Journal of Informetrics 11 (2017) 625–640

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

The Herrero-Villar approach to citation impact

Pedro Albarrán[a], Carmen Herrero[b], Javier Ruiz-Castillo[c,*], Antonio Villar[d]

[a] *Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, Spain*
[b] *Departamento de Fundamentos del Análisis Económico, Universidad de Alicante & IVIE, Spain*
[c] *Departamento de Economía, Universidad Carlos III, Spain*
[d] *Departamento de Economía, Universidad Pablo de Olavide & IVIE, Spain*

ARTICLE INFO

Article history:
Received 19 December 2016
Received in revised form 26 April 2017
Accepted 26 April 2017
Available online 17 May 2017

Keywords:
Citation impact
Additive rules
Tournaments
Relative research performance
Worth

ABSTRACT

This paper focuses on the evaluation of research institutions in terms of size-independent indicators. There are well-known procedures in this context, such as what we call additive rules, which provide an evaluation of the impact of any research unit in a scientific field based upon a partition of the field citations into ordered categories, along with some external weighting system to weigh those categories. We introduce here a new ranking procedure that is not an additive rule – the HV procedure, after Herrero & Villar (2013) – and compare it those conventional evaluation rules within a common setting. Given a set of ordered categories, the HV procedure measures the performance of the different research units in terms of the relative probability of getting more citations. The HV method also provides a complete, transitive and cardinal evaluation, without recurring to any external weighting scheme. Using a large dataset of publications in 22 scientific fields assigned to 40 countries, we compare the performance of several additive rules – the Relative Citation Rate, four percentile-based ranking procedures, and two average-based high-impact indicators – and the corresponding HV procedures under the same set of ordered categories. Comparisons take into account re-rankings, and differences in the outcome variability, measured by the coefficient of variation, the range, and the ratio between the maximum and minimum index values.

1. Introduction

In a globalized and highly interconnected world, comparative exercises have become more and more frequent in many aspects of life. Research is no exception and there seems to be growing interest in the evaluation of the scientific influence. Citation analysis has become one of the key tools for evaluating the scientific performance of research units (individual authors, research groups, departments, universities, countries, etc.). Citation impact indicators differ depending on the evaluation approach, the motivation, and the way of transforming citations into specific evaluation formulae. In this paper, we contribute to the literature of citation analysis that focuses on the *ranking of research units* by *size-independent measures of citation impact* when size is measured by the number of publications, i.e. measures that take the relative citation frequencies as the basis for the evaluation (see, among many others, Bornmann, De Moya Anegón, & Leydesdorff, 2012, Fairclough & Thelwall, 2015, and Glänzel, Thijs, & Debackere, 2014). We propose a new procedure that evaluates the citation impact of a set of research units according to the criterion pioneered in Herrero & Villar (2013) (*HV* in the sequel).

[*] Corresponding author.
E-mail address: jrc@eco.uc3m.es (J. Ruiz-Castillo).

http://dx.doi.org/10.1016/j.joi.2017.04.008
1751-1577/© 2017 Elsevier Ltd. All rights reserved.

This new evaluation protocol can be thought of as a two-step procedure in which we first define a partition of the range of citations into a series of categories that gather publications of similar merit, and then evaluate the research units' relative citations distributions embedded in these categories. The key informational item to compare research units is, therefore, the shares of the publications into the different categories. The comparison of these distributions is made in terms of the following principle: each research unit will be compared with all others in terms of the probability of getting a greater citation impact. We shall see that this procedure can also be formulated in terms of a series of tournaments in which each research unit is confronted with all others repeatedly.[1]

There are well-established evaluation procedures that also rely on the assessment of the research units' relative citations distributions by categories using different principles. We shall consider here three types of these indicators that will be used as reference for comparison with the *HV* evaluations. The first type is the *Relative Citation Rate, RCR* (Schubert & Braun, 1986, and Vinkler, 1986). The second type, promoted since 2010 by a group of highly qualified professional leaders in scientometrics, corresponds to what Bornmann & Mutz (2011) call the *percentile rank approach*.[2] The third type consists of the FGT family of high-impact indicators, introduced in Albarrán, Ortuño, and Ruiz-Castillo (2011a), which are real valued functions defined over the subset of publications with citations above a *critical citation line* (CCL), and whose properties are inherited from a class of economic poverty indicators introduced by Foster, Greeer and Thorbecke (1984).[3]

All these evaluation procedures have in common an additive structure, as shown in Section 2. That is, the evaluation of the different research units is given by a weighted sum of the relative citations distribution by categories, where the weights measure the importance of each category. They can be described as implementing the following protocol. First, publications are distributed into a set of categories that gather those publications regarded as being of similar merit. Second, each of those categories is given a weight that determines the rate of substitution between the corresponding categories. And third, the evaluation of each research unit is obtained as a weighted sum of its relative frequencies aggregated by categories.

The additive structure of these evaluation procedures is very appealing because it provides a relatively simple construct that is easy to interpret and rather immediate to compute. The main shortcoming of these indicators is that the evaluation turns out to depend critically on the choice of the weights with which we ponder the publications. Quite often there is no good reason to choose a particular weighting system, which makes the evaluation exercise somehow arbitrary because both the evaluation of the individual units and their ranking depend on those weights. The new evaluation approach presented in this paper avoids this inconvenience because no weighting of categories is involved and still provides a complete, transitive and cardinal evaluation.

Any comparison between alternative ranking procedures should involve not only their rationale and their properties but also the empirical differences they give rise to in applications. Following this idea, we consider here an empirical analysis based on a dataset indexed by Thomson Scientific, and consisting of 4.4 million articles published in 1998–2003, and the citations they received during a five-year citation window for each year in that period. Articles are classified into the 20 natural sciences and the two social sciences distinguished by this firm. Using these data, we compare the *HV* ranking procedure and the ranks provided by a group of additive procedures in four scientific fields.[4] We compare ranking procedures both from an ordinal point of view (changes in the ranking) and from a cardinal point of view (differences in the spread of the evaluations, as measured by the coefficient of variation, CV, the range, and the ratio between the maximum and minimum). We study a partition of the world into 39 countries and a residual geographical area.

The remainder of this paper is organized into four Sections and an Appendix. Section 2 presents a selection of additive ranking procedures. Section 3 describes the alternative approach for the evaluation of research units in a single field by adapting the ideas in Herrero & Villar (2013) to this context. The empirical Section 4 develops a comparison between this new ranking procedure and the selected additive procedures in the following fields: Clinical Medicine, Physics, Engineering, and Economics & Business. These fields have been selected endeavoring to ensure diversity and relevance, while keeping the set of empirical comparisons within reasonable limits. Section 5 contains discussion. The Appendix A includes some examples and descriptive statistics. Also, to facilitate reading of the text, some statistical results are relegated to a Supplementary Material section.

[1] The recourse to tournaments as an evaluation procedure has been applied in related contexts, such as the Google Page Rank algorithm to rank web pages (Altman & Tennenholtz, 2005 and Page et al., 1998), as well as the invariant method (Palacios-Huerta and Volij, 2004), the Eigenfactor (Bergstrom, 2007, and West et al., 2010), and the recent paper by Kóczy & Nichifor (2013) that have been used to rank scientific journals. The closest contribution to ours is Carayol & Lahatte (2014), which uses the idea of tournaments for ranking research units when citation impact and quantity both matter.

[2] See also the Integrated Impact, or the *I3* indicator in Leydesdorff et al. (2011), Leydesdorff and Bornmann (2011), Leydesdorff (2012), Wagner & Leydesdorff (2012), and Rousseau (2012). In their search for standards for applying bibliometric methods in the evaluation of research institutes or individuals, Bornmann & Williams (2013), as well as Bornmann, Marx, and co-authors point to the percentile rank approach as the obvious choice (Bornmann & Marx, 2013, 2014, and Bornmann, Mutz, Nehaus, & Daniel, 2008, Bornmann et al., 2014).

[3] For empirical applications of members of the FGT family, see Albarrán, Ortuño, and Ruiz-Castillo (2011b), Albarrán, Ortuño, and Ruiz-Castillo (2011c), Herranz & Ruiz-Castillo (2012, 2013), and Perianes-Rodriguez & Ruiz-Castillo (2016).

[4] The study has been made for all 22 fields, obtaining similar results. We here report the results for four fields for the sake of parsimony. All remaining results can be obtained from the authors upon request.

2. The additive approach to the evaluation of citation impact

2.1. Framework

A research unit is an institution consisting of a collection of researchers with a common affiliation. It may refer to a University, a University Department, a lab, or even a country. Our reference problem is that of evaluating the citation impact of *U research units*, indexed by $u = 1, \ldots, U$, in a given research field during a given time span, in terms of their *publications* in a given set of journals. Let $A = \{1, 2, \ldots, J\}$ denote a set of scientific journals considered appropriate to disseminate the research outcomes in the corresponding field. Each journal consists of a collection of dated issues, each of which contains a number of research articles. For a given time span and a given set of research units, we call ***publication*** to an article published in a journal in A, during the selected time span, for some member/s of the reference research units. So articles published in journals outside A, in a different period from the one selected, or in which no researcher belongs to a research unit in U, are not considered publications in our framework. In this Section, we adopt the simplifying assumption that each publication can be attributed to one and only one research unit (this amounts to assuming no co-authorship between people of different units and a single unit affiliation for each author).

Let B denote the set of publications. Each publication contains a number of different references to other publications. The number of citations of a publication b is the number of publications that contain b in their references. The relevance of the contribution of a research unit is related to the *citations* of its publications.

We represent a research unit u by a mapping $q_u : \boldsymbol{N} = \left\{0, 1, 2, \ldots\right\} \to \boldsymbol{N}$, where $q_u(r)$ is the number of publications of research unit u with exactly r citations. Hence, a research unit is represented by its publication record as given by q_u. For instance, $q_u(0) = 2$, $q_u(5) = 3$ means that this research unit has two publications with zero citations and three publications with five citations. When there is no publication with r citations we have $q_u(r) = 0$. Note that the total number of publications of unit u is given by $\Sigma_{r\varepsilon N} q_u(r)$. Note also that, in the absence of information on the authors that work in a given institution, the notion of productivity recently defended by Abramo and D'Angelo (2016), as well as the joint ranking of scientists and institutions studied in Bouyssou and Marchant (2011), are beyond the scope of this paper.

We shall focus here on those evaluation procedures that are size-independent, that is, those that pay attention to the relative frequencies of citations, rather than the absolute ones.[5] This is the case when evaluation procedures satisfy the property of *replication invariance*: replicating any finite number of times the citations of all research units does not change their evaluations. Then, we define the ***relative citations distribution*** of unit u, $D^u : \boldsymbol{N} \to \boldsymbol{Q}$, where $\boldsymbol{Q}$ stands for the set of rational numbers, as follows:

$$D^u(r) = \frac{1}{\sum_{r \in N} q_u(r)} q_u(r).$$

Similarly, we can represent a scientific field as a mapping $Q : \boldsymbol{N} \to \boldsymbol{N}$ with $Q(r) = \Sigma_{u=1}^{U} q_u(r)$, and the corresponding relative citations distribution, $D : \boldsymbol{N} \to \boldsymbol{Q}$, as

$$D(r) = \frac{1}{\sum_{r \in N} Q(r)} Q(r).$$

An evaluation problem is thus given by $P = \left\{D^u\right\}_{u=1}^{U}$. For any problem P, our aim is to rank the different research units in terms of their citations by means of a vector valued evaluation function F, with $F(P) \in R^U$, so that $F_u(P) > F_v(P)$ if and only if unit u is considered as better or precedes unit v.

2.2. The additive approach

Let us present now the family of additive evaluation procedures that constitute one of the standard ways of dealing with the evaluation problem in this scenario. This family can be described as a two-step procedure defined over the relative citations distributions, as follows.

Step 1. For a given evaluation problem, let r^* denote a natural number such that $q_u(r) = 0$, for all $r > r^*$ (i.e. no publication gets more than r^* citations). We now establish a partition of the set $\{0, 1, 2, \ldots, r^*\}$ in terms of G consecutive intervals $I_1 = [0, r_1]$, $I_2 = [r_1 + 1, r_2]$, ..., $I_G = [r_{G-1} + 1, r^* + 1]$. Those intervals determine a partition of the publications into a collection of categories, $\Theta = \{\theta_g\}_{g=1}^{G}$, which gather publications of similar merit. That is, all publications whose citations belong to the same interval I_g are considered as equally valuable and thus assigned to category θ_g. Each of those categories is given a weight, ω_g, in a monotonous way (a better category has associated with it a higher weight).

Step 2. The relative citations distribution of each research unit is embedded into these categories, as follows. We denote by $c_g^u = \Sigma_{r\varepsilon I_g} D^u(r)$ the sum of relative frequencies of all publications within category θ_g. That is, c_g^u is the fraction of publications

[5] The h-index introduced by Hirsh (2005) and its many variants, the scoring rules in Marchant (2009), and all ranking procedures in Carayol & Lahatte (2014) are not size-independent. Therefore, they all lie outside the scope of this paper.

produced by unit u that belong to category θ_g. We can describe the relative citations distribution of unit u embedded in the set of categories Θ, by an array $D_\Theta(u) = (c_1^u, \ldots, c_G^u)$. The evaluation formula adopts the following format, for $u = 1, 2, \ldots, U$:

$$F_u(P) = \Sigma_g \omega_g^F c_g^u \tag{1}$$

In this formulation, value judgments are introduced in the configuration of the categories, $\Theta = \{\theta_g\}_{g=1}^G$, and the choice of weights, ω_g^F, in the understanding that the higher the merit the higher the weight. Notice that with the same set of categories, different weights give rise to different evaluation rules.

Among the different evaluation procedures that adopt this format, in this paper we consider three different types often found in the literature: (A) The *Relative Citation Rate*, *RCR*, based on the comparison of mean citations; (B) The percentile evaluation procedure, based on the classification of the field publications according to percentiles (or percentile classes), and some associated weights; (C) The FGT family of high-impact indicators, defined over the subset of publications with citations above a critical citation level *CCL*.

2.2.1. Relative citation rate indicator

Given an evaluation problem $P = \left\{D(u)\right\}_{u=1}^{U}$, let μ_u and μ denote the mean citation of distributions D^u and D, respectively. The *RCR* procedure is a citation ranking in terms of the ratio:

$$RCR(u) = \frac{\mu_u}{\mu}. \tag{2}$$

This corresponds to a particular specification of (1) with $I_g = g$ (a degenerate interval) from g = 0 to g = r*, and $\omega_g^{RCR} = \frac{g}{\mu}$. Here, as any citation level r receives implicitly a different weight, the associated categories correspond to the distinct values in the field citations distribution.

2.2.2. Percentile evaluation procedures

In the percentile approach the set of ordered categories, $\Theta = \{\theta_g\}_{g=1}^G$, is such that θ_g consists of publications in a certain percentile or percentile class and the *typical* choice of weights is given by $\omega_g^\pi = g$. Note that in this case the weights do not change from problem to problem. The evaluation formula is therefore given by:

$$\pi(u) = \Sigma_g g c_g^u \tag{3}$$

We now present three traditional examples of percentile procedures.

1 The *C100* procedure, where categories correspond to the 100 percentiles in *C*, and the weights range from 1 to 100 for each percentile in ascending order. Thus, $\omega_g^{C100} = g$,for $g = 1, \ldots, 100$.

2 The *NSF6* procedure, adopted in the *Science & Engineering Indicators* of the National Science Foundation of the U.S. (National Science Foundation, 2010; see also Leydesdorff, Bornmann, Mutz, & Opthof, 2011, and Schreiber, 2012). This procedure considers six categories, given by: θ_1 = publications in the interval [1,50th]; θ_2 = publications in the interval (50th, 75th]; θ_3 = publications in the interval (75th, 90th]; θ_4 = publications in the interval (90th, 95th]; θ_5 = publications in the interval (95th, 99th], and θ_6 = publications in the interval (99th, 100th]. Here, again, $\omega_g^{NSF6} = g$, for $g = 1, \ldots, 6$.

3 The *Top k% dichotomous procedure,* where the reference set is partitioned into two categories, θ_1, the top *k%* most cited publications, and θ_0, the rest. In this case, $\omega_g^{Topk} = g$, ranks research groups according to the proportion of papers in θ_1 .

Dichotomous procedures rely on the idea that only the upper part of the distribution matters and are used in the influential *Leiden Ranking* for universities and the *SCImago Institutions Ranking*.[6] Specifically, the 2015 version of the Leiden Ranking includes the cases *k%* = 50%, 10%, and 1%, while the *Excellence Rate* indicator corresponding to *k%* = 10% is one of the elements of the composite indicator used by SCImago to rank research institutions.

2.2.3. FGT high-impact indicators

The dichotomous procedure presented above is a particular instance of those evaluation methods that focus on high impact publications and disregard the rest. The latter can be formulated in terms of a CCL z, so that any citation below z is considered as irrelevant. Thus, only publications with citations over z (high impact publications) are considered as important.

[6] The *Leiden Ranking* (www.leidenranking.com) is a university ranking compiled by the Centre for Science and Technology Studies, Leiden University, The Netherlands, using the Thomson Reuters bibliographic database Web of Science as the source of the publication and citation data. SCImago is a research group from the *Consejo Superior de Investigaciones Científicas*, the University of Granada, Extremadura, Carlos III (Madrid), and Alcalá de Henares in Spain. The *SCImago Institutions Rankings* (SIR; www.scimagoir.com) is a bibliometric ranking of research institutions based on Elsevier's Scopus database. The web-based evaluation tool *InCites* (Thomson Reuters), http//incites.thomsonreuters.com/, also uses the percentile rank approach.

For each z, the *FGT family* of evaluation procedures, H^{β} with $\beta \geq 0$, is the family of high-impact indicators introduced in Albarrán et al. (2011a). It reads as follows:

$$FGT^{\beta}(u) = \Sigma_{r>z} c_r^u \left(\frac{r-z}{z}\right)^{\beta}. \tag{4}$$

This is an additive rule with θ_0 = publications with citations in the interval [0,z], θ_1 = publications with $(z+1)$ citations, . . ., θ_v = publications with $(z+v)$ citations, . . ., i.e., as many categories as distinct citations above z, plus one, the category formed by all citations below z.[7] Then, $\omega_0^{FGT\beta} = 0$, and$\omega_g^{FGT\beta} = \left(\frac{g-z}{z}\right)^{\beta}$ for $g \geq 1$.

Given any CCL z, let $k\%$ be the proportion represented by the high-impact publications in citation distribution D. Then, for $\beta = 0$, the first member of the FGT family coincides with the *Top k%* indicator. For $\beta = 1$, the second member of the FGT family is called the *Average of the Normalized citation Gaps* (*ANGz* or *ANG k%*) indicator. Note that, following the terminology used in economic poverty analysis, any *Top k%* indicator only captures the *incidence* of the high-impact aspect of a citation distribution, whereas the *ANG k%* indicator captures both *the incidence and the intensity* of this phenomenon.

3. The HV evaluation procedure

3.1. The evaluation formula

Let $\Theta = \left\{\theta_g\right\}_{g=1}^{G}$ be a given selection of categories, as in the former section. For any unit u, let $D_{\Theta}(u) = (c_1^u, \ldots, c_G^u)$ stand for the vector of publication shares of unit u into the different categories. The basic principle of the *HV* procedure to assess the citation impact of a research unit refers to the probability that a representative publication of that unit is in a higher category than a representative publication of another unit. We shall refer to those probabilities as *domination probabilities*. For any pair of units, u and s, we thus define the scalar $p_{us} \in [0, 1]$ as the probability that a publication in unit u belongs to a higher category than a publication in unit s. If categories are ordered in an ascending order, we can easily calculate that probability as follows:

$$p_{us} = c_u^G \left(c_s^{G-1} + \ldots + c_s^1\right) + c_u^{G-1}\left(c_s^{G-2} + \ldots + c_s^1\right) + \ldots + c_u^2 c_s^1.$$

Similarly, p_{su} denotes the probability that a publication in unit s belongs to a higher category than a publication in unit u. By construction, the probability of two publications belonging to the same category is given by $e_{us} = e_{su} = 1 - p_{us} - p_{su}$.

We now define the *relative advantage* of research unit u with respect to unit s, RA_{us}, as the ratio between the probability that u dominates s and the sum of the probabilities that u is being dominated by some other unit. That is,

$$RA_{us} = \frac{p_{us}}{\sum_{k \neq u} p_{ku}}.$$

Whenever there are only two groups, $RA_{us} = p_{us}/p_{su}$. The *overall relative advantage* of unit u, RA_u, can thus be described by a weighted average of the relative advantage of u with respect to all other research units. That is,

$$RA_u = \sum_{s \neq u} \lambda_s RA_{us}.$$

Since the weights in this average reflect the relevance of the different units, it is only natural to require choosing them consistently, that is $\lambda_s = RA_s$. In this way, each unit enters the evaluation of the overall relative advantage of any other unit with its own overall relative advantage. That means that the evaluation of a problem P is given by a vector$v \in \mathbf{R}_{+}$, such that:

$$v_u = \sum_{s \neq u} v_s RA_{us} = \frac{\sum_{s \neq u} v_s\, p_{us}}{\sum_{k \neq u} p_{ku}}, \quad u,\ s,\ k = 1, 2, \ldots, U. \tag{5}$$

Herrero & Villar (2013) show that such a vector, called the **worth**, always exists, and it is strictly positive and unique (up to normalization) under very general conditions.

Notice that the worth vector has a degree of freedom, so that we can freely normalize its values. The easiest (and most natural way) of comparing the performance of the different units among themselves, and with respect to the field as a whole, is by adding to the problem $P = \left\{D(u)\right\}_{u=1}^{U}$ the world field citation distribution D as if the field as a whole were an additional unit, and normalize the worth vector of this problem by making the worth of the whole field equal to one. In this way, we are immediately informed not only about the order of the different units, but also whether a certain unit is above or below the "world's average".

Some relevant features of this evaluation procedure are the following:

[7] The categories above the CCL can be written as θ_v= publications with ($int(z) + v$) citations, where $int(z)$ is the integer part of z, to take into account that the percentile defining the CCL can be in practice a non-integer number.

1 For any given problem, we obtain a *complete ranking* of the units, as well as a *cardinal evaluation* of their relative performance.
2 Given a problem *P*, a unit u is *irrelevant* whenever it happens that for any other unit s different from u, we have $p_{us} = 0$. If a unit is irrelevant, then the corresponding component of the worth vector is zero.
3 The worth *is not an additive rule*. That is, we cannot interpret that the worth provides an evaluation in which each category has associated a weight endogenously determined. As shown in Example 1 in the Appendix A, in general it is not possible to express the worth as a weighted sum of the different categories. This is due to the multilateral nature of the comparison between individual units.[8]
4 In the case of only two categories, though, the *rankings provided by the HV k% and the Top k% procedures coincide, even if the corresponding cardinal evaluations do not*. Indeed, without loss of generality, if we have U units so that the proportion of publications in the top k percentiles are, respectively, $c_1^1 \geq c_1^2 \geq \ldots \geq c_1^U$, the worth vector of these units is given by

$$v_u = \frac{\left(1 - c_1^U\right) c_1^u}{c_1^U \left(1 - c_1^u\right)}, forallu \neq U; v_U = 1, \tag{6}$$

so that $v_1 \geq v_2 \geq \ldots \geq v_U$. Thus, both procedures order the units in the same way. Furthermore, under a common normalization, the *components of the worth vector grow faster* than the values attached to the units by the traditional dichotomous procedure, which in this case are

$$Top_u = \frac{c_1^u}{c_1^U}, foru \neq U, Top_U = 1.$$

It is also interesting to note that the difference between the worth values and the *Top k%* values vanishes when k goes to zero. Example 2 in the Appendix A illustrates this fact.

Remark 1. The computation of the worth vector can be directly obtained through a friendly and freely available algorithm, hosted in the website of the Instituto Valenciano de Investigaciones Económicas, Ivie (http://www.ivie.es/valoracion/index.php). *The worth can be obtained directly from the matrix of relative frequencies of publications by categories and can be plugged into the algorithm as an excel table, thus saving much time and effort. By default the algorithm normalises the worth vector by making the mean value equal to one and requires ordering categories from best to worst.*

Using the categories presented before for the selected additive rules, we find the following examples of the *HV* evaluation procedures:

1 The *RCR* indicator corresponds to the case in which each point in the support of the field citation distribution is considered as a different category. The corresponding partition in the *HV* procedure will be called the *HV max* procedure.
2 The *HV100* procedure considers the categories corresponding to the 100 percentiles in *D*, as in the *C100* procedure.
3 The *HVNSF6 procedure* is the counterpart of the *NSF6* procedure
4 The *HV k%* procedure is the counterpart of the *Top k%* indicator. As mentioned before, the rankings provided by the *HV k%* and the *Top k%* procedures coincide, even though the corresponding evaluations do not.
5 Finally, given a CCL z, we had the *ANG k%* evaluation procedure with the following categories: θ_0 = publications with citations in the interval [0,z] or in the bottom (100–k) percentiles, θ_1 = publications with $(z+1)\}$ citations, . . ., θ_v = publications with $(z+v)$ citations, . . ., i.e., as many categories as distinct citations above z, plus one, the category formed by all citations below z. We call *HV k% min-max* the procedure associated to the corresponding set of categories. In the terminology used in Section II.2.C for the FGT indicators, the *HV k%* procedure only compares the *incidence* of the high-impact aspect of the different citation distributions, whereas the *HV k% min-max* procedure compares both *the incidence and the intensity* of the high-impact phenomenon.

Example 3 in the Appendix A illustrates the results of applying some of the procedures described in points 1, 4 and 5 above on a given dataset.

3.2. The HV procedure as a tournament

The worth can also be regarded as the stationary distribution of a Markov process induced by a comparative evaluation of the performance of the different research units in terms of the following tournament. Start by selecting arbitrarily one research unit, say unit u. Then select randomly another unit to compete with unit u, say unit s. Now pick randomly a publication from each of those units and confront them in terms of the category to which they belong. If the publication of one of those units belongs to a higher category than the other, then that unit remains in the contest while the other one is dismissed. If both publications belong to the same category, then the one initially chosen (unit u in this case) remains in the

[8] Note that the weights that appear endogenously in (5) refer to research units and not to categories.

contest, to be confronted with another unit s randomly selected. Two new publications are chosen at random from each of those units and are confronted as before.

By repeating indefinitely this process, we find that the probability that unit u remains in the contest is given by:

$$\frac{\sum_{s\neq u}(p_{us}+e_{us})}{G-1}$$

That is, the sum of the probability that unit u gets better or equal results than all other units times the probability of being chosen at random. Similarly, the probability of unit u being beaten by unit s is given by:

$$\frac{p_{su}}{G-1}$$

That is, the probability that unit s gets a better outcome than unit u times the probability of s being chosen.

This protocol defines, therefore, a Markov process whose stochastic matrix is:

$$M=\frac{1}{G-1}\begin{pmatrix}\sum_{s\neq 1}(p_{1s}+e_{1s}) & p_{12} & \dots & p_{1G}\\ p_{21} & \sum_{s\neq 2}(p_{2s}+e_{2s}) & \dots & p_{2G}\\ \dots & \dots & \dots & \dots\\ p_{G1} & p_{G2} & \dots & \sum_{s\neq G}(p_{Gs}+e_{Gs})\end{pmatrix}.$$

This matrix has a positive dominant eigenvector, which corresponds to a stationary distribution, whose u-th component is:

$$v_u=\frac{\sum_{s\neq u}p_{us}v_s}{\sum_{s\neq u}p_{su}}.$$

That is, the corresponding worth. This structure explains why the worth is so well behaved and its computation presents no particular difficulty. We can also immediately deduce the uniqueness (and strict positiveness) of the worth vector when matrix M is irreducible. A sufficient condition for that to happen is that $p_{us}>0$ for all u, s. That is, when there are no irrelevant units.

In this context the worth tells us the "number of times" that a research unit beats another (more precisely, the fraction of time that, in the long run, each unit will keep competing in the tournament).[9]

4. Empirical results

4.1. Data

The data considered here refer only to research articles or, simply, articles. We begin with a set of 4,472,332 distinct articles published in the period 1998–2003, as well as the citations they receive using a common, five-year citation window for each year in that period. Each of these articles is assigned by Thomson Reuters to one of 22 broad fields. We consider 38 countries that have published at least 10,000 articles in all sciences in the period 1998–2003, plus Luxembourg, which is included in order to cover the 15 countries in the European Union before the accession in 2004, as well as one residual geographical area for the Rest of the World (RW hereafter). Therefore, the total number of research units is $U=40$.

Articles are assigned to countries according to the institutional affiliation of their authors on the basis of what had been indicated in the by-line of the publications. So far we have assumed that there is no co-authorship, so that each article belongs to a single research unit. However, international cooperation, namely, the existence of articles written by authors belonging to two or more countries poses a technical difficulty that, as is well known, admits different solutions (Waltman & Van Eck, 2015). In this paper, we follow a multiplicative strategy that extends as much as necessary the citation distributions of the research units in our dataset.[10] In this way we arrive at what we call the *geographical extended count* consisting of 5,452,445 articles, a total which is 21.9% larger than the original 4.5 million articles. The distributions of the number of articles by field in the original and the geographically extended count are very similar (for details, see Albarrán, Perianes-Rodriguez, & Ruiz-Castillo, 2015).

Columns 1 and 2 in Table A1 in the Appendix A include the distribution of the total number of articles by country. The U.S. publishes 26.7% of the total, while the European Union is responsible for approximately one third. The remaining 23

[9] There is a variant of this procedure, called the *balanced worth* (see Herrero & Villar, 2017), in which the probability of remaining in the tournament is equally split when there is a tie. This cancels the extra prize given here to the stronger units, as those winning more often will still have more time competing. Yet we understand that the premium is interesting in this evaluation context.

[10] Perianes-Rodriguez & Ruiz-Castillo (2015) find that a move from the alternative fractional approach to the multiplicative approach does not cause dramatic differences in co-authorship patterns and citation impact values. Nevertheless, *ceteris paribus*, the gainers with a move from the fractional to the multiplicative approach are characterized by (i) a low co-authorship rate for citation distributions as a whole, but a high co-authorship rate in the upper tail of these distributions; (ii) a low citation impact performance, and (iii) a small number of solo articles.

countries and the RW publish about 40% of the total. As indicated in the Introduction, for the sake of parsimony we only present the data for the following four fields: Clinical Medicine, Physics, Engineering, and Economics & Business. Clinical Medicine is the field with the highest number of published articles; Physics and Engineering represent two of the main fields among the natural sciences with different characteristics, and Economics & Business is the only separately identified field among the social sciences. Consequently, columns 3–6 in Table A1 include the distribution by country of the total number of articles in each of these four fields.

In each field, we study the following ranking procedures introduced in Section II.2 and III.3, namely:

(i) RCR and *HV max*;
(ii) C100 and *HV100*;
(iii) NSF6 and *HVNSF6*;
(iv) Top k% and *HV k%*;
(v) ANG k% and *HV k% min-max*.

Given a field citation distribution and a percentile k, we fix the CCL z so that high-impact articles coincide with those above the k-th percentile.

Remark 2. The use of percentiles in practice and the construction of the corresponding categories is well known in the literature and is explained in depth, among others, by Waltman and Schreiber (2013)*. Here we define the percentile categories using the standard statistical approach used by the Science and Engineering Indicators report of the* National Science Board (2012)*. In this approach, the set of, say, top 10% publications is defined such that it includes at most 10% of the publications in a field.*

4.2. Results

As far as ranking procedures are concerned, in groups (iv) and (v) above we consider the values $k\% = 10\%$, 1%, so that in each of these two groups there are four procedures. Since groups (i) to (iii) involve two procedures each, there are 14 procedures altogether. Index values and country ranks for the aforementioned 14 procedures, within the four fields, are presented in Tables SM1 and SM2 in the Supplementary Material. Country index values are presented relative to the field value, so that the value one can serve as a benchmark for evaluating research units in the usual way.

Comparing alternative evaluation procedures usually involves two key elements: (1) the extent of re-rankings, and (2) the importance of cardinal differences between those evaluations (see *inter alia* Waltman et al., 2012, and Ruiz-Castillo & Waltman, 2015). As a first approximation, the first aspect might be partially revealed by rank correlation coefficients, and also by the Ulam distance between rankings. As discussed by Gordon (1979), the Ulam distance between rankings can be interpreted as the number of elements (countries in our case) that have to be changed in two ranking vectors in order to have full agreement between them.[11] Consequently, the value of the Ulam distance regarding our evaluation problem lies between 0 and 39. Kendall rank correlation coefficients can be found in Table SM3 in the Supplementary Material, while Ulam distances can be found in Table 1.

Here we are mostly interested in the following questions:

1 How different are the rankings delivered by the additive procedures and the corresponding *HV* evaluations?
2 How different are the rankings as a function of the number of categories (the fineness of the grid) and their associated scores? We are interested, in particular, in comparing indicators of incidence with indicators of incidence and intensity of the high-impact phenomenon.

We find that Kendall correlation coefficients between the *Top k%* and *HV k%* procedures for $k\% = 10\%$, 1%, are both equal to one in all fields and that the Ulam distance between them is zero, as it should be, since we have already mentioned that they provide identical rankings.

There are some other key comparisons characterized by extremely high Kendall coefficients and low Ulam distances, indicating that, in those cases, there are few re-rankings. For instance, among the indicators that look at the whole distribution (*RCR, HV max, C100* and *HV100*), the last three yield extremely similar rankings. Also, the rankings of the *HV k% min-max* are almost identical to the corresponding *Top k%* and *HV k%*. This implies that, in our data, the additional flexibility provided by an extension of the category set in the upper tail of the field citations distribution in the *HV k% min-max* procedure plays a limited role in ranking the countries.

In general, although the rank correlations are relatively high among all procedures, the Ulam distances often show a substantial number of re-rankings, revealing that the Ulam protocol distinguishes more finely between procedures than the Kendall coefficient. Consider the following three cases. Firstly, when comparing indicators that use several categories across

[11] In order to compute the Ulam distance, we have used several components from the Fortran 90 library SUBSET by John Burkardt (Florida State University), available at his website https://people.sc.fsu.edu/~jburkardt/f_src/subset/subset.html.

Table 1
(Ulam) Distance between rankings according to each procedure.

Part A. Clinical Medicine.

	RCR	HV max	C100	HV100	NSF6	HVNSF6	Top 10%	HV	HV10% min-max	ANG	Top	HV	HV1% min-max	ANG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1. RCR	0													
2. HV max	17	0												
3. C100	17	3	0											
4. HV100	17	0	3	0										
5. NSF6	13	13	12	13	0									
6. HVNSF6	17	10	8	10	12	0								
7. Top 10%	14	22	22	22	19	20	0							
8. HV10%	14	22	22	22	19	20	0	0						
9. HV10% min-max	14	22	22	22	19	20	1	1	0					
10. ANG10%	19	22	23	22	22	21	17	17	17	0				
11. Top 1%	22	25	25	25	22	24	21	21	21	15	0			
12. HV1%	22	25	25	25	22	24	21	21	21	15	0	0		
13. HV1% min-max	22	25	25	25	22	24	21	21	21	15	0	0	0	
14. ANG1%	25	27	27	27	26	24	26	26	25	23	20	20	20	0

Part B. Economics.

	RCR	HV max	C100	HV100	NSF6	HVNSF6	Top 10%	HV	HV10% min-max	ANG	Top	HV	HV1% min-max	ANG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1. RCR	0													
2. HV max	18	0												
3. C100	18	4	0											
4. HV100	18	0	4	0										
5. NSF6	15	15	14	15	0									
6. HVNSF6	18	14	12	14	13	0								
7. Top 10%	21	22	22	22	22	23	0							
8. HV10%	21	22	22	22	22	23	0	0						
9. HV10% min-max	21	21	21	21	22	22	2	2	0					
10. ANG10%	22	25	25	25	25	24	23	23	23	0				
11. Top 1%	25	26	26	26	25	27	25	25	25	22	0			
12. HV1%	25	26	26	26	25	27	25	25	25	22	0	0		
13. HV1% min-max	25	26	26	26	25	27	25	25	25	22	0	0	0	
14. ANG1%	27	28	29	28	27	28	30	30	30	29	16	16	16	0

Part C. Engineering.

	RCR	HV max	C100	HV100	NSF6	HVNSF6	Top	HV	HV10% min-max	ANG	Top	HV	HV1% min-max	ANG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1. RCR	0													
2. HV max	19	0												
3. C100	20	3	0											
4. HV100	19	0	3	0										
5. NSF6	14	16	17	16	0									
6. HVNSF6	18	9	10	9	16	0								
7. Top 10%	12	21	21	21	17	21	0							
8. HV10%	12	21	21	21	17	21	0	0						
9. HV10% min-max	12	21	21	21	17	21	4	4	0					
10. ANG10%	21	25	25	25	23	23	18	18	17	0				
11. Top 1%	22	23	22	23	20	22	22	22	22	18	0			
12. HV1%	22	23	22	23	20	22	22	22	22	18	0	0		
13. HV1% min-max	22	23	22	23	20	22	22	22	22	18	1	1	0	
14. ANG1%	26	29	29	29	28	29	29	29	28	26	24	24	24	0

Part D. Physics.

	RCR	HV max	C100	HV100	NSF6	HVNSF6	Top	HV	HV10% min-max	ANG	Top	HV	HV1% min-max	ANG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1. RCR	0													
2. HV max	23	0												

Table 1 (*Continued*)

Part D. Physics.

	RCR	HV max	C100	HV100	NSF6	HVNSF6	Top	HV	HV10% min-max	ANG	Top	HV	HV1% min-max	ANG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
3. C100	23	0	0											
4. HV100	23	0	0	0										
5. NSF6	17	13	13	13	0									
6. HVNSF6	23	5	5	5	14	0								
7. Top 10%	17	22	22	22	17	22	0							
8. HV10%	17	22	22	22	17	22	0	0						
9. HV10% min-max	16	23	23	23	19	23	3	3	0					
10. ANG10%	22	25	25	25	25	25	20	20	20	0				
11. Top 1%	22	24	24	24	23	24	22	22	21	20	0			
12. HV1%	22	24	24	24	23	24	22	22	21	20	0	0		
13. HV1% min-max	22	24	24	24	23	24	22	22	21	20	0	0	0	
14. ANG1%	23	27	27	27	25	26	25	25	24	22	21	21	21	0

Table 2
The variability of the 14 procedures in terms of the range, the ratio Max/Min, and the coefficient of Variation (CV). Selected scientific fields.

	Clinical Medicine			Physics			Engineering			Economics & Business		
	Range	Max./Min.	CV	Range	Max./Min.	CV	Range	Max./Min.	CV	Range	Max./Min.	CV
1. RCR	1.06	5.43	0.334	1.12	3.55	0.264	0.99	2.84	0.217	1.11	9.39	0.335
2. HV max	1.26	8.48	0.331	0.99	2.92	0.243	1.08	3.26	0.212	1.08	9.40	0.330
3. C100	0.79	3.15	0.184	0.47	1.65	0.116	0.47	1.65	0.095	0.67	2.62	0.167
4. HV100	1.26	8.49	0.331	0.99	2.93	0.243	1.08	3.26	0.212	1.08	9.40	0.331
5. NSF6	0.51	1.81	0.141	0.46	1.62	0.113	0.44	1.56	0.094	0.52	1.92	0.129
6. HVNSF6	1.34	9.43	0.398	1.22	3.92	0.301	1.20	3.65	0.243	1.16	13.79	0.374
7. Top 10%	1.27	11.03	0.469	1.85	*	0.443	1.51	5.17	0.352	1.38	*	0.496
8. HV10%	1.34	12.62	0.499	2.03	*	0.480	1.70	5.98	0.387	1.44	*	0.522
9. HV10% min-max	1.31	11.96	0.485	1.96	*	0.464	1.63	5.68	0.374	1.42	*	0.514
10. ANG10%	1.45	34.26	0.519	2.14	*	0.509	1.65	5.69	0.436	1.50	*	0.664
11. Top 1%	1.58	*	0.578	2.43	*	0.565	2.35	*	0.587	1.61	*	1.062
12. HV1%	1.58	*	0.581	2.46	*	0.570	2.38	*	0.593	1.62	*	1.066
13. HV1% min-max	1.58	*	0.579	2.44	*	0.567	2.37	*	0.591	1.61	*	1.065
14. ANG1%	1.76	*	0.604	2.37	*	0.692	2.09	*	0.600	2.96	*	1.440

*Since the minimum value is zero, the ratio Max/Min becomes infinite.

the whole distribution (*RCR, C100, NSF6* and the corresponding *HV* indicators) with indicators that focus on the upper tail (*Top k%, ANG k%* and the corresponding *HV* indicators), Ulam distances go from not much more than 10 up to well above 20. This highlights that focusing on the upper part of the field citations distribution has important implications for the evaluation of countries' performance. