Research Policy 42 (2013) 886–894

Contents lists available at SciVerse ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Effects of the durability of scientific literature at the group level: Case study of chemistry research groups in the Netherlands

Rodrigo Costas*, Thed N. van Leeuwen, Anthony F.J. van Raan

Leiden University, Centre for Science and Technology Studies (CWTS), Wassenaarseweg 62A, 2333AL Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 12 June 2012
Received in revised form
29 November 2012
Accepted 30 November 2012
Available online 30 December 2012

Keywords:
Research evaluation
Research groups
Durability of scientific publications
Obsolescence

ABSTRACT

In this study an analysis of the effects of the different types of durability on the bibliometric performance at the group level is presented. The scientific production during the period of 1991–2000 of a set of 158 Dutch research groups in chemistry is studied considering several bibliometric indicators in the perspective of the durability of the publications in terms of the citations received. Two citation windows have been considered for the analysis of the effect of the enlargement of the citation period, one including the citations received in the same period of publications (1991–2000) and a second one including eight years more (1991–2008). In addition, qualitative indicators provided by a committee of experts who evaluated the research groups have been analyzed in order to study the relationship between qualitative indicators and quantitative measures, in particular these of durability. Results show that production with "normal" durability is the most rewarded both according to bibliometric indicators and qualitative assessments given by experts. We also find that publications with a delayed pattern do not represent a major problem in the assessment of research groups, as those groups with a higher share of this type of publications do not improve their assessment when the citation window is substantially enlarged. Several discussions are presented regarding the importance of durability analysis in the framework of research assessment situations.

1. Introduction

An important element of concern in research evaluation is the effect that the durability of the scientific literature and particularly the "delayed recognition" (Cole, 1970; Garfield, 1980) or "scientific prematurity" (Stent, 1972) can have on the indicators used in the evaluation of research and the development of research management and policy. Garfield (1970) claimed that "critics of citation indexes sometimes question their utility because many great discoveries were unnoticed by contemporaries and therefore not cited". In this sense, experts in research assessment frequently face comments from researchers claiming that their publications "need more time" to become properly acknowledged. This is the reason why several researchers have studied the problem of delayed recognition and the so-called "sleeping beauties" (van Raan, 2004) showing that although delayed recognition actually does exist, it is not a very frequent phenomena in scientific publishing, thus being more a myth than a real problem (Glänzel and Garfield, 2004). Nevertheless, it is still a topic that challenges researchers in scientific communication nowadays (Wang et al., 2012).

The interest of studying the durability of publications is important for the practice of research evaluation, and in the words of Hook (2002) the "identification and dissection of the factors that contribute to "delay" are not only of interest to scientists, historians, philosophers, and sociologists. Their recognition may also lead to useful scientific and personal practices and be of value to those making science and technology policy". From a critical perspective, Stent (1972) cited by Hook (2002), considers that a discovery can be considered as premature if it cannot be connected by a series of simple logical steps to canonical knowledge of the time and this disconnectedness is the reason why it is not appreciated by the relevant practitioners in the field at the time it is presented. In the views of this author it is even appropriate that the scientific community ignores (if not actually rejects) work that is premature, until it can be properly connected. In this view, delayed recognition is somehow the necessary price that both scientists as well as society must pay at the time to prevent being overwhelmed by attention to perhaps false and useless leads. Following Garfield and Malin (1968) and Costas et al. (2011) it can be suggested that situations of severe patterns of delayed recognition could be also linked to the own fault of researchers as they are not able to communicate their ideas in a proper way.

Recently a new methodology developed by Costas et al. (2010b) for the analysis of durability of scientific publications introduced a flexible tool for the analysis of the aging of publications. This

* Corresponding author.
E-mail address: rcostas@cwts.leidenuniv.nl (R. Costas).

0048-7333/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.respol.2012.11.006

methodology uses a classification of all papers in three general types of durability ("Flash-in-the-pan", "Normal" and "Delayed" papers, their definition is given in Section 3.3). Thus it provides a response to the claim of Garfield for a "handy yardstick" to measure the durability of scientific publications. It also allows a more flexible and systematic identification of citation patterns related to delayed recognition. This new methodology has been already used to test the existence of the so-called "Mendel syndrome" (Garfield, 1979; van Raan, 2004) in the analysis of individual scientists (Costas et al., 2011). The results showed that the potential cases of "Mendelism" are rare and that enlarging the citation windows does not necessary imply a significant improvement in the assessment of these researchers.

Building up on these previous developments in the analysis of *individual scientists*, we focus in this current study on the analysis of *research groups* in the field of chemistry. Research groups can be considered as the basic unit of the research system and their analysis is very common in bibliometric literature (Nederhof and van Raan, 1993; Bordons et al., 1995; Rey-Rocha et al., 2002; Calero et al., 2006). The analysis of the durability of the production at the level of research groups has never been analyzed before. This is important particularly in the light of the comparison of durability indicators with qualitative indicators, because it could provide new insights into the effects of literature obsolescence in the assessment of research performance.

2. Objectives

The main objective of this paper is to combine the analysis of various indicators (both quantitative and qualitative) used in research evaluation at the group level, and to study their relationship with the three types of durability mentioned in the foregoing section.

The main research questions that we want to answer are: can the assessment of research groups be significantly affected by the different durability types observed in their output? Does international collaboration have any relationship with the durability of scientific publications of research groups? Can experts in peer review assessment panels be able to somehow perceive the durability of the publications of the research groups that they are assessing?

3. Data and methods

In this paper outcomes are presented from a study of publication output and international impact of academic chemistry researchers in the Netherlands. The study was performed on behalf of the International Review Committee on Chemistry in the Netherlands (VSNU, 2002). This Committee was established in 2001 by the Association of Universities in the Netherlands (VSNU) for a quality assessment of academic chemistry research (van Raan, 1996). Ten universities were involved in this research assessment procedure: Radboud University Nijmegen, Leiden University, University of Groningen, Delft University of Technology, Eindhoven University of Technology, Twente University, Utrecht University, University of Amsterdam, Vrije Universiteit Amsterdam, and Wageningen University Research Center.

The period of analysis is 1991–2000 for source publications. Their citation impact has been collected for the same period (1991–2000) and also an additional period of citations has been considered: 1991–2008.

