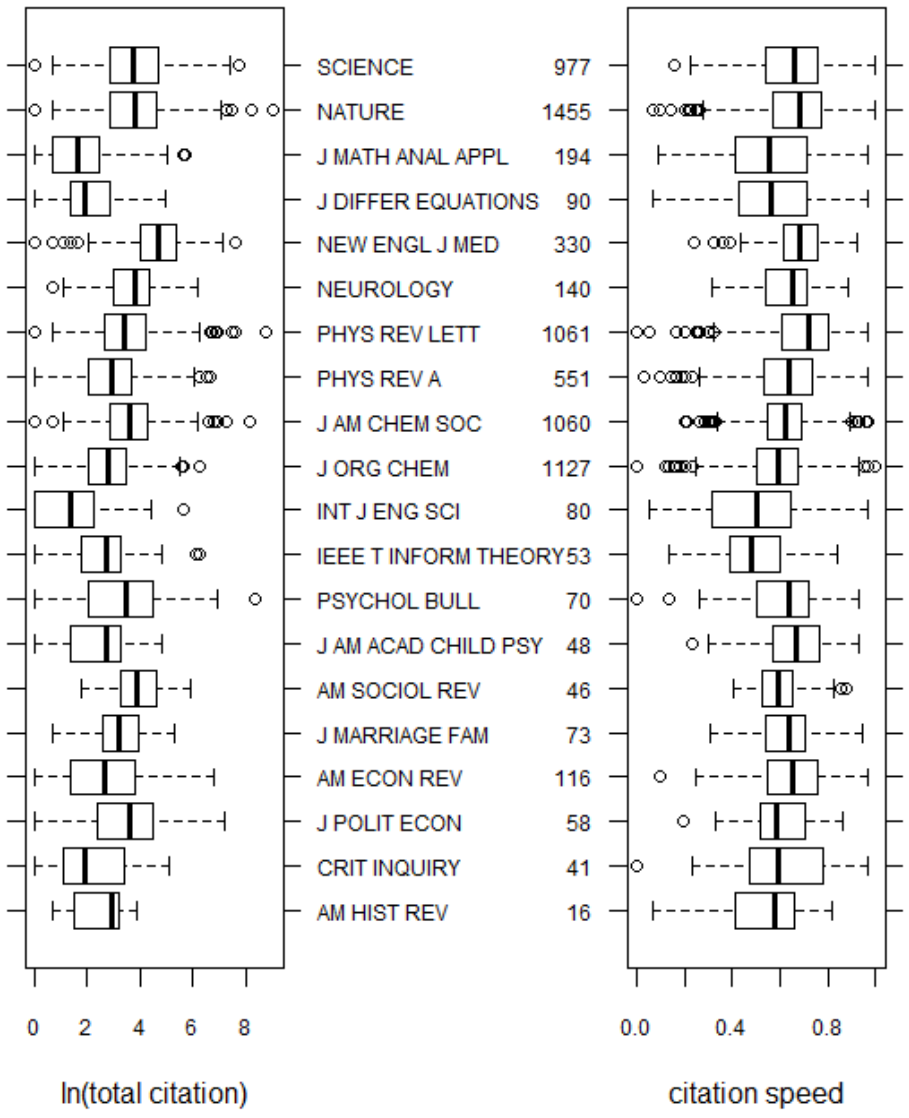

Fig. 7 Total citation counts and citation speed comparison for selected journals: Based on dataset 2. Number on the right side of journal title abbreviation is the number of papers. The bar in the middle of the box is the median (i.e. the second quartile), the left and right boundary of the box indicate the first and third quartile respectively, the left and right bar outside the box are the theoretical minimum and maximum respectively, and circles are considered as outliers

Furthermore, we investigate within journal citation ageing differences. If we plot the cumulative citation ratio trend of each paper, within each journal, the ratio trends at the paper level would be very heterogeneous and spread the whole plotting area, indicating that citation ageing characteristics are primarily specific to the individual article rather than to the journal.

