Scientometrics
DOI 10.1007/s11192-014-1368-6

Comparing scientific performance among equals

C. O. S. Sorzano · J. Vargas · G. Caffarena-Fernández · A. Iriarte

Received: 15 November 2013

Abstract Measuring scientific performance is currently a common practice of funding agencies, fellowship evaluations and hiring institutions. However, as has already been recognized by many authors, comparing the performance in different scientific fields is a difficult task due to the different publication and citation patterns observed in each field. In this article, we defend that scientific performance of an individual scientist, laboratory or institution should be analysed within the corresponding context and we provide objective tools to perform this kind of comparative analysis. The usage of the new tools is illustrated by using two control groups, to which several performance measurements are referred: one group being the Physics and Chemistry Nobel laureates from 2007 to 2012, the other group consisting of a list of outstanding scientists affiliated to two different institutions.

Keywords Scientific performance · Relative measurements · Control groups

Introduction

Measuring scientific performance is currently at the heart of many funding calls, job openings and fellowship applications. The rationale behind it is that the limited resources available in science should be allocated to the best performing scientists, so that the overall output is maximized. However, measuring performance is not an easy task, since different scientific aspects can be evaluated and comparison across disciplines is a major issue. There is the general agreement that scientific outcome can be measured by productivity (mostly, number of papers) and by impact (number of citations to those papers).

C. O. S. Sorzano (✉) · J. Vargas
National Center of Biotechnology (CSIC), Darwin 3, Campus de Cantoblanco, 28049 Madrid, Spain
e-mail: coss@cnb.csic.es

C. O. S. Sorzano · G. Caffarena-Fernández · A. Iriarte
Department of Information and Telecommunication Systems, University CEU San Pablo, Urbanización Montepríncipe S/N, Boadilla del Monte, 28668 Madrid, Spain

Published online: 29 July 2014

Additionally, it has been argued that patents and licensed patents should also be used as an indicator for scientific production evaluation. However, not all scientific fields are prone to patenting, as for example theoretical physics or mathematics, and there are significant differences in the way scientific articles and patents are cited.

Measuring performance

Many scientific productivity indicators have been traditionally proposed to evaluate the performance of individuals, laboratories, or institutions. Up to 2005, the indicators were based on simple functions (number of citations, number of citations of the single most highly cited paper) or statistical measurements (average number of citations, average number of papers published per year, median number of citations, relative frequencies and quartiles (Glänzel 2006). One of the main disadvantages of these indicators is that they do not reflect the full impact of scientific research, or that they are disproportionately affected by a single publication of major influence (Panaretos and Malesios 2009). In 2005, Hirsch proposed a different indicator to quantify scientific excellence (Hirsch 2005). Since then, the so-called *h-index* or *Hirsch-index* has succeeded in becoming a used yardstick when assessing the productivity of individual scientists. A scientist has index h when a number h of his publications have at least h citations each. It is thus an indicator easy to compute that combines productivity and impact in a single number. However, it has received a lot of criticism and a number of variants and alternative indices have been proposed to avoid its inconveniences. For example, the *g-index* (highest number of papers of a scientist that received g^2 or more citations) was introduced to avoid the disadvantage of the *h-index* that once a paper is cited more than h times, it does not matter it being cited 100, 1,000 or 10,000 times, nor it continuing to be cited. The *hg-index* (geometric mean of the *h*- and *g*-indices of a researcher) tries to keep a balance between the advantages and disadvantages of the *h*- and *g-indices*; while taking into account the number of citations of the most cited papers of an author, it also moderates the impact that a highly cited paper could have in the *g-index*. Other indices have been proposed that extend the *h-index* to take into account other sensible variables. For example, the *m-quotient* (h-index divided by the number of years since the first publication of the researcher) facilitates the comparison between scientists with different lengths of academic careers. On the other hand, the h_m- and h_i-indices account for the number of co-authors of the publications. The reader is referred to the excellent reviews by Lehman et al. (2008), Moed (2009), Alonso et al. (2009), Bornmann et al. (2008), Harzing (2010) and Panaretos and Malesios (2009) for a full description of most of the numerous possibilities to measure scientific performance.

Performance among equals

As stated in the previous section, a huge quantity of different proposed scientific productivity indicators have been proposed. All of them are sensitive to different aspects of scientific productivity. In this work, we will focus on the fact that none of the above mentioned indicators is adequate to compare the performance between different scientific fields. Whichever the performance index used, an obvious drawback to an "absolute" scale of performance is that different scientific disciplines have different publication and citation patterns. Even the definition of "field" is not well established (two electrical engineers, one working in Radio Frequency devices and another in Image Processing, have very different publication patterns; among engineers working in Image Processing, those working in the pure development of methods and those in their applications also have very different

publication patterns; likewise, among engineers working on the application of Image Processing methods, those working in Neuroimaging present a very different scientific environment compared to those applying those methods to Material Sciences).