The study is based on 18,160 papers in chemistry covered by the Web of Science (WoS). These papers were published by 600 senior researchers, who were associated with chemistry research programmes on December 31, 2000. The names of the senior scientists were provided by VSNU. The researchers were aggregated into about 158 research groups. For each group the full time staff members were selected.

In a first step, for each senior scientist all relevant publications from 1991 to 2000 were extracted from our Web of Science based publication data system. This includes all publications listing the researcher either as first author or as co-author. In a verification-round, researchers were asked to verify whether publication lists were correct and complete. We also performed a test ourselves, aimed at identifying and deleting publications authored by other scientists having similar names. As a result, we are confident that we obtained a highly valid publication data for all chemistry groups in this study.

In the following paragraphs the different sets of indicators used in the analysis are described.

3.1. Qualitative indicators of assessment (VSNU, 2002)

In the first place we describe the set of indicators resulting from the review of Dutch chemical research, these are the qualitative indicators provided by the Review Committee.[1] An assessment of each of the following aspects was required for each research programme and group. These aspects of the evaluation procedure are discussed in more detail here, to provide insight into the working method of the committee, and the level of detail of the decision making. It should be noted that the members of the Review Committee were asked to draw up a preliminary conclusion on the basis of the self evaluation report before the first meeting, and the bibliometric report was handed over just before the first meeting of the Review Committee The report describing the evaluation process and the outcomes clearly states that "In view of some restrictions of the method of bibliometric analysis, the Committee based its assessments primarily on the self-evaluations provided by the Faculties, on the site visits and on the Committee Members knowledge of the field. The bibliometric results were consulted to check the outcome of that process; only in cases of unresolved disagreement, experts more familiar with the specific area were asked for additional comment. However, it should be pointed out that in the vast majority of programmes the correlation between the two types of assessment was good" (page 16 of the Report, VSNU, 2002).

3.1.1. Quality

Academic quality is based on the quality of the output of the research group: dissertations, academic publications, professional publications (where relevant), patents (where relevant), other academic products (tests, prototypes, software). Scores were from 1 (low quality) to 5 (excellent). More precisely, a score of '5' means that according to the review committee the group belongs to the top 5% in the world.

Aspects of the assessment include academic level of the publications, with respect to publication media (e.g., journal status), originality and coherence of the research, and contribution to the development of the discipline or area. Due regard is given to the international standing of (the members of) a research group in assessing the quality of its achievements. Note is taken of participation in international cooperative projects, membership of editorial boards of international journals, academic awards, invitations to international conferences, visiting professorships, research funding acquired from NWO, the Dutch national research council.

[1] The aim of the VSNU procedure was ambitious: evaluation within the next five years of all main disciplines (e.g., physics, chemistry, biology, psychology, sociology, linguistics, in total about 25 major disciplines) in all thirteen Dutch universities. Also, a certain 'foresight' element was included: an assessment of each group in terms of its 'long term viability' (van Raan, 1996).

3.1.2. Productivity

The committee assesses academic productivity by relating the output (the number of publications in total and in each category) to the input of human resources. The committee has used a uniform frame of reference for all disciplines. In order to do justice to those groups with missions supplementary to strictly 'curiosity-driven' research, due attention should also be given to other forms of academic output. In the comment accompanying the assessment, the review committee therefore compared the academic quality and productivity with the objectives or 'mission' of the research programme as submitted by the research group themselves through the self-evaluation report written for the review process.

3.1.3. Relevance

In this evaluation element the following questions were considered by the committee: what significance does the research have for the development of the academic field? Are the issues and the approaches chosen with insight, given the international situation of scholarship in the discipline concerned? In academic fields with a strong strategic/applied research background the issue of academic relevance is inseparable from that of societal/technological impact. As in the case of productivity, here too the context should be taken into account and the relevance should be specified in the description of the group's mission. The minimum requirements for these missions will be a basic academic framework (is the work of a kind that can be expected from an academic group?). Furthermore, missions could include: a contribution to the front lines of science; support for academic areas; support for applied sciences. Here again, the committee's view of the relevance of the research are set against the research group's own viewpoint as expressed in the research group's mission.

3.1.4. Viability

The assessment must also take into account the direction in which the research programme is developing. The committee commented on the viability of the issues chosen by the group and the research approach in the international academic arena. The cohesion of the programme elements should also be a part of the assessment of academic viability.

3.2. Standard bibliometric indicators

In this section we discuss the main bibliometric indicators calculated for this study. The first indicator gives the total number of papers published by the research group during the entire period (***P***). We considered only papers classified as *normal articles*, *letters*, *notes*, and *reviews* (from 1996 onwards, notes are no longer used as a separate document-type, and in general notes are treated as normal articles). The second indicator is the number of citations received, ***C*** (excluding self-citations).

Next, two international reference values are computed. A first value represents the mean citation rate of the *journals* in which the research group has published (***JCSm,*** the mean **J**ournal **C**itation **S**core). The second value relates to the *fields* in which the research group has published (***FCSm***, the mean **F**ield **C**itation **S**core). Our definition of subfields is based on the classification of scientific journals into *categories* developed by ISI. Although this classification is certainly not perfect, it is at present the only classification readily available to us in the WoS context, fitting the multidisciplinary nature of the ISI citation indexes. Both the ***JCSm*** and ***FCSm*** take into account the type of paper (e.g., normal article, review, and so on), as well as the specific years in which the research group's papers were published. For example, with respect to the calculation of ***FCSm***, the number of citations received during the period 1991–2000 by a *letter* published by a research group in 1991 in field X is compared to the average number of citations received during the same period (1991–2000) by all *letters* published in the same field (X) in the same year (1991). Generally, a research group publishes its papers in several fields rather than in one. Therefore, we calculated a weighted average ***FCS*** indicated as ***FCSm***, with the weights determined by the number of papers published in each field. Self-citations are excluded from the computation of ***FCSm***. When a journal is classified in multiple subfields, citation scores are computed according to their number of field assignments. Basically, a paper in a journal classified in N subfields is counted as $1/N$ paper in each subfield, and so are its citations and ***FCSm*** scores.