Moreover, we control for the quality of the paper (as indicated by total number of citations) and compare cumulative citation ratio trends of papers with exactly the same number of total citations and published in the same journal. In Fig. 8, we can still find very heterogeneous citation ageing patterns. Therefore, even within the same journal and controlling for paper quality, the assumption of homogeneity in citation ageing is not valid.

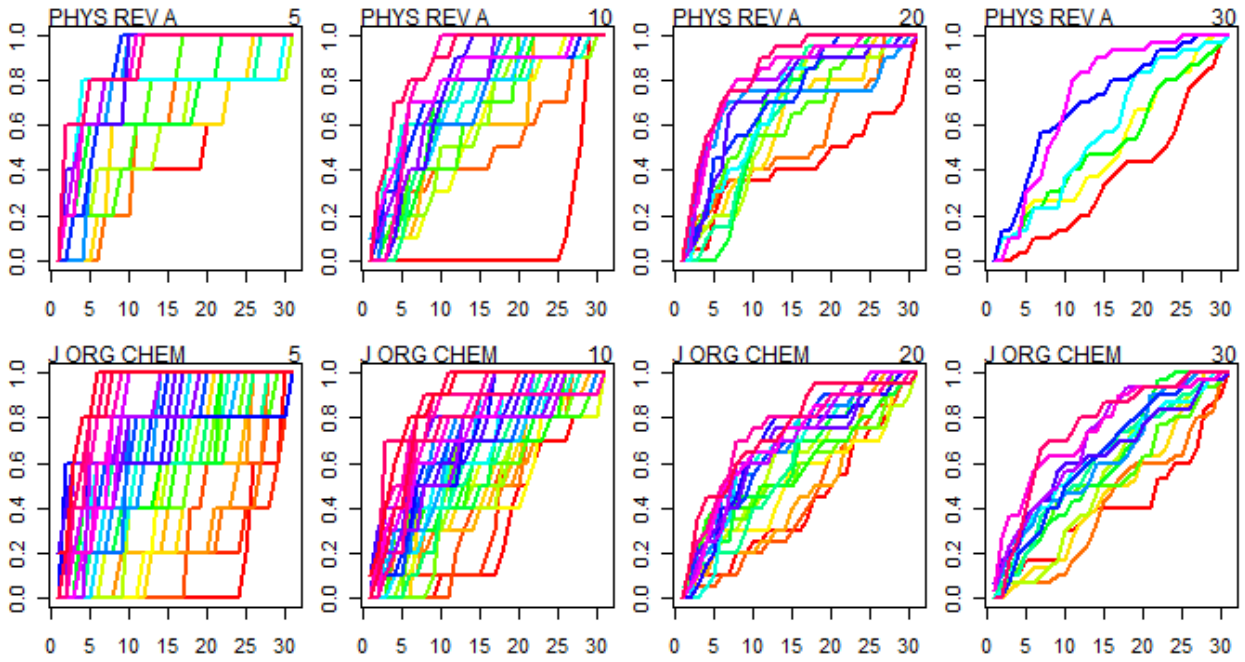

Fig. 8 Within journal citation ageing comparison after controlling for paper quality: Based on dataset 2. X-axes are year. Y-axes are the ratio between cumulative non-self-citation counts in year ***x*** and total non-self-citation counts. Only two journals are plotted (i.e. Physical Review A and Journal of Organic Chemistry), because: First, we only select specialized journals to rule out the possibility that the heterogeneity of citation ageing can be explained by diverse field coverage, and second, many other specialized journals have limited number of papers with exactly the same total citations and therefore cannot be compared. The number at the top-right corner of each plot indicates total citations, for instance, the top-left plot shows the citation ageing of papers published in PHYS REV A and having five citations in total

Field normalization

Size and ageing differences are two sources of field variations preventing cross-field comparison of citation-based indicators. However, field normalization methods pay attention exclusively to the size differences but overlook the ageing differences. Therefore, we test if field

normalization can improve the accuracy of using short citation time windows. The Spearman correlations between normalized cumulative non-self-citation counts in each time window and normalized total non-self-citations are plotted in Fig. 9. Two normalizations used here are: field and document type normalization (i.e. citation count/mean citation count of the same field and document type) and journal and document type normalization (i.e. citation count/mean citation count of the same journal and document type). The results suggest that these two normalizations cannot improve the accuracy of using a short time window. Instead, journal and document type normalization performs much worse. This finding is in line with the observation of remarkable difference in citation ageing within the same field. Field normalization may help to eliminate the between-field differences caused by the size and ageing differences, but is still unable to eliminate the within-field ageing differences.