This need to compare among equals has already been pointed out by Hirsch himself. How "equal" should be is a matter of discussion, but current bibliographic tools as Scopus, ISI Web of Science, or Google Scholar have very limited capacity to clearly define the reference group to which a certain scientist, laboratory or institution should be compared. This fact disagrees with the generally accepted scientific principle that any numerical measurement should be compared to a "control" group in order to know its actual relevance (Is an *h*-index of 30 high or low? To which standards? It depends on what the distribution of the *h*-index among the equals of that researcher is). Unfortunately, there are no standardized ways of establishing what is high or low when it comes to defining an appropriate control group. The major players in bibliographic analysis should take into account this significant drawback faced by many scientists when applying for grants, fellowships, positions, etc. The jury evaluating the proposal may not be sufficiently related to the applicant's field, the application thus becoming either undervalued or overvalued. As an example, in Table 1, we show the best scored scientist in very different topics (Astrophysics, Biology, Sociology, Engineering and Literature) obtained by Google Scholar. As can be seen from Table 1, the scientific productivity indicators are by no means comparable, taking into account that these individuals are top-quality scientists in their own fields.

There have been a number of attempts to compensate for the differences among various disciplines. Podlubny (2005) observed how in the distribution of the scientific citations across a wide range of scientific fields published in the 2004 report of the National Science Foundation, the ratio of the total number of citations of any two broad fields of science remains close to constant. Based in this law, the normalization of the total number of citations with respect to Mathematics is suggested as a tool to compare performance in different fields of science. Iglesias and Pecharromán (2007) calculate the average number of citations/papers in different ISI fields and use it to calculate a multiplicative correction to the *h-index* for each field, provided in a very practical table. Imperial and Rodríguez-Navarro (2007) empirically observed that the highest *h-index* values attained for a given area correlate well with the impact factor of journals in that area. Based on this observation, they propose what they call the *reference h-index*, defined as $h_R = 16 + 11f$, f being the impact factor of the top journals that characterize the specific area or subarea. Despite the fact that the above mentioned works provide efficient tools to compare scientists working in different fields, it can be objected that the standardizations they propose are all based in assumptions derived from empirical observations. The results they obtain may be thus strongly conditioned by the statistical significance of the specific observations in which they are based. The patterns derived from such observations can be caused by specific trends or circumstances that might not apply anymore by the time the tools they provide are used.

Coleman et al. (2012) and El Emam et al. (2012) developed benchmarks for researchers who publish in logistic and medical informatics journals respectively, based on their peers performance. Using a sample of authors and publications in a selection of journals from the respective fields, these works identify a set of productivity thresholds allowing classifying individual authors of a field at various levels. A more general approach along the same line is proposed in Radicchi et al. (2008), where the statistical distributions of citations in the different disciplines are rescaled on a universal curve. With such purpose a relative indicator $cf = c/c0$ is considered, where cf is the number of citations received and $c0$ is the

Table 1 Scientific productivity indicators of the best scored scientific in very different topics

Name	Topic	Citations	*h-index*
William H. Press	Astrophysics	131,722	64
Eugene Koonin	Biology	98,272	156
Pierre Bourdieu	Sociology	308,893	198
David R. Nelson	Engineering	32,019	85
Gianfranco Delle Fave	Literature	7,442	45

average number of citations per article of the discipline. This gives the numerical value of *cf* the direct interpretation as the relative citation performance of the publication, also known as the "item oriented field normalized citation score" (Lundberg 2007), an analogue for a single publication of the popular Centre for Science and Technology Studies, Leiden (CWTS), field-normalized citation score or "crown indicator" (Moed et al. 1995). The work by Radichi et al. (2008) allows both classifying scientists in groups and ranking them, as well as mixing researchers working in different disciplines. However, this approach has not been massively adopted, and it can somehow be described as a "coarse grained" discipline correction, in the sense that it considers very broad disciplines. Notwithstanding, it points out in the right "statistical direction". Papers of the control group define a reference statistical distribution that can be used to assess the quality of a paper. Based on this premise, that is the standard methodology in statistical data analysis, we propose a way of referring scientific performance to the scientific performance of a control group and even a methodology to compare among different control groups.

Furthermore, it needs to be pointed out that the previous approaches focus just in accounting for the differences among research areas. However, a fair comparison between two scientists should consider other factors as well (e.g. whether they have been working on science a similar number of years, in a similar scientific environment with the same access to resources, funding access, scientific opportunities and equipment, institutional networking, …). This can be particularly important for young researchers with high potential, but trying to compete for funding with established labs. For this reason, we would like to draw your attention to the fact that the methodology that we are to introduce next enables to take into account the dependence that the scientific performance is known to have on age, gender, scientific environment, country and any other factor for which a control group can be defined.

Materials and methods

Fine grain performance measure among equals

In the following, we introduce a new methodology to measure scientific performance. We will refer to this method as "fine grained", because each component in the control group contributes to the score of the element being evaluated.

Let us assume that we have identified our control group, it does not matter how general it is (papers of scientists working in Physics, or scientists in U.S. developing new materials for X-ray photon detection in synchrotrons). In our opinion a good criteria to define the control group would be as follows: assuming that we want to evaluate a number of papers (these papers can be from a given scientist, a set of scientists, a department, a university, or may have been collected by any other criterion), and considering all those papers

referenced by the collection of papers to be evaluated and their authors, then the control group would be formed by the papers of these referred authors in journals of the same topic as the collection of papers being evaluated. However, any other criterion could be devised as long as the control group serves for the meant comparison purposes and the control group to which everyone is compared to is well defined. The construction of the control group has to pay careful attention to database issues like removal of duplicated items, incorrect entries, etc. But these issues fall out of the scope of this paper. Let us also assume that we have selected a performance measure that highlights a certain aspect we are interested in (total number of citations, total number of papers, number of citations per year, number of citations per year in the last 5 years, any of these measurements normalized with whichever correction factor we think is relevant). Any numerical index can be used and it may include normalization by age of the paper or the scientist, number of authors or any other factor that we may consider applicable. Let us call $X_i; i =, 2, \ldots, N$ the numerical indices of the N elements in the control group. We can use the values in X_i to obtain the empirical statistical distribution of the index within the control group. Note that we do not need to fit the distribution of the X_i values to any a priori distribution (Gaussian, power law, etc.). To assess the importance of a particular paper P with respect to the control group, we measure the chosen index X_P and the empirical probability in the control group of having an index smaller (or equal) than the element index ($\text{score}_p = \Pr(X < X_p)$ or $\text{score}_p = Pr(X \leq X_p)$; the difference between the two will be commented below). Note that this score index definition corresponds to the percentile value of X_P. This probability is obviously between 0 and 1 (the higher the value, the better the performance of the index of the P element being evaluated). Note that these probabilities can be calculated because the empirical distribution of X is given by the X_i measurements observed in the control group.

The previous measure evaluates the performance of a single paper. To assess the importance of a set of papers (for example, the papers of a scientist, or a group of scientists in a laboratory or an institution) with respect to the control group, we may compute the sum of all scores (gathering productivity and impact in a single measure) or compute any centrality measure (like its mean, median or trimmed mean). In the following, we will call the *P-index* to the index resulting of the sum of all scores with respect to our defined control group ($p = \sum_{p=1}^{N_p} \Pr(X \leq X_p)$ or $p = \sum_{1}^{N_p} \Pr(X_i < X_p)$ being N_p the number of papers considered). Note that the P-index summarizes the quality and impact of a set of performance scores and that it is bound by the total number of papers (as also is the *h-index*).

Fine grain performance measures among groups

The performance presented above covers the case of comparison among equals. Next, we will show how the *P-index* introduced in the previous subsection can be used to compare scientists, groups or institutions whose control groups differ.

Given two control groups, we first compute the *P-index* of each element in the first group using as control group the rest of elements in the group. We perform a similar analysis for the elements in the second group. The ranking of these elements resembles a preference learning problem. However, the standard input to these algorithms [some examples are Ranked Pairs (Tideman 1987 and Schulze's method 2003] are matrices whose entries are the frequency with which an option is preferred over a different option. We find that the problem of ranking the different scientists, groups of papers or groups of scientists does not fully fit in this setting. Alternatively, the field of Decision Making

Theory solved a similar problem using Analytic Hierarchy Process (AHP; Saaty 1988). One of the subproblems in AHP is to rank a number of categorical features by assigning them a numerical value according to their absolute importance. This numerical assignment and ranking is performed by pairwise comparisons of the different features and the calculation of the ratio of their relative importance. These ratios are provided by domain experts and the information of several of such experts can be combined. This methodology has been quite successful in Decision Making and the theoretic papers setting its foundations have received over 20,000 citations and found applicability in most scientific fields. AHP is a method to solve the so-called Multi-Criteria Decision Making. There are other methods to solve this sorting problem like those proposed by Geoffrion et al. (1972) and Köksalan and Sagala (1995). However, their grade of success in other fields is much smaller than that of AHP.