On the basis of the above international reference values, two 'normalized' impact indicators are calculated. First is the indicator ***CPP/JCSm***. This indicator compares the average number of citations to the output of a university (***CPP***) to the journal mean citation scores ***JCSm***, by calculating the ratio for both. Self-citations are excluded in the calculation of the ratio ***CPP/JCSm*** to prevent that ratios are affected by divergent self-citation behavior. Next, in calculating ***CPP/FCSm***, the average number of citations to the output of a university (***CPP***) is compared to the field mean citation scores ***FCSm***, by calculating the ratio for both. Again, self-citations are excluded in the calculation of the ratio ***CPP/FCSm***. If the ratio ***CPP/FCSm*** is above (below) 1.0, it means that the output of the research group is cited more (less) frequently than an 'average' publication in the subfield(s) in which the research group is active. Thus ***FCSm*** constitutes a *world subfield average* in a specific (combination of) subfield(s). In this way, one obtains an indication of the international position of a research group, in terms of its impact compared to a 'world' average. This 'world' average is calculated for the total population of articles published in WoS indexed journals assigned to a particular subfield or journal category. As a rule, about 80 percent of these papers are authored by scientists from the United States, Canada, Western Europe, Australia and Japan. Therefore, this 'world' average is dominated by the Western world.

Another important international reference value is ***JCSm/FCSm***, an indicator for scientific status of the publication journals. If this indicator is above (below) 1.0, the mean citation score of the journal set in which the research group has published exceeds the mean citation score of all papers published in the subfield(s) to which the journals belong. In this case, one can conclude that the research group publishes in journals with a relatively high (low) impact.

3.3. Indicators of durability

A general methodology for the classification of the durability of scientific papers has been used (Costas et al., 2010b). This methodology aims to classify documents according to their citation histories in three general durability types:

- *Normal type*: these are the documents with the typical distribution in their citations over time according to their fields.
- *Flash-in-the-pan type*[2]: documents that tend to receive citations soon after their publication but they are not cited in the long term.
- *Delayed type*: documents that receive the main part of their citations later than normal documents.

The methodology of classification of papers by durability takes into consideration the distribution of scientific publications by ISI Subject Categories (i.e., the 'fields' as defined in the WoS), for more details regarding the methodology and the multi-assignation of papers we refer to Costas et al. (2010b). It is important to bear in mind that as a final result all cited documents are classified in one durability type, making it possible to calculate the percentages of

[2] This concept was suggested by Garfield and Malin (1968) and Zuckerman and Miller (1980), and it was also described by van Dalen and Henkens (2005).

delayed, normal and flash in the pan papers of each research group. In this sense, it is important to take into account that our approach is more targeted to the study of collections ('oeuvres') of publications and not very much to the detection of very delayed publications (or delayed breakthroughs) as this is more the role of the 'sleeping beauties' methodology previously described by van Raan (2004).

3.4. Indicators of concentration as measures of the multidisciplinary nature of research

The study of the multidisciplinarity of scientific publications is a further challenging aspect of bibliometric analysis (Morillo et al., 2001; Porter et al., 2007; Porter and Rafols, 2009; Rinia et al., 2001) as it presents important problems in its definition, delimitation and calculation. In this paper a similar approach as taken by Morillo et al. (2001) has been used, based on the number of different ISI Subject Categories (fields) in which the journals used for publishing the papers of the research groups are distributed (in relation to the total number of fields). The Pratt's and Gini's concentration indexes (Pratt, 1977) have been calculated for the publications of every research group. These indicators provide a measure of the concentration of papers in fields, ranging between 0 and 1. Values closer to 1 mean that the groups have a higher degree of concentration. Values closer to 0 mean that the groups have a high degree of dispersion of papers across research fields (that the groups have a more multidisciplinary character). Finally, the percentage of papers in the three main fields of each group (%Top3 fields) was also considered as a measure of concentration of papers in the main disciplines of activity of the groups.

3.5. Statistical tests

Several statistical tests have been used to determine the significant differences across the different groups analyzed in this paper. The main indicators that have been tested are CPP, CPP/FCSm, CPP/JCSm and JCSm/FCSm.

The tests used are non-parametric and as such they do not have the stringent assumptions as for parametric techniques. They are also more suitable for smaller samples (Pallant, 2007). However, non-parametric statistics tend to be less sensitive than parametric statistics and are more likely to fail to detect differences between groups that actually exist. Besides, the use of statistical tests is also not free from criticism and limitations (Schneider, 2011) and the best advice is to use them with care and not as a normative element of 'truth' but only as indications where the sampling error is lower.

The main test used in this paper is the *U*-Mann–Whitney test, which is a technique used to test for differences between two independent groups on a continuous measure. Instead of comparing means of the two groups, the *U*-Mann–Whitney test actually compares medians. It converts the scores on the continuous variable to ranks, across the two groups. It then evaluates whether the ranks for the two groups differ significantly. As the scores are converted to ranks, the actual distribution of the scores does not matter. We have considered differences as significant when $p < 0.05$.

4. Results

4.1. Relationships among all indicators

In a first step the relationships among all the different indicators used for the analysis of the research groups have been studied through Factor Analysis. The results of this analysis are presented in Table 1. Six main dimensions are obtained explaining 81% of the total variance.

Table 1
Factor analysis of main indicators at group level.

Rotated Component Matrix[a]

	Dimensions					
	1	2	3	4	5	6
N fields	.934	.043	−.094	−.112	−.088	.086
P	.861	.318	.219	.061	−.031	−.040
C	.706	.355	.187	.306	.174	−.058
Full time personnel (FTE)	.597	.217	.235	.198	−.087	.095
Quality	.113	.845	.016	.225	.038	.002
Relevance	.180	.822	.092	.077	−.012	.144
Viability	.197	.808	.029	.084	.192	.154
Productivity	.426	.478	−.116	−.016	.060	−.098
Pratt index	.389	.034	.881	.031	.038	.035
%Top3 fields	−.438	.013	.835	.064	.085	−.069
Gini index	.540	.034	.792	.009	.020	.059
CPP/JCSm	.114	.097	.086	.924	−.021	.073
CPP/FCSm	.055	.191	.003	.920	.290	.109
% Delayed papers	.061	.009	−.107	−.032	−.854	.268
JCSm/FCSm	−.077	.240	−.105	.421	.614	.094
% Flash in the pan	.015	−.096	.061	−.138	.247	−.914
% Normal papers	.139	.239	.215	.103	.600	.656

Extraction method: Principal Component Analysis.
Rotation method: Varimax with Kaiser Normalization.
[a] Rotation converged in 7 iterations.

