Does the durability of scientific literature affect research performance assessment at the group level?

Rodrigo Costas and Thed van Leeuwen
Centre for Science & Technology Studies (CWTS)
Leiden University, the Netherlands

1. Introduction
A crucial element in Science Policy is the effect that the durability of the scientific literature as well as the so-called ‘delayed recognition’ (Garfield, 1980) or ‘scientific prematurity’ (Stent, 1972) can have over the indicators used in the evaluation and development of research policies. 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, for the experts in research assessment it is very common to face comments from researchers claiming that their publications “need more time” for getting properly acknowledged. Some researches have also studied delayed recognition and the so-called “sleeping beauties” (Glanzel et al, 2003; van Raan, 2004) proving that although delayed recognition actually does exist, it is not very frequently occurring phenomenon, and thus more a myth than a reality (Glanzel & Garfield, 2004).

Recently, a new methodology developed by Costas et al (2010b) for the analysis of durability of scientific publications provides a more flexible tool for the analysis of the ageing of publications, considering all papers and based on a classificatory system of three general types of durability (“Flash in the pan”, “Normal” and “Delayed” papers), providing thus a response to the claim of Garfield (1990) for a “handy yardstick” to measure durability and delayed recognition from a broader perspective and contributing to a more flexible identification of citation patterns of delayed recognition.

In this study we focus on a national research performance assessment in the field of chemistry and related fields, initiated by the Association of Universities in the Netherlands (VSNU). We present a comparative analysis of indicators of research performance developed by CWTS (Moed et al, 1995), with qualitative indicators generated by the review committee. The analysis of the durability of the production of research teams is a challenging approach, as well as the contrast of durability indicators with qualitative indicators, which could provide new insights into the phenomenon of obsolescence in the assessment of research performance.

2. Objectives
The main objective 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 suggested. We try to answer the question if the assessment of research groups can be significantly affected by the durability of their papers.

3. Data & Methodology
In this paper outcomes are presented from a past study on the publication output and international impact of academic chemistry researchers in the Netherlands, as part of a quality assessment of academic chemistry research. Using bibliometric techniques, the study assessed the publication output (18,160 papers) and citation impact of senior scientists (n=600) affiliated with chemistry research programs (n=158) at ten universities in the Netherlands. The period of analysis is 1991-2000 for source publications in the Web of Science (WoS). Their citation impact has been collected for the same period (1991-2000) and an extra period of citation impact has been considered: 1991-2008.

In the first place we describe the qualitative indicators provided by the review committee. An assessment of each of the following aspects was required for each research program and group: **Quality** (quality of the output of the research group: dissertations, academic
publications, professional publications), *Productivity* (academic productivity by relating the output to the input of human resources), *Relevance* (what significance does the research have for the development of the academic field?), and *Viability* (some account of the direction in which the research program is developing). These four indicators were rated on a five point scale.

Next we present the main bibliometric indicators used in the results for this abstract. More indicators were calculated, those will be presented in the full paper. 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*. The second indicator is the number of citations received, \(C\), which leads to *CPP*, the average number of citations per publication. Next we present *CPP/FCSm*, the average number of citations to the oeuvre of a group \((CPP)\) is compared to the field mean citation scores *FCSm*, by calculating the ratio for both\(^1\). If the ratio of *CPP/FCSm* is above (below) 1.0, this means that the oeuvre of the research group is cited more (less) frequently than an 'average' publication in the subfield(s) in which the research group is active. Self-citations are excluded in the calculation of the impact indicators mentioned above to prevent that impact scores are affected by divergent self-citation behavior.

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 types: *Normal-type*: these are the documents with the typical distribution in their citations over time; *Flash in the pan-type*: documents that tend to receive citations immediately after their publication but they are not cited in the long term (this concept was suggested by Garfield & Malin (1968) and Zuckerman & Miller (1980), and it was developed and described by van Dalen & Henkens (2005); and finally, *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. As a final result, all documents are classified in one durability type, making it possible to calculate the percentages of delayed, normal and flash in the pan papers of each research group.

### 4. Results of the study

As this abstract offers only a limited space to present findings of the study, we made a selection of the relationship between impact measures and durability indicators, and the relationship between qualitative indicators and durability indicators. In Figure 1, the distribution of *CPP/FCSm* scores by durability types of the chemistry groups in the study is presented. We show the *CPP/FCSm* of the total set of publications aggregated by the three types of durability and for all the research groups under analysis. Statistical significant differences (we applied the Wilcoxon Signed Rank test, the outcome of which is \(p<0.000\)) were observed for the whole set as well as for the *CPP/FCSm* normal, while we do not find statistical significant differences for *CPP/FCSm* delayed and *CPP/FCSm* flash in the pan. As a consequence, it can be said that quantitatively the *CPP/FCSm* scores of groups are negatively affected by delayed and flash in the pan publications as the *CPP/FCSm* total (normal+delayed+flash in the pan) is significantly lower than that if only normal papers were considered. However, these differences are not important enough to change the position of the groups in their ranks according to these indicators (as shown by Pearson’s correlation).
In Figure 2 we present the relationship between qualitative indicators and the three types of durability. Remarkably, we observe differences between the three types of durability, and the peer judgments made by the committee. For the set of groups with proportionally more Delayed publications, the highest peer rate is observed for Relevance, which can be explained by the fact that peers recognize the relevance although the publications of these groups are not among the most immediately visible in the field. The peer judgments on the groups with more Flash in the pan publications show the highest scores on the indicator Viability and the lowest in Relevance, while the groups with more Normal publications display the highest peer scores on Quality, Relevance, and Viability, and a relative low score on Productivity.

References
Although the CPP/FCSm indicator has been subject of a debate in the bibliometric community, the indicator was actively used in the period in which we conducted our study, and the users were familiar with the indicator. Therefore, we chose to work with this indicator instead of the improved version, referred to as the MNCS (mean normalized citation score) indicator, see references below for a further elaboration of this indicator.