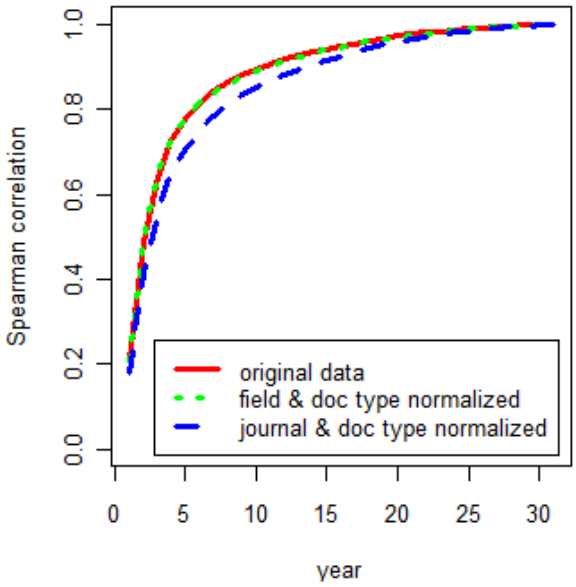

Fig. 9 Spearman correlations: Based on dataset 2

Are all these findings relevant today?

All these analyses are based on papers published more than 30 years ago, so one question is: Are these findings still relevant to today's research evaluation? It is possible that citation behavior has changed so much in the last 30 years that short citation time window is no longer that problematic. To address this question, we further compare citation behavior between papers

in different cohorts, that is, we compare citations of all WoS journal papers published in 1980, 1990, and 2000. For each research field and cohort, we count the mean non-self-citations in each year. As shown in Fig. 10, in all research fields except humanities, recent papers have higher mean citation counts than older papers. Although the mean citation count rises as cohort year increases, it does not peak earlier nor decline faster, so our findings do not exaggerate the problem of using short citation time windows. On the contrary, citations in many fields seem to peak later and decline slower, so it is possible that the problem is even worse than before. However, we do not have a sufficient long time period to study more recent papers or evaluate rigorously how the citation ageing patterns have changed over time. In sum, we conclude that our findings are still relevant today and can help to inform choice of citation time windows in research evaluation practices.

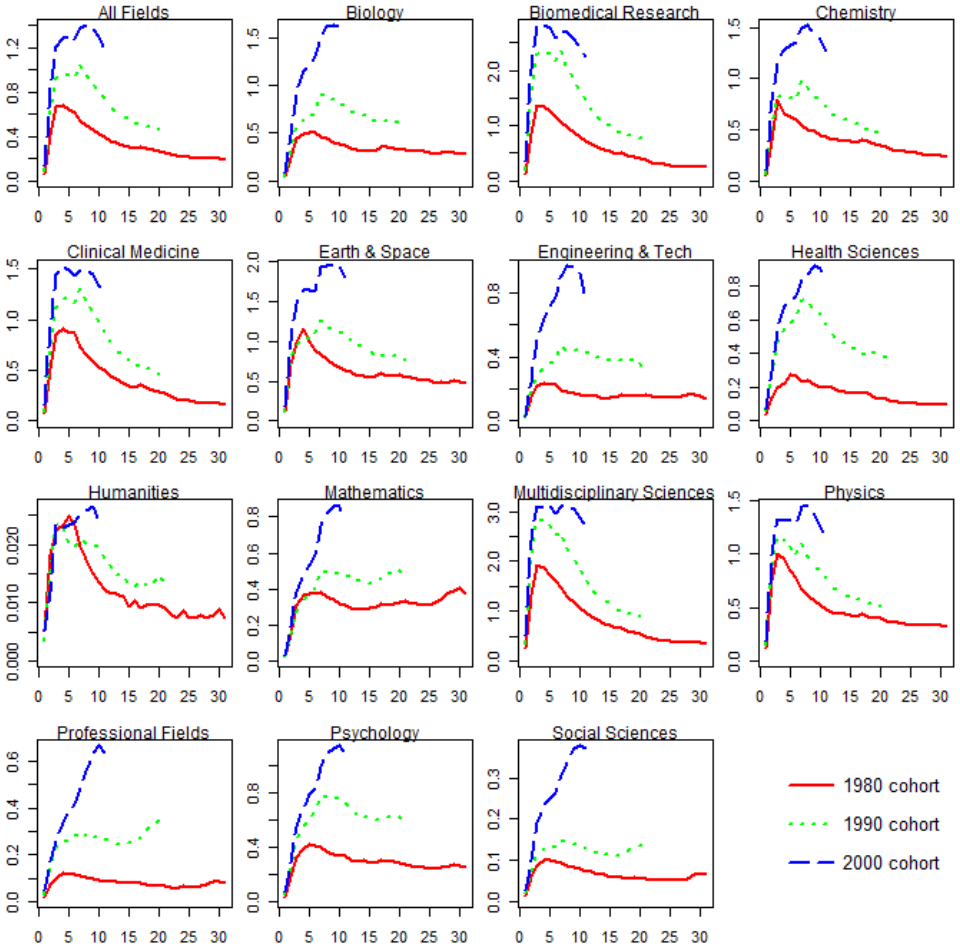

Fig. 10 Paper cohort comparison: 1980 cohort data are based on dataset 1. 1990 and 2000 cohort data also include all WoS journal papers published in 1990 and 2000 respectively. X-axes are year after publication. Year 1-31

correspond to 1980-2010, 1990-2020, and 2000-2030 for cohort 1980, 1990, and 2000 papers respectively. Y-axes are average non-self-citation counts, that is, citation counts in this plot for a given year is not cumulative citations till this year, but only citations received in this year

Discussion

We calculate the correlations between cumulative non-self-citation counts in short time windows and total non-self-citations. The Spearman correlation rises from 0.266 in year 1 to 0.756 in year 3, and then slowly reaches 1 in year 31. Furthermore, the correlation is higher for all papers than for highly cited papers, and if we look at the top 10% most cited papers, more than 30% of the papers recognized as elite in year 5 will not be elite in year 31. This time window accuracy evaluation aims to inform research. Unfortunately, there is no rule of thumb to use in deciding what level of correlation is acceptable. The choice depends on the accuracy requirement, timeliness demand, and data availability. Furthermore, it also depends on the purpose of the evaluation: Whether to detect the research front or to assess research impacts, and whether to identify the 'elites' or to evaluate the 'masses.' To identify the research front and current impact, using short time window is theoretically justified while total citation is irrelevant to the quest (Garfield 1986; Leydesdorff 2009). In addition, findings of this paper suggest a longer time window for screening out elites because the accuracy of using shorter citation time window is worse for elites than for lowly cited papers.

The Spearman correlations and percentages of 'final' top 10% most cited papers are reported in Table 1 in the Appendix to inform choice of citation time windows. If a research evaluation project evaluates general impact of all papers and views a correlation of 0.8 as adequate, then a four-year window may be sufficient. However, if a project aims to identify top researchers by looking at their share in top 10% cited papers and takes 20% as the highest acceptable error rate, then a citation time window of at least nine years is required. In addition, maybe researchers should report the potential errors in their evaluations when using short time windows, providing a paragraph such as: "Although a citation window of five years is used here, note that the Spearman correlation between these citation counts and long term (31 year) citation counts will be about 0.87. Furthermore, the potential error of using a five-year time window will

be higher for highly cited papers because papers in the top 10% most cited papers in year 5 have a 32% chance of not being in the top 10% in year 31."

In addition, there are significant differences in citation ageing between different research fields. For studies on one specific field, a tailored citation time window is preferred. For example, if 0.8 is viewed as an adequate Spearman correlation for the evaluation, then a three-year time window is sufficient for the biomedical research fields and multidisciplinary sciences, while a seven-year time window is required for the humanities and mathematics. Table 2 in the Appendix reports the Spearman correlations by field to inform choice of citation time windows for each research field.

Furthermore, compared with significant between-group citation ageing differences, more attention should be given to the within-group variations. Many subfields in the slowest ageing field may still age faster than many subfields in the fastest ageing field. This finding also applies at the paper level. Even in the same journal and controlling for paper quality, papers show very different ageing patterns. Therefore, although the group means are significantly different, this difference is not a powerful predictor of citation ageing at the paper level. These findings imply that narrowing down research fields to finer units would not improve the ageing homogeneity within the unit. In line with these findings, field normalization cannot improve the accuracy of using short time windows. These findings reveal a more fundamental risk in using citation-based indicators: the citation behavior is so heterogeneous that there is little common ground for reliable comparisons, and the heterogeneity cannot be controlled or reduced by the set of variables at our disposal, such as research field and document type.

Although the citation behavior covers many aspects other than ageing, our findings regarding the field ageing homogeneity can still inform field normalization studies. Field normalization can be done at various levels: field, subject categories, journals, and so on. Evaluatees complain about using field normalization when their subfield is in a disadvantaged position in the field and advocate for subfield normalizations. The question is when to stop further level refinement. In one extreme case, every paper is somehow unique and we can normalize at the paper level, that is, normalize every paper by itself, then the evaluation cannot make any distinctions between papers. Besides this argument against a too fine level for normalization, our findings further suggest that homogeneity does not increase as the level goes finer. Citation ageing patterns are specific to the individual paper rather than to the journal,

subfield, or field. Therefore, normalization at finer level is still unable to achieve its goal of improving homogeneity for a fairer comparison.

There are also limitations of this paper: First, the accuracy of using short citation time windows is investigated at the individual paper level but not the author or institution level. If we can assume that the shares of slow and fast ageing papers are the same for all focal authors and institutions to be evaluated, then using short citation time windows would penalize every evaluatee equally and therefore is less problematic for evaluation purposes. This assumption is more likely to be true at the institution than the author level, and previous literature also found that using short citation time windows changed evaluation results considerably at the author level (Costas et al. 2011; Abramo et al. 2012b), but not that much at the institution level (Glänzel 2008; Abramo et al. 2012a). Second, the analysis is based on papers published in 1980. Although we have demonstrated that the findings are still relevant today, we do not have a sufficient long time period to study more recent papers or evaluate rigorously how the citation ageing patterns have changed over time. Third, our field classification is not perfect and our investigation on field normalization is not exhaustive. The NSF field classification scheme is adapted in this paper, which is not perfect or the only option. Furthermore, there are also convincing arguments for field delineation at the paper level rather than the journal level. In addition, many other advanced field normalized indicators are proposed in literature but not tested in this paper.

Acknowledgements: The author would like to thank Stefan Hornbostel, Sybille Hinze, and William Dinkel for their efforts in the early stage of project initiation and research design, Diana Hicks and Daniel Sirtes for their suggestions which were most helpful in improving the paper, Jasmin Schmitz, Haiko Lietz, Marion Schmidt, Pei-Shan Chi, and Jana Schütze for their many helpful ideas and collegial support, and two anonymous reviewers for their critical and constructive comments. The research underlying this paper was supported by the German Federal Ministry for Education and Research (BMBF, project number 01PQ08004A). The data used in this paper are from a bibliometrics database developed and maintained by the Competence Center for Bibliometrics for the German Science System (KB) and derived from the 1980-2011 Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and Arts & Humanities Citation Index (AHCI) prepared by Thomson Reuters (Scientific) Inc. (TR®), Philadelphia, Pennsylvania, USA: © Copyright Thomson Reuters (Scientific) 2012. The author thanks the KB team for its collective effort in the development of the KB database.

Appendix

Table 2 Accuracy of using short citation time windows (based on dataset 1)

	Spearman correlations with total citations (31 years)					
Year	all papers	total cites>0	total cites>6	total cites>18	total cites>36	Percentage of 'final' top 10% papers
1	0.266	0.208	0.197	0.177	0.153	10.3%
2	0.592	0.484	0.425	0.368	0.323	39.1%
3	0.754	0.637	0.544	0.465	0.402	59.7%
4	0.830	0.720	0.614	0.525	0.453	68.3%
5	0.872	0.773	0.665	0.572	0.497	68.3%
6	0.900	0.813	0.707	0.613	0.537	73.9%
7	0.918	0.840	0.739	0.646	0.572	74.6%
8	0.931	0.862	0.766	0.676	0.603	79.8%
9	0.940	0.879	0.790	0.704	0.631	80.9%
10	0.948	0.893	0.811	0.729	0.659	82.2%
11	0.954	0.906	0.830	0.753	0.685	83.8%
12	0.959	0.916	0.847	0.776	0.711	82.1%
13	0.963	0.925	0.863	0.797	0.736	84.0%
14	0.967	0.933	0.877	0.817	0.760	85.9%
15	0.970	0.940	0.890	0.836	0.782	87.8%
16	0.973	0.947	0.902	0.854	0.804	86.8%
17	0.977	0.954	0.915	0.873	0.829	88.6%
18	0.980	0.961	0.927	0.890	0.851	90.6%
19	0.983	0.967	0.938	0.906	0.872	92.5%
20	0.985	0.972	0.948	0.921	0.891	91.7%
21	0.987	0.976	0.957	0.934	0.909	93.8%
22	0.989	0.980	0.964	0.946	0.925	93.1%
23	0.991	0.984	0.971	0.955	0.938	95.1%
24	0.992	0.987	0.977	0.965	0.951	94.6%
25	0.994	0.989	0.982	0.973	0.962	96.7%
26	0.995	0.992	0.986	0.980	0.972	96.3%
27	0.996	0.994	0.990	0.986	0.981	98.4%
28	0.997	0.996	0.994	0.991	0.988	98.1%
29	0.998	0.997	0.996	0.995	0.994	100.0%
30	0.999	0.999	0.999	0.998	0.998	100.0%
31	1.000	1.000	1.000	1.000	1.000	100.0%

Table 3 Spearman correlation with total citations by field (based on dataset 1)

Year	Biology	Biomedical Research	Chemistry	Clinical Medicine	Earth & Space	Engineering & Tech	Health Sciences	Humanities	Mathematics	Multidisciplinary Sciences	Physics	Professional Fields	Psychology	Social Sciences
1	0.174	0.295	0.229	0.258	0.284	0.203	0.244	0.199	0.171	0.422	0.307	0.227	0.234	0.227
2	0.464	0.657	0.547	0.602	0.622	0.466	0.488	0.407	0.386	0.735	0.637	0.489	0.534	0.486
3	0.656	0.812	0.739	0.767	0.777	0.636	0.647	0.541	0.571	0.827	0.777	0.634	0.707	0.636
4	0.752	0.873	0.811	0.844	0.851	0.734	0.741	0.637	0.684	0.874	0.840	0.733	0.802	0.731
5	0.810	0.906	0.852	0.886	0.888	0.792	0.813	0.711	0.750	0.900	0.873	0.791	0.858	0.792
6	0.848	0.930	0.881	0.915	0.910	0.835	0.861	0.768	0.795	0.919	0.898	0.836	0.892	0.834
7	0.874	0.943	0.899	0.930	0.925	0.861	0.887	0.804	0.826	0.931	0.914	0.861	0.915	0.864
8	0.893	0.953	0.914	0.942	0.937	0.880	0.908	0.832	0.848	0.940	0.927	0.881	0.930	0.884
9	0.907	0.960	0.926	0.950	0.945	0.895	0.923	0.852	0.868	0.948	0.937	0.897	0.941	0.902
10	0.918	0.966	0.935	0.957	0.952	0.906	0.933	0.869	0.883	0.953	0.945	0.908	0.949	0.914
11	0.927	0.971	0.943	0.962	0.957	0.916	0.941	0.882	0.896	0.958	0.951	0.919	0.956	0.924
12	0.935	0.974	0.949	0.966	0.961	0.924	0.951	0.893	0.908	0.962	0.956	0.927	0.961	0.932
13	0.942	0.978	0.955	0.970	0.965	0.932	0.958	0.904	0.917	0.965	0.961	0.939	0.965	0.940
14	0.947	0.980	0.960	0.973	0.969	0.938	0.964	0.912	0.925	0.968	0.965	0.947	0.969	0.946
15	0.952	0.982	0.965	0.975	0.972	0.944	0.968	0.918	0.933	0.971	0.969	0.954	0.972	0.952
16	0.957	0.985	0.969	0.978	0.975	0.950	0.973	0.927	0.940	0.974	0.972	0.960	0.976	0.957
17	0.963	0.987	0.974	0.981	0.979	0.956	0.975	0.934	0.947	0.977	0.976	0.965	0.979	0.961
18	0.968	0.989	0.978	0.983	0.981	0.961	0.979	0.941	0.953	0.980	0.979	0.969	0.981	0.966
19	0.972	0.991	0.981	0.986	0.984	0.966	0.981	0.948	0.959	0.983	0.982	0.973	0.984	0.969
20	0.976	0.992	0.984	0.989	0.987	0.971	0.983	0.955	0.964	0.985	0.985	0.976	0.986	0.973
21	0.980	0.993	0.986	0.990	0.989	0.974	0.985	0.960	0.969	0.987	0.987	0.980	0.988	0.976
22	0.983	0.994	0.989	0.992	0.991	0.977	0.986	0.965	0.973	0.989	0.989	0.983	0.990	0.979
23	0.986	0.995	0.991	0.993	0.993	0.980	0.989	0.969	0.978	0.991	0.991	0.985	0.991	0.981
24	0.988	0.996	0.992	0.995	0.994	0.984	0.989	0.973	0.981	0.992	0.992	0.987	0.992	0.984
25	0.990	0.997	0.994	0.996	0.995	0.986	0.991	0.977	0.984	0.993	0.994	0.989	0.993	0.986
26	0.992	0.998	0.995	0.997	0.996	0.988	0.993	0.981	0.987	0.995	0.995	0.991	0.995	0.988
27	0.994	0.998	0.996	0.997	0.997	0.990	0.994	0.986	0.990	0.996	0.996	0.993	0.996	0.990
28	0.996	0.999	0.997	0.998	0.998	0.993	0.996	0.990	0.992	0.997	0.997	0.994	0.997	0.993
29	0.997	0.999	0.998	0.999	0.999	0.995	0.997	0.993	0.995	0.998	0.998	0.996	0.998	0.995
30	0.999	1.000	0.999	0.999	0.999	0.998	0.998	0.997	0.998	0.999	0.999	0.998	0.999	0.997
31	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

References

Abbott, A. (2009). Italy introduces performance-related funding. [News Item]. *Nature, 460*(7255), 559-559, doi:10.1038/460559a.

Abramo, G., Cicero, T., & D'Angelo, C. A. (2011). Assessing the varying level of impact measurement accuracy as a function of the citation window length. *Journal of Informetrics, 5*(4), 659-667, doi:10.1016/j.joi.2011.06.004.

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012a). A sensitivity analysis of research institutions' productivity rankings to the time of citation observation. *Journal of Informetrics, 6*(2), 298-306, doi:10.1016/j.joi.2011.11.005.

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012b). A sensitivity analysis of researchers' productivity rankings to the time of citation observation. *Journal of Informetrics, 6*(2), 192-201, doi:10.1016/j.joi.2011.12.003.

Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics, 63*(3), 567-581.

Aksnes, D. W. (2003a). Characteristics of highly cited papers. *Research Evaluation, 12*(3), 159-170.

Aksnes, D. W. (2003b). A macro study of self-citation. *Scientometrics, 56*(2), 235-246, doi:10.1023/a:1021919228368.

Aversa, E. S. (1985). Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature. *Scientometrics, 7*(3), 383-389.

Costas, R., Van Leeuwen, T. N., & van Raan, A. F. J. (2010). Is scientific literature subject to a 'Sell By Date'? A general methodology to analyze the 'durability'of scientific documents. *Journal of the American Society for Information Science and Technology, 61*(2), 329-339.

Costas, R., van Leeuwen, T. N., & van Raan, A. F. J. (2011). The "Mendel syndrome" in science: durability of scientific literature and its effects on bibliometric analysis of individual scientists. *Scientometrics, 89*(1), 177-205.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the Science citation index to cybermetrics*. Lanham, MD: Scarecrow Press.

Garfield, E. (1980). Premature discovery or delayed recognition–Why. *Current Contents, 21*, 5-10.

Garfield, E. (1985a). The articles most cited in the SCI from 1961 to 1982. 7. Another 100 citation-classics–the Watson-Crick double helix has its turn. *Current Contents, 20*, 3-12.

Garfield, E. (1985b). The articles most cited in the SCI from 1961 to 1982. 8. Ninety-eight more classic papers from unimolecular reaction velocities to natural opiates-the changing frontiers of science. *Current Contents, 33*, 3-11.

Garfield, E. (1986). Letter to editor. *Information Processing & Management, 22*(5), 445.

Glänzel, W. (2008). Seven myths in bibliometrics. About facts and fiction in quantitative science studies. In H. Kretschmer, & F. Havemann (Eds.), *Proceedings of WIS 2008* (pp. 1-10). Berlin, Germany.

Glänzel, W., Debackere, K., Thijs, B., & Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics, 67*(2), 263-277, doi:10.1007/s11192-006-0098-9.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics, 58*(3), 571-586.

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of information Science, 21*(1), 37-53.

King, D. A. (2004). The scientific impact of nations. [Article]. *Nature, 430*(6997), 311-316, doi:10.1038/430311a.

Levitt, J. M., & Thelwall, M. (2008). Patterns of annual citation of highly cited articles and the prediction of their citation ranking: A comparison across subjects. *Scientometrics, 77*(1), 41-60.

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluations. *Journal of the American Society for Information Science and Technology, 59*(2), 278-287, doi:Doi 10.1002/Asi.20743.

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology, 60*(7), 1327-1336, doi:10.1002/asi.21024.

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. [Letter]. *Journal of Informetrics, 4*(4), 644-646, doi:10.1016/j.joi.2010.05.003.

Line, M. B. (1993). Changes in the use of literature with time: obsolescence revisited. *Library Trends, 41*(4), 665-683.

Moed, H. F., Burger, W., Frankfort, J., & van Raan, A. (1985). The application of bibliometric indicators: Important field- and time-dependent factors to be considered. *Scientometrics, 8*(3), 177-203, doi:10.1007/bf02016935.

Moed, H. F., van Leeuwen, T. N., & Reedijk, J. (1998). A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation, 54*(4), 387-419.

Porter, A. L. (1977). Citation Analysis: Queries and Caveats. *Social Studies of Science, 7*(2), 257-267, doi:10.1177/030631277700700207.

Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E, 83*(4), 046116.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. [Article]. *Proceedings of the National Academy of Sciences of the United States of America, 105*(45), 17268-17272, doi:10.1073/pnas.0806977105.

Rogers, J. D. (2010). Citation analysis of nanotechnology at the field level: implications of RD evaluation. *Research Evaluation, 19*(4), 281-290.

Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics, 36*(3), 311-324, doi:10.1007/bf02129597.

Stent, G. (1972). Prematurity and uniqueness in scientific discovery. *Scientific American, 227*(6), 84-93.

Van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467-472.

Walters, G. D. (2011). The citation life cycle of articles published in 13 American Psychological Association journals: A 25-year longitudinal analysis. *Journal of the American Society for Information Science and Technology, 62*(8), 1629-1636, doi:10.1002/asi.21560.