- *Dimension 2* is clearly linked to the qualitative indicators provided by the review committee (Quality, Relevance, Viability and – to a lower extent – Productivity). Remarkably the indicator 'Productivity' is the one with the lowest loadings in all dimensions. This dimension explains 15% of total variance.
- *Dimension 3* is related with the multidisciplinarity and concentration measures, proving that the three measures suggested (Gini, Pratt, and % Top 3 Fields) are closely related. This dimension explains 13% of all total variance
- *Dimension 4* is the factor where we see the correlation of the field normalized relative impact indicators. Particularly CPP/FCSm and CPP/JCSm are closely related, and to a lower extent also related with the journal relative impact indicator JCSm/FCSm. Of the total variance 9% is explained in this dimension.
- In *dimensions 5 and 6* we find the three indicators of durability. In dimension 5 we see how the percentage of delayed papers is negatively correlated with the quality of the journals (JCSm/FCSm). This can be understood as delayed papers published more frequently in journals of lower impact (this was already highlighted in Costas et al., 2010b). Next, the percentage of normal papers is positively correlated with the JCSm/FCSm and negatively correlated with the percentage of delayed papers. In other words, normal papers are published in better journals, and the higher their percentage, the lower the percentage of delayed papers. The same idea goes for dimension 6, where the higher percentage of flash in the pan publications correlates negatively with the share of normal publications.

4.2. Research groups by durability

In this analysis we investigate to what extent the different durability type of the chemistry groups can influence a bibliometric research performance analysis. The trends in the scores of CPP/FCSm considering the three different types of durability and the two citation periods are shown in Fig. 1. The two periods of citations (1991–2000 and 1991–2008) are used to determine the evolution and change of the three durability types from the first period to the second.

We observe that delayed publications of the research groups clearly improve in their CPP/FCSm from one period to another, while flash-in-the-pan publications decrease in their relative

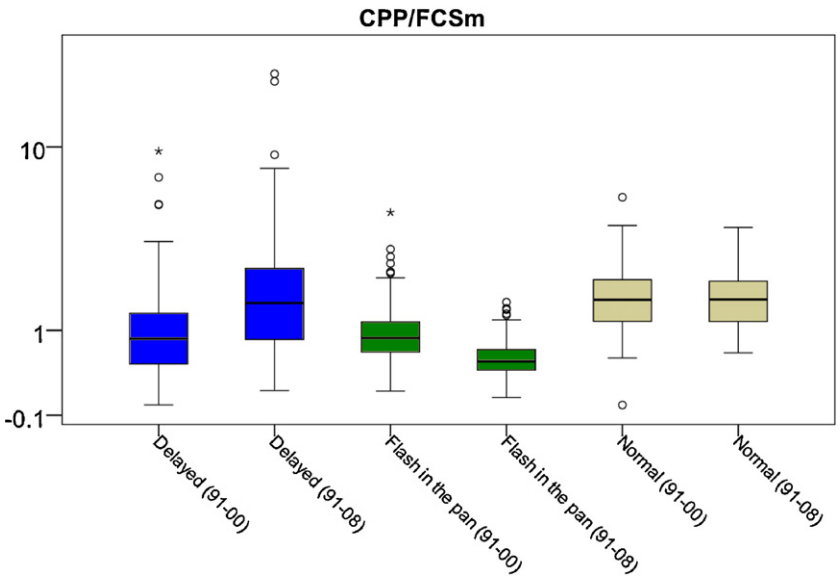

Fig. 1. Groups by durability type and different citations periods.

impact from the shortest citation window to the largest. Normal documents stay at a similar level of CPP/FCSm.

It is remarkable that in the citation period 1991–2000 delayed and flash-in-the pan papers have a similar level of impact (no significant differences were found), while in the longer period (1991–2008) delayed papers have clearly outperformed the flashes-in-the-pan. Furthermore, in the period 1991–2000 normal papers have the highest impact compared to the other types ($p < 0.000$), while in the longer period delayed and normal papers have a similar level of CPP/FCSm.

These results indicate that there is a higher impact performance of delayed outputs when the citation window is enlarged, while the contrary happens for flashes-in-the-pan. Regarding these results, one may argue that if a research group has a considerable number of delayed publications and not many flash-in-the-pan publications, the assessment of its performance could be prejudiced by the use of standard bibliometric indicators with relatively short citation windows. This possibility is tested in the following analysis.

Are there groups with significantly different levels of delayed and/or flash-in-the-pan in their outputs? This may significantly affect their assessment. To find this out, we performed a *k*-means cluster analysis[3] in order to classify groups by their durability types (the same method was applied in Costas et al. (2011) for the classification of individual researchers, so we refer for the discussion of the methodology to this paper).

A final 4-cluster solution was obtained. In general in all clusters normal papers are the majority, but the clusters are named to the durability type that is next to the normal papers, except when delayed papers and flashes-in-the-pan are below a certain percentage threshold, in this case the cluster is 'just' normal. We add "+" to the durability that marks this difference. Thus, we have a first cluster including 51 groups (33%) – labeled "+Delayed", a second one with 42 (27%) – "+Flash in the pan", a third one with 59 (38%) – "+Normal" – being this the biggest cluster, and finally a fourth one

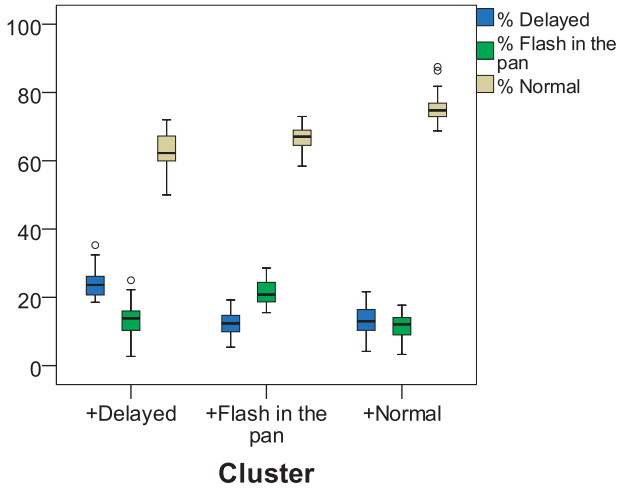

Fig. 2. Distribution of percentages by durability types and clusters of research groups.

including only 2 groups (1% being not included in the subsequent figures and analysis). The distribution by percentages of durability types of the three main clusters obtained is shown in Fig. 2.[4]

Fig. 2 presents three different patterns for the three clusters previously obtained through the *k*-means analysis.