In the following we will use the methodology of AHP. To construct the relative importance matrix, H, let us assume that, among the two groups, we have a total of M elements that we need to rank together. If we construct a matrix of the form:

$$H = \begin{pmatrix} 1 & \frac{p_1}{p_2} & \cdots & \frac{p_1}{p_M} \\ \frac{p_2}{p_1} & 1 & \cdots & \frac{p_2}{p_M} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{p_M}{p_1} & \frac{p_M}{p_2} & \cdots & 1 \end{pmatrix}$$

where P_i is the *P-index* of the i-th scientist (computed with respect to its corresponding control group) divided by the maximum *P-index* in its group (this normalization assumes that the leaders of the different groups are equally performing). Then AHP produces a numerical rank by computing the eigenvector corresponding to the largest eigenvalue of H. This provides an objective way of ranking the different scientists with relative scores. In the following, we will refer to the eigenvector components as *W-index*.

Results and discussion

For the sake of illustrating the method to measure performance among equals introduced in section "Fine grain performance measure among equals", we have chosen a first control group, composed by the 13 Physics Nobel laureates between 2007 and 2012. Again for illustration purposes, we have chosen as performance measure the number of citations given by Scopus for the works published in the last 6 years divided by the age in years of the paper and the number of authors. Such data can be easily obtained from Scopus by introducing the name of the author in the search engine, going to the "Documents" link, and ticking the checkboxes of the corresponding years (2007–2012 in this case), in order to restrict the search to the desired temporal window. In this particular example, the control group includes N papers (N = 473) and the number of citations per year and number of authors ($X_i; i =, 2, \ldots, N$) ranges from 0 to about 205.2 (interestingly, 19.6 % of the papers received 0 citations, despite the fact that the scientific quality of these scientists is out of question). Table 2 shows the results for the 13 physicists considered in this example when they are compared according to the *P-index* (comparison among equals). For comparison purposes, we also report the *h-index (*provided by Scopus in the "Author Information" section) and what we will call the *T-index* $T = \sum_1^{N_p} X_p$ (being N_p the number of

publications considered for each author), which is one of the most common ways of aggregating impact. It needs to be noted that different performance measures (e.g. without normalization by age and/or number of authors) might produce different author rankings. It must also be noted that choosing $Pr(T < T_p)$ instead of $Pr(T \leq T_p)$ would entail different consequences. For example, papers with 0 citations under the first scoring scheme receive $score = 0$, while in the second scheme they receive a score different from 0 (in this particular case $score = 0.196$ since 19.6 % of the papers in the control group have no citations). Depending on whether we want to reward a published paper (even if it had no impact) or not, we must choose one or the other scheme.

Focusing in the results obtained under the criteria chosen, it can be observed in Table 2 how large *P-indices* correspond to scientists with many papers, consistently of a top-quality. There is in fact, a strong correlation both between the values of the *P-* and *T-indices* (specifically, the Pearson correlation coefficient is $r = 0.82$) and between the values of the *P-* and *h*-indices obtained ($r = 0.85$). However, the *P-index* overcomes some of the limitations of the *T-* and *h*-indices in the measurement of scientific performance. While allowing, as the *h-index*, to account for both the number of publications and citations (this will depend on the performance measurement used) of an author, the *P-index* takes into account all the citations of all the publications of an author. Moreover, the *P-index* is a time-variant index. It can grow not only if the scientist increases his number of publications but also if his papers have an impact sustained over the time whereas his counterparts do not. The *P-index* can as well decrease if a scientist "rests on his laurels" while his group mates make their progress or if the impact of his publications remains less cited than his partners'. This allows discriminating inactive from trendsetter scientists.

But the main novelty of the new *P-index* is that its value is relative to a specific control group. In order to illustrate this aspect, we have performed a second experiment. The performance measure is again the number of citations normalized by the paper age and the number of authors, but this time the control group is formed by the most productive scientists (between 2007 and 2012) of the University of Manchester in the same Scopus areas as the Nobel Prizes Konstantin Novoselov and Andre Geim; we have chosen this university because both scientists are affiliated to this institution. Both scientists belong thus to the two groups, Nobel Prizes and Manchester University, that have been studied. In order to get the data fitting these criteria, a search by affiliation has been performed. Scopus allows restricting the authors search to a particular research center just by introducing the corresponding university or institution (Manchester University in this case) in the Affiliation box of the search engine. Then, the results have been filtered by areas by ticking the corresponding checkboxes that match the areas of publication associated to Konstantin Novoselov and Andre Geim by Scopus. Finally, the resulting authors have been ordered by descending number of publications just by selecting this criterion in the results page. As can be observed in the results shown in Table 3 (corresponding to 45 scientists with 6,220 papers), the relative order of both authors is kept when they are evaluated with respect to the second group (as expected, since the measure of comparison is the same), but the value of their respective *P-indices* change with respect to the first experiment, according to our claim that scientific performance depends on the specific context. Strangely enough, the two Nobel prizes do not occupy the first positions of this second ranking and are not the most cited authors either, two facts that once again prove the single-index evaluators to be limited. However, their *P-indices* are higher than when assessed with respect to the other Nobel Prizes, a fact that is in agreement with the higher scientific standards, and therefore fiercer competition, that can be expected in such a distinguished group.

Table 2 *P-index* and *T-index* and *h-index* analysis for the Physics Nobel laureates since 2007

Name	*P-index*	*T-index*	*h-index*
K. Novoselov	67.63	184.06	54
A. Geim	63	404.1	62
A. Riess	24.92	15.06	52
A. Fert	24.12	24.21	50
M. Kobayashi	21.4	5.16	32
S. Perlmutter	18.87	5.98	39
B. Schmidt	11.63	3.16	48
G. Smith	6.07	1.54	5
Y. Nambu	3.34	1.93	5
T. Maskawa	2.02	0.23	1
P. Grunberg	1.99	2.7	17
C. Kao	0.87	0.27	1
W. Boyle	0.86	1	1

The differences among the *P-indices* of Konstantin Novoselov and Andre Geim vary as well from one experiment to another. However, since they have been computed in different scales, a fair comparison between the distances separating both authors within their respective rakings cannot be performed. This is precisely the type of problems that want to be addressed with the computation of the *W-index* introduced in section "Fine grain performance measures among groups". In order to apply the *W-index* as a tool for the evaluation of scientists that belong to different control groups we have chosen first the Physics Nobel laureates (13 physicists with 473 papers) and the Chemistry Nobel laureates (11 chemists with 350 papers) between 2007 and 2012, as differing control groups. We have measured the publication and impact pattern using the normalized *P-index* within each group (physicists are compared to the Physics control group, and chemists are compared to the Chemistry control group). Then, we have computed the corresponding *W-indices*. Table 4 shows the ranking obtained for the $M = 13 + 11 = 24$ scientists, by applying the methodology described in section "Fine grain performance measures among groups", in which their belonging to different groups is taken into account. Although the classification obtained by ordering the authors according to their normalized *P-index* matches exactly the *W-index* based ranking, this does not need to be always the case, and the use of the *W-index* is recommended over the *P-index* when evaluating different groups. The fourth column of the table shows an item-oriented normalized indicator, computed as the sum of the number of citations received by the accounted papers, normalized by the average number of citations per article of the corresponding, physics, or chemistry, discipline. The normalization factors (8.74 for the physicists and 10.74 for the chemists) have been obtained from Thomson Corp. Item-oriented normalization is a normalization procedure that is becoming standard when comparing different disciplines. Note that there is no direct relationship between the *W-index* and the item-oriented normalized index, meaning that these two indices are actually measuring different aspects of scientific performance. It may be argued that the item-oriented normalized index is a more coarse grain measurement since it uses a single average number of citations per discipline while underlying the *W-index* is the whole distribution of citations within the control group.

The *W-index* allows not only to rank the scientists but also to put their performance indices in a common scale in which their relative distances can be measured (i.e. not only

Table 3 *P-index*, *T-index* and *h-index* analysis for the Manchester affiliated scientists

Name	*P-index*	*T-index*	*h-index*
K. Harder	200.75	1.1	45
K. Johns	196.58	1.05	50
Christian Schwanenberger	181.08	1	43
G Hesketh	146.85	0.87	38
G. Lafferty	140.63	0.87	56
A. Das	140.33	0.9	36
S. Soldner-Rembold	138.06	1.13	47
Y. Peters	130.36	0.82	31
K. Petridis	106.37	0.47	31
D. Bailey	102.37	0.61	48
W. Yang	102.05	0.65	25
L. Suter	95.64	4.12	29
K. Novoselov	**88.11**	**184.06**	**54**
P. Rich	87.4	0.56	28
C. McGivern	78.86	0.53	23
A. Geim	**78.79**	**404.11**	**62**
K. Alwyn	76.67	0.48	24
M. Vesterinen	75.67	1.24	20
T. West	74.84	0.39	29
G. Jackson	71.8	0.33	20
F. Tuna	71.45	21.17	25
S. Snow	66.72	0.37	29
J. Pater	66.62	0.23	45
F. Loebinger	66.11	0.21	43
G. Brown	60.2	0.17	27
V. Chavda	60.2	0.17	27
J. Howarth	56.78	0.16	24
J. Lane	56.56	0.16	26
M. Ibbotson	55.4	0.16	43
J. Almond	53.59	0.22	27
G. Timco	50.81	7.79	30
E. McInnes	48.7	10.29	30
D. Cullen	26.66	0.42	22
V. Markevich	23.85	1.3	18
J. Billowes	23.05	1.44	17
R. Bishop	21.55	5.53	16
M. Ford	17.95	0.11	32
G. King	11.11	0.87	16
B. Varley	8.74	0.32	23
J. Durell	5.89	0.37	21
J Dowker	4.04	2.54	10
A. Turcot	2.12	0.01	40
A. Donnachie	0.98	0.33	19
M. Naisbit	0.9	0	28
B. Middleton	0.3	0	7