A relevant characteristic of these clusters is that groups classified in the "+Normal" cluster present a level of around 80% of normal papers, while the other two cluster have shares of normal papers between 60% and 70%, a measure that supports the "rule of thumb" previously suggested in Costas et al. (2011) that those cases with less than 60–70% of normal publications could be considered as potential candidates of suffering from the "Mendel syndrome".

Once these three clusters have been detected, we want to know how our standard bibliometric indicators assess their performance. The analysis of the three main indicators of normalized impact (CPP/JCSm, CPP/FCSm and JCSm/FCSm) for the period 1991–2000 is presented in Fig. 3. The idea is to see how the groups would be

[3] The *k*-means algorithm (MacQueen, 1967) is one of the simplest and most widely applied non-hierarchical clustering techniques (Kaufaman and Rousseeuw, 1999). The algorithm focuses on partitioning a population into k sets. This process gives partitions (clusters) which are reasonably efficient in the sense of within-class variance. In this paper, we use this algorithm as it is implemented in SPSS 17.0 with the purpose of detecting groups (clusters) of authors who are different in their percentages of Normal, Delayed and Flash in the pan publications.

[4] Four groups were excluded from this particular analysis as they had zero values in any of the three durability types.

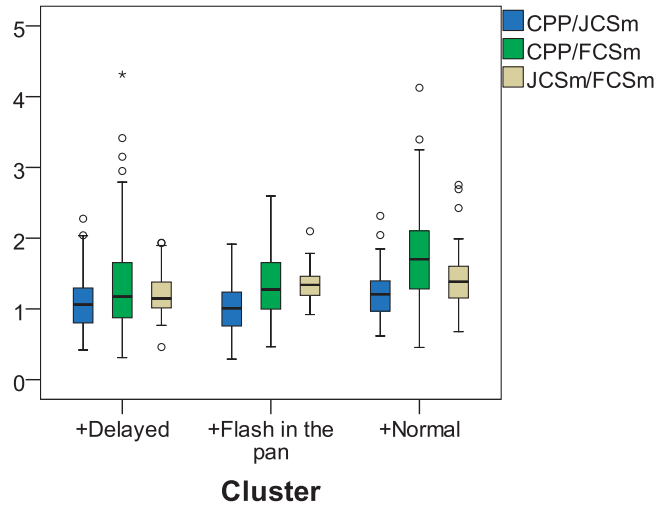

Fig. 3. Relative impact indicators by clusters of durability.

bibliometrically valued considering their indicators at the time of the evaluation (in 2001).

As is shown in Fig. 3, "+Normal" groups present the highest scores in most of the indicators. Between the "+Delayed" and "+Flash-in-the-pan" groups we find only statistical significant differences for the JCSm/FCSm ($p < 0.000$), showing that "+Flash-in-the-pan" groups have published in better journals than "+Delayed" groups.

"+Flash-in-the-pan" groups present lower levels of CPP/FCSm and CPP/JCSm than "Normal" groups (*U*-Mann–Whitney test $p < 0.05$) but not in JCSm/FCSm, which shows that flash-in-the-pan papers are published in journals of the same impact level as normal papers.

According to these results, it can be confirmed that there are some differences in the performance of the groups depending on their durability cluster. The next step is to analyze whether the CPP/FCSm values of these three clusters is affected by the enlargement of the citation period. In other words, do the groups benefit or are they negatively affected when the citation window is enlarged? In Fig. 4 this question is analyzed by comparing the CPP/FCSm of all documents for the three clusters of groups, this time taking into account the two periods of citations.

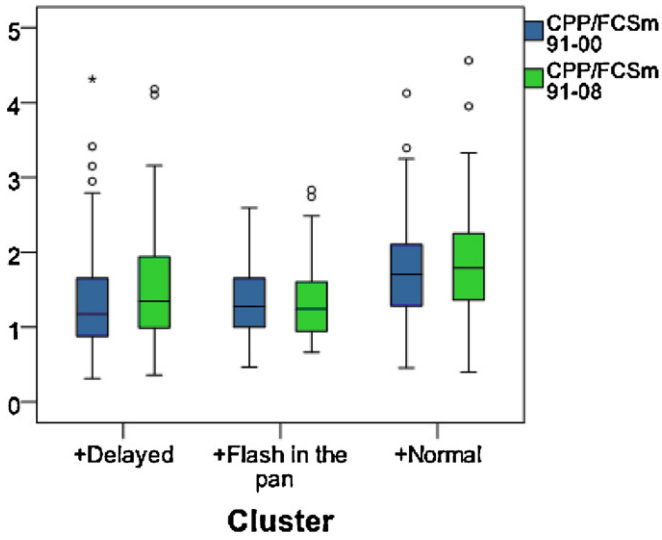

Fig. 4. Distribution of CPP/FCSm values by cluster (periods 1991–2000 and 1991–2008).

The comparison within each of the three clusters shows that from one period to another the "+Delayed" cluster is the one that improves the most, meaning that groups with more delayed papers will benefit by an increase in length of the citation period. On the other hand, groups in the "+Flash-in-the-pan" cluster slightly decrease in the CPP/FCSm of their publications while "+Normal" groups slightly improve.

The comparison across clusters shows that in both citation periods the "+Normal" groups present the highest scores (*U*-Mann–Whitney test $p < 0.05$). This means that those groups with a high level of 'normal paper production' tend to perform the strongest, regardless the citation window employed. The second important aspect is the comparison between "+Delayed" and "+Flash-in-the-pan", in this case we see inverse patterns depending on the period. Although there are not statistically significant differences, it is remarkable the pattern that with the shortest period of citations the "+Flash-in-the-pan" groups are those that have a slightly higher CPP/FCSm, while in the longer period the "+Delayed" groups are the ones that present slightly higher scores. In a way, it can be assumed that the enlargement of the citation window improves the impact of the "+Delayed" groups and decreases that of "+Flash-in-the-pan" groups. These results suggest that durability of publications does have an effect on the performance assessment of groups, but this influence is generally rather small.

4.3. Qualitative indicators vs. durability types

The distribution of the percentages of the three types of durability is studied in contrast with the *qualitative* indicators provided by the review committee (Fig. 5), the main values can be seen in Appendix I. It is important to realize that the scores for Productivity and Viability were ranging from 2 to 5, while the scores for Quality and Relevance were ranging between 3 and 5. Reasons for a score of 2 for Viability could be the relative low number of FTE's available for a research group, while for Productivity the number of publications coming out of a group was considered as 'on average too low'.

For the indicators of Quality, Relevance and Viability there is an increasing trend in the share of normal papers with the higher scores of the committee, while a decreasing trend is observed for the other two types of durability. For the indicator of Productivity no clear differences and no clear pattern can be mentioned. In any case, these results support the idea that quality in research is mainly linked to production with a normal durability character, thus supporting the idea that quality research is published in articles that are assimilated and cited by the contemporary colleagues in a regular period of time.

Finally, the three previous *clusters* of durability have also been studied considering the four qualitative indicators (Fig. 6).

Although no statistical significance has been found (only between "+Flash-in-the-pan" and "+Normal" clusters and the qualitative indicator of Relevance, $p < 0.05$), there are several interesting patterns that deserve some comments. In the first place, the most remarkable and clear pattern is that "+Normal" groups present the highest scores in three qualitative indicators (with the only exception of Productivity). This again supports the idea that "normal" publications (i.e. publications that are cited within a 'normal' period of time) are the type of publication that is more appreciated also on the basis of qualitative assessments by peers.

Regarding the Productivity indicator, "+Normal" groups show the lowest scores while "+Delayed" groups the highest. This suggests that groups with higher shares of delayed papers might also produce other types of research output not covered by the WoS, for instance document types such as books, theses, etc., thus beyond the horizon of the journal-output based bibliometric analysis, but positively assessed by the experts.

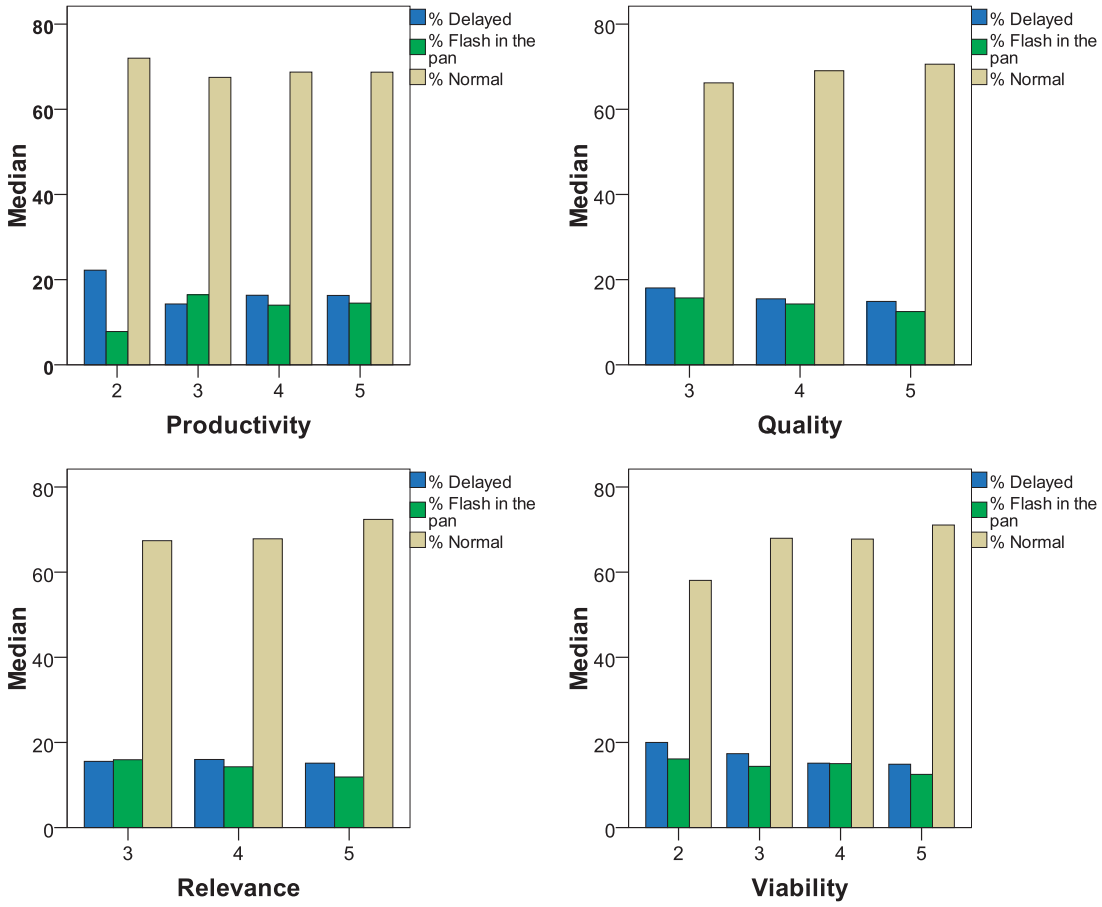

Fig. 5. Distribution of durability types of research groups by qualitative indicators.

With respect to the Quality indicator it is interesting to remark that the peer committee assessed "+Delayed" groups as better compared to "+Flash-in-the-pan" groups. The Relevance indicator presents a similar although more clear pattern as compared to the previous one, with "+Delayed" groups presenting a higher degree of relevance according to the peer committee as compared to "+Flash-in-the-pan" groups.

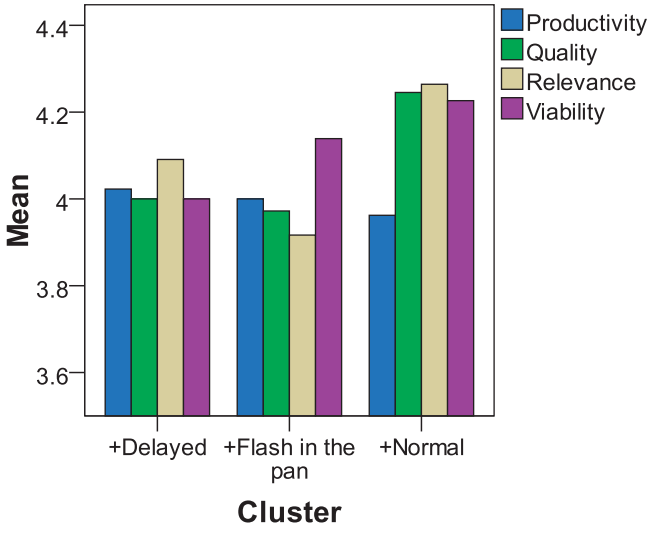

Fig. 6. Distribution of qualitative scores by durability cluster.

Finally, the Viability indicator of the clusters presents an interesting increasing pattern from "+Delayed" groups to "+Normal" groups, having "+Flash-in-the-pan" groups higher scores in this indicators as compared to "+Delayed". This may indicate that "+Delayed" groups are regarded as initially less viable in the direction of their research, and therefore they scored lower in this aspect as compared to the other two clusters, while "+Flash-in-the-pan" groups may have been appreciated as more viable in the research in the short-run, therefore scoring higher in this indicator.

5. Discussion and conclusions

There is a popular belief among scientists that their work can suffer form an important delay in being recognized by their peers (Garfield, 1980). For this reason, on several occasions citation analysis has been rejected as a valid tool for supporting research assessment, as scientists could be prejudiced by bibliometric indicators when too short citation windows are chosen. However, it is important to stress here that the chosen period of ten years is an adequate citation window in chemistry for the study of delayed patterns, the years added function mainly for testing the influence of longer windows on publications that can be characterized as 'delayed'.

More recently the so-called "Mendel syndrome" (van Raan, 2004; Costas et al., 2011) has been proved to be very unlikely in affecting the bibliometrically measured performance at the individual level. Moreover, the cases where we identified a 'Mendel syndrome' do not significantly improve their performance even when longer periods of citations are applied.

In this paper the same idea has been tested but this time at the level of the *research groups* and with a longer additional citation window (8 extra years more). The results confirm that even with such a long citation window, the phenomenon of delayed publications has only a minor effect on the overall performance of the research groups. In fact, it can not be sustained that at the group level, enlarging the citation window would improve the performance assessment of those groups with more delayed publications. Furthermore, the analysis of the results confirms that publications with a normal durability character are those that are more cited but also those more valued by peer review assessments, suggesting a similar conclusion as in Costas et al. (2011).

5.1. Durability measures and other bibliometric and qualitative indicators

Three of the five dimensions obtained in the factor analysis of the relations between all the different (qualitative and quantitative) indicators roughly correspond with the three dimensions obtained by Costas et al. (2010a). This finding reinforces the idea of the existence of a first dimension of size-dependent indicators (total number of publications and citations, total number of fields and total number of FTE's). A second dimension relates with the average impact of publications (including CPP/FCSm and CPP/JCSm), and a third dimension with the average impact of the journals of publication (JCSm/FCSm). Interestingly, we found that this latter indicator correlates negatively with the percentage of delayed publications. The other three new dimensions obtained in this study correspond to (1) the dimension covering all four qualitative (peer) indicators (Quality, Relevance, Viability and Productivity); (2) the dimension of concentration/dispersion of fields across fields (Pratt index, %Top3 fields and Gini index); and finally (3) the dimension where we find the inverse correlation of the percentage of flash-in-the-pan and normal publications.

It is remarkable that in general the distribution of durability types across publications is relatively independent of the quantitative (bibliometric) and qualitative (peer judgments) indicators. However, the factor analysis suggests that the share of delayed papers is negatively correlated with publications in high-impact journals, as the correlation between the percentage of Delayed papers and JCSm/FCSm is high, but negative. This aspect of publication of delayed papers in journals of lower impact and visibility is an element that was also suggested by Costas et al. (2010b). This finding can be connected to the idea that publications in non-core journals need more time to be 'detected' and cited by the scientific community.

5.2. Research groups by durability measures

In general, the production of research groups with a delayed pattern tends to increase with the enlargement of the citation window, while a contrary pattern (a decrease in the field-normalized impact) is found for to the flash-in-the-pan 'production'. This is in line with the findings at the individual level (Costas et al., 2011) that the field-normalized impact for normal publications does not improve very much when enlarging the citation window.

Three clusters based on the classification by durability of scientific publications have been established, being this the same distribution of clusters also established at the individual level (Costas et al., 2011), with a cluster of research groups with more normal publications, another cluster with distinctly proportionally more flash-in-the-pan papers, and a third group with proportionally more delayed publications.

A more in-depth analysis of the clusters based on durability shows that research groups that have proportionally more normal publications (i.e. the cluster "+Normal") tend to perform the best in the field-normalized bibliometric indicators, regardless of the period of citation impact measurement. In other words, research groups with a production of papers that follows a standard obsolescence pattern in citations (standard for the field under study, chemistry) tend to be more cited as compared to groups with publications with different obsolescence patterns. Research groups with proportionally higher levels of delayed publications tend to *increase* their field-normalized impact with the enlargement of the citation window. But this is not enough to outperform the groups with a higher share of normal publications, and only just enough to level off those groups with more flashes-in-the-pan. In other words, waiting (in terms of the citation window) for these groups results in a better performance but not sufficient to significantly change their position as compared to the other research groups.

5.3. Qualitative indicators vs. durability measures

Considering the qualitative indicators in relation to durability types, it is important to note a tendency in which research groups with the highest qualitative scores are also the groups with the highest levels of normal papers. This may indicate that "normal science" is also positively valued from point of view of the experts in the committee, in line with the quantitative findings. On the other hand, groups with a higher degree of delayed publications were considered by the experts being more productive. This suggests that these groups may have also important production in other channels of communication (e.g. books, book chapters, etc.) that is not covered by the WoS but known to the review committee through the self-evaluation reports produced by the groups. Besides, the peer review committee assessed that these groups presented a higher quality and relevance as compared to those groups with more flash-in-the-pan production, but a lower viability. This implies that the peer review committee positively assessed the potential interest of the research lines of groups with a higher degree of delayed publications but considered their work to be less viable. This may suggest that the near-future expectations of these groups, although regarded as relevant, were considered less positively by the peers. Finally, those groups with a more flash-in-the-pan pattern were the lowest in nearly all qualitative indicators with, remarkably, the exception of viability. This suggests that the peer committee considered the line of research of these groups to be not very relevant yet but it still considered as viable.

Our results suggest a consistent relationship between peer review assessments and the durability of scientific publications, in the sense that for the experts of the committee the normal production was the part of the output that was most appreciated. The differences in the scores given by the experts between flash-in-the-pan and delayed production also suggest a certain consistency, with delayed groups regarded as relevant but less viable, and flash-in-the-pan as less relevant but more viable.

5.4. Final conclusions and implications of the study

In general, we can conclude that the different types of durability of publications have almost no effect on the assessment of the performance of research groups. It appears that both by bibliometric indicators and by peer review assessments, research with a standard durability is valued and rewarded more positively. This supports the statement made by Small (1998) that the right idea at the right time is incorporated into science, while the right idea at the wrong time is not. This reinforces the suggestion that it is important to communicate new results and important conclusions in a way that they can be understood and assimilated by contemporary colleagues.

Our study also shows the consistency of peer review assessments with bibliometric indicators and especially with indicators

of durability, in the sense that qualitative indicators also tend to reward "normal" science. In other words, peers judge research positively if the potential impact of the publications can be expected in a regular period of time. This indicates that the assessment with bibliometric indicators based on a citation window of 'normal' length will correlate well the qualitative assessment. However, it is remarkable that the peer review was also able to detect (on the basis of the qualitative indicators) several relevant differences which may be related with the durability of the publications of the research groups.

In sum, the main finding of this study that (1) the effects of the different types of durability are not very relevant at the group level, and that (2) the selection of an adequate citation window (e.g., 3–5 years) (Vlachy, 1985; Glänzel et al., 2003; Costas et al., 2011) is in most cases sufficient to provide reliable indicators and fair assessments that correlate well with peer review based qualitative assessments. Of course, sleeping beauties and delayed publications can still appear. But if they do not represent an important share of the production of the unit of analysis (e.g. more than 30%, which is not very likely), we can assume that in general their effect will be not very influential. Nevertheless, combination of durability indicators with peer review assessments can detect potential cases of groups with certain level of delayed patterns that would require a more thorough assessment.

We would like to conclude with a policy relevant recommendation. An improvement in the application of bibliometric tools in research assessment procedures could consist of the inclusion of an indicator of the degree of 'normal durability' publications, in order to alert peers in a review committee for the possible differences between groups on this aspect. This could then be considered a possible reason for more in-depth scrutiny. Overall, this could improve informed peer review, by focusing on yet another dimension of scientific publishing, and the way scientific results are communicated *and* perceived by the scientific environment.

References

Bordons, M., Zulueta, M.A., Cabrero, A., Barrigón, S., 1995. Research performance at the micro level: analysis of structure and dynamics of pharmacological research teams. Research Evaluation 5 (2), 137–142.

Calero, C., Buter, R., Cabello-Valdés, C., Noyons, E., 2006. How to identify research groups using publication analysis: an example in the field of nanotechnology. Scientometrics 66 (2), 365–376.

Cole, S., 1970. Professional standing and the reception of scientific discoveries. American Journal of Sociology 76 (2), 286–306.

Costas, R., van Leeuwen, T.N., Bordons, M., 2010a. Self-citations at the meso and individual levels: effects of different calculation methods. Scientometrics 82 (3), 517–537.

Costas, R., van Leeuwen, T.N., van Raan, A.F.J., 2010b. 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, 177–205.

Garfield, E., Malin, M.V., 1968. Can Nobel Prize winners be predicted? 135th Annual Meeting, American Association for the Advancement of Science, Texas. http://www.garfield.library.upenn.edu/papers/nobelpredicted.pdf (accessed: 07.04.12).

Garfield, E., 1970. Would Mendel's work have been ignored if the Science Citation Index was available 100 years ago? Current Contents 2, 69–70.

Garfield, E., 1979. Is citation analysis a legitimate evaluation tool? Scientometrics 1 (4), 359–375.

Garfield, E., 1980. Premature discovery or delayed recognition—why? Essays of an Information Scientist 4, 488–493.

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., Garfield, E., 2004. The myth of delayed recognition. The Scientist 18 (11), 8.

Hook, E.B., 2002. Prematurity in Scientific Discovery: On Resistance and Neglect. University of California Press, USA.

Kaufaman, L., Rousseeuw, P.J., 1999. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, NY, p. 113.

MacQueen, J.,1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 281–297.

Morillo, F., Bordons, M., Gomez, I., 2001. An approach to interdisciplinarity through bibliometric indicators. Scientometrics 51 (1), 203–222.

Nederhof, A.J., van Raan, A.F.J., 1993. A bibliometric analysis of six economics research groups: a comparison with peer review. Research Policy 22, 353–368.

Pallant, J., 2007. SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows, third edition. McGraw-Hill, England.

Porter, A.L., Cohen, A.S., Roessner, J.D., Perreault, M., 2007. Measuring research interdisciplinarity. Scientometrics 72 (1), 117–147.

Porter, A.L., Rafols, I., 2009. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics 83 (3), 719–745.

Pratt, A.D., 1977. A measure of class concentration in bibliometrics. Journal of the America Society for Information Science 28 (5), 285–292.

Rey-Rocha, J., Martín-Sempere, M.J., Garzón, B., 2002. Research productivity of scientists in consolidated vs. non-consolidated teams: the case of Spanish university geologists. Scientometrics 55 (1), 137–156.

Rinia, E.J., van Leeuwen, Th.N., Bruins, E.E.W., van Vuren, H.G., van Raan, A.F.J., 2001. Citation delay in interdisciplinary knowledge exchange. Scientometrics 51 (1), 293–309.

Schneider, J.W., 2011. Caveats for Using Statistical Significance Tests in Research Assessments. http://arxiv.org/abs/1112.2516

Small, H., 1998. Citations and consilience in Science. Scientometrics 43 (1), 143–148.

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

van Dalen, H.P., Henkens, K., 2005. Signals in science—on the importance of signaling in gaining attention in Science. Scientometrics 64 (2), 209–233.

van Raan, A.F.J., 1996. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics 36, 397–420.

van Raan, A.F.J., 2004. Sleeping beauties in science. Scientometrics 59 (3), 467–472.

Vlachy, J., 1985. Citation histories of scientific publications. The data sources. Scientometrics 7 (3-6), 505–528.

VSNU, October 2002. Chemistry and Chemical Engineering, Series Assessment of Research Quality. VSNU, Utrecht, 173 pp.

Wang, J., Ma, F., Chen, M., Rao, Y., 2012. Why and how can sleeping beauties be awakened? The Electronic Library 30 (1), 5–18.

Zuckerman, H., Miller, R.B., 1980. Indicators of science: notes and queries. Scientometrics 2 (5–6), 347–353.