Entries in bold indicates Nobel laureates

to determine which scientist is "better" but how much "better" than others he is). As has been previously mentioned, this type of benchmark can be extremely useful in many practical situations in which the performance of several scientists is assessed by taking into account their differences either in terms of academic disciplines, career length, scientific environment or any other relevant factor (these factors have to be used in the normalization of the underlying X_p measurements). For example, Table 5 shows a joint scaled ranking, always according to our methodology and to the evaluation criterion chosen, of the most productive scientists from the Manchester University (in the same Scopus areas as its affiliated Physics Nobel prizes Konstantin Novoselov and Andre Geim) and from the Iowa State University (in the Scopus areas of its affiliated Chemistry Nobel Prize Dan Shechtman). The first group gathers 45 scientists with 6,220 papers, whereas the second consists of 45 scientists with 849 papers. In this way, we have two rankings including the three above mentioned Nobel Prizes, one that accounts for their different field of work and another that accounts as well for their different affiliation. The results obtained in Table 5 illustrate how the comparisons among scientists may vary if their different contexts are taken into account, and how the *W-index* can be used to consider and measure such

Table 4 *W-index* normalized, *P-index* and item-oriented normalized analysis for the Physics and Chemistry Nobel laureates since 2007

Name	Nobel prize	*W-index*	Normalized *P-index*	Item-oriented normalized index
K. Novoselov	Physics	0.48	1	21.06
R. Heck	Chemistry	0.48	1	3.19
A. Geim	Physics	0.45	0.93	101.08
R. Tsien	Chemistry	0.24	0.51	53.84
T. Steitz	Chemistry	0.24	0.5	26.37
A. Suzuki	Chemistry	0.23	0.47	11.97
A. Riess	Physics	0.18	0.37	8.03
A. Fert	Physics	0.17	0.36	2.77
M. Kobayashi	Physics	0.15	0.32	0.59
S. Perlmutter	Physics	0.13	0.28	0.68
E. Negishi	Chemistry	0.13	0.28	13.7
V. Ramakrishnan	Chemistry	0.12	0.25	11.25
A. Yonath	Chemistry	0.11	0.23	4.65
M. Chalfie	Chemistry	0.11	0.23	8.27
B. Schmidt	Physics	0.08	0.17	4.13
G. Smith	Physics	0.04	0.09	0.34
G. Ertl	Chemistry	0.04	0.09	17.37
Y. Nambu	Physics	0.02	0.05	3.8
D. Shechtman	Chemistry	0.02	0.05	1.74
O. Shimomura	Chemistry	0.02	0.04	0.28
T. Maskawa	Physics	0.01	0.03	0.03
P. Grunberg	Physics	0.01	0.03	0.31
C. Kao	Physics	0.01	0.01	0.03
W. Boyle	Physics	0.01	0.01	0.11

Table 5 *W-index* normalized and *P-index* analysis for the Manchester and Iowa State affiliates scientists

Name	Affiliation	*W-index*	Normalized *P-index*
K. Harder	Manchester	0.31	1
J. Clutter	Iowa	0.31	1
K. Johns	Manchester	0.3	0.98
C. Schwanenberger	Manchester	0.28	0.9
G. Hesketh	Manchester	0.22	0.73
G. Lafferty	Manchester	0.22	0.7
A. Das	Manchester	0.21	0.7
S. S. Rembold	Manchester	0.21	0.69
Y. Peters	Manchester	0.2	0.65
K. Petridis	Manchester	0.16	0.53
D. Bailey	Manchester	0.16	0.51
W.C. Yang	Manchester	0.16	0.51
L. Suter	Manchester	0.15	0.48
K. Novoselov	**Manchester**	**0.13**	**0.44**
P. Rich	Manchester	0.13	0.44
F. Xiu	Iowa	0.12	0.4
C. McGivern	Manchester	0.12	0.39
A. Geim	**Manchester**	**0.12**	**0.39**
K. Alwyn	Manchester	0.12	0.38
M.Vesterinen	Manchester	0.12	0.38
T. West	Manchester	0.11	0.37
R. Fernandes	Iowa	0.11	0.36
G. Jackson	Manchester	0.11	0.36
F. Tuna	Manchester	0.11	0.36
S. Snow	Manchester	0.1	0.33
J. Pater	Manchester	0.1	0.33
F. Loebinger	Manchester	0.1	0.33
G. Brown	Manchester	0.09	0.3
V. Chavda	Manchester	0.09	0.3
J. Howarth	Manchester	0.09	0.28
J. Lane	Manchester	0.09	0.28
M. Ibbotson	Manchester	0.09	0.28
A.Baran	Iowa	0.08	0.27
J. Almond	Manchester	0.08	0.27
G. Timco	Manchester	0.08	0.25
W. Yuhasz	Iowa	0.08	0.25
E. McInnes	Manchester	0.07	0.24
R. Dhaka	Iowa	0.07	0.24
A. Russell	Iowa	0.07	0.22
Y. Han	Iowa	0.07	0.22
D. Pratt	Iowa	0.07	0.22
T. Kempel	Iowa	0.06	0.2
F. Wei	Iowa	0.06	0.19

Table 5 continued

Name	Affiliation	*W-index*	Normalized *P-index*
N. Singh	Iowa	0.05	0.16
A. Struck	Iowa	0.05	0.16
M. Vuckovic	Iowa	0.05	0.15
X. Sheng	Iowa	0.05	0.15
J. Lamsal	Iowa	0.04	0.14
N. Shen	Iowa	0.04	0.13
A. MCullen	Manchester	0.04	0.13
M. Vannette	Iowa	0.04	0.13
A. Stefanescu	Iowa	0.04	0.13
V. Markevich	Manchester	0.04	0.12
J. Billowes	Manchester	0.04	0.11
T. Nagai	Iowa	0.04	0.11
S. Sivasankar	Iowa	0.03	0.11
V. Smetana	Iowa	0.03	0.11
R. Bishop	Manchester	0.03	0.11
Y. Wu	Iowa	0.03	0.09
M. Ford	Manchester	0.03	0.09
S. Nelson	Iowa	0.03	0.09
W. Keith	Iowa	0.02	0.08
S. Prell	Iowa	0.02	0.08
L. Willson	Iowa	0.02	0.08
O. Pestovsky	Iowa	0.02	0.08
A. Ruiz-Martinez	Iowa	0.02	0.07
C. Chen	Iowa	0.02	0.07
K. Yamamoto	Iowa	0.02	0.07
S. Thimmaiah	Iowa	0.02	0.07
S. Xu	Iowa	0.02	0.07
C. Kerton	Iowa	0.02	0.06
G. King	Manchester	0.02	0.06
D. Urner	Manchester	0.02	0.05
D. Shechtman	**Iowa**	**0.01**	**0.05**
J. Xu	Iowa	0.01	0.05
B. J. Varley	Manchester	0.01	0.04
R. Huang	Iowa	0.01	0.04
F. Margetan	Iowa	0.01	0.04
X. Lin	Iowa	0.01	0.03
S. Bahadur	Iowa	0.01	0.03
L. Brasche	Iowa	0.01	0.03
J. Durell	Manchester	0.01	0.03
S. Song	Iowa	0.01	0.02
J. Dowker	Manchester	0.01	0.02
D. Carter Lewis	Iowa	0	0.01
A. Turcot	Manchester	0	0.01

Table 5 continued

Name	Affiliation	*W-index*	Normalized *P-index*
Y. Li	Iowa	0	0.01
S. Lin	Iowa	0	0.01
A. Donnachie	Manchester	0	0
M. Naisbit	Manchester	0	0
N. Anderson	Iowa	0	0
B. Middleton	Manchester	0	0

Entries in bold indicates Nobel laureates

variation. Relevant differences can, in fact, be observed among the *W-indices* obtained for the three reference scientists (Konstantin Novoselov, Andre Geim and Dan Shechtman) when referred to their field in Table 4 (*W-indices* of 0.48, 0.45, 0.02 respectively) or their affiliation in Table 5 (*W-indices* of 0.13, 0.12 and 0.01). Specifically, it can be observed how Dan Shechtman reduces his distance with respect to Konstantin Novoselov, and Andre Geim when his affiliation to a less renowned institution (in the 2012 Academic Ranking of World Universities, http://www.arwu.org, Manchester is ranked 40th in the world, whereas the Iowa State University is out of the rank) is taken into account. Such tendency is also reflected in the normalized *P-indices* of the three authors, but since the values belong to two distinct scales (Physics and Chemistry), the differences among the *P-indices* cannot be used as a reliable distance measure.

As a final remark, we want to point out that the aim of the above experiments is by no means to assess the scientific value of the contribution of any of the authors used in the examples, but to illustrate the tools introduced/presented, that could be applied in a real research assessment decision. For this reason the reference groups chosen have been those that allowed us to easily classify some of the authors according to several different contexts, and may not be groups that make the most radical differences between their components. For example, the publication and citation practices in the areas of Physics and Chemistry Nobel Prizes belong are not so discordant and, although the Manchester University is usually ranked before the Iowa State University, both are unquestionably prestigious universities. However, if these, somehow naive, examples already show how the belonging to different reference groups can condition the scientific evaluation, more significant differentiations can be expected in real situations, where more critical factors may come into play.

Conclusions

Measuring performance among different scientists, laboratories or institutions is an important issue when trying to get the most from the money invested in science. With more or less success many different performance measures have been proposed (probably, several of them are needed to provide a complete picture of scientific performance). However, the absolute scale of each one of these measures is ill-defined since it strongly depends on the scientific discipline, scientific environment, or scientific age (is a *h-index* of 30 high or is it low? It depends on what we compare it to). We have proposed a way of referring scientific performance to the scientific performance of any control group and a methodology to compare among different control groups. In particular we have defined:

- *The P-index*: that compares any performance score (h-index or any other that highlights those features we are interested in) to the distribution of the same performance score computed in a control group. The P-index is calculated using a fine-grained approach based on individual papers and, consequently, can be used to carefully distinguish among similar performances.
- *The W-index*: that allows integrating two lists of scientists, departments or institutions with different control groups.

Both together allow us to effectively compare an author, department or institution to his equals and compare authors, departments or institutions with different scientific backgrounds. Finally, it needs to be pointed out that, to be optimally effective, maximizing the use of the resources devoted to science should also consider the monetary income (e.g., it is not enough to measure the number of citations per year of a given institution, the number of citations per year and invested dollars should also be evaluated). If measuring scientific outcome is a controversial issue, measuring this productivity by invested money can raise bitter debates among scientific stakeholders, which was not the aim of this article.

References

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics, 3*(4), 273–289.

Bornmann, L., Mutz, R., & Daniel, H. D. (2008). Are there better indices for evaluation purposes than the h index?A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology, 59*, 830–837.

Coleman, B. J., Bolumole, Y. A., & Frankel, R. (2012). Benchmarking individual publication productivity in logistics. *Transportation Journal, 51*(2), 164–196.

El Emam, K., Arbuckle, L., Jonker, E., & Anderson, K. (2012). Two h-index benchmarks for evaluating the publication performance of medical informatics researchers. *Journal of Medical Internet Research, 14*(5). doi: 10.2196/jmir.2177.

Geoffrion, A. M., Dyer, J. S., & Feinberg, A. (1972). An interactive approach for multi-criterion optimization, with an application to the operation of an academic department. *Management Science, 19*, 357–368.

Glänzel, W. (2006). On the Opportunities and Limitations of the H-index. *Science Focus, 1*(1), 10–11.

Harzing, A. (2010). *The publish or perish book*. Melbourne, Australia: Tarma Software Research.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 16569–16580.

Iglesias, J. E., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics, 73*(3), 303–320.

Imperial, J., & Rodríguez-Navarro, A. (2007). Usefulness of Hirsch's h-index to evaluate scientific research in Spain. *Scientometrics, 71*(2), 271–282.

Köksalan, M. M., & Sagala, P. N. S. (1995). Interactive approaches for discrete alternative multiple criteria decision making with monotone utility functions. *Management Science, 41*, 1158–1171.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics, 76*, 369–390.

Lundberg, J. (2007). Lifting the crown-citation z-score. *Journal of Informetrics, 1*, 145–154.

Moed, H. F., Debrun R. E., & Vanlleuwen T. (1995). New bibliometric tools for the assessment of national research performance-database description, overview of indicators and first applications. *Scientometrics*, *33*, 381–422.

Moed, H. F. (2009). New developments in the use of citation analysis in research evaluation. *Archivum immunologiae et therapiae experimentalis*, *57*(1), 13–18.

Panaretos, J., & Malesios, C. (2009). Assessing scientific research performance and impact with single indices. *Scientometrics*, *81*(3), 635–670.

Podlubny, I. (2005). Comparison of scientific impact expressed by the number of citations in different fields of science. *Scientometrics, 64*(1), 95–99.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America, 45*, 17268–17272.

Saaty, T. (1988). What is the *Analytic hierarchy process*? *Mathematical models for Decision Support, 48*, 109–121.

Schulze, M. (2003). A new monotonic and clone-independent single winner election method. *Voting Matters, 17*, 9–19.

Tideman, T. N. (1987). Independence of clones as a criterion for voting rules. *Social Choice and Welfare, 4*, 185–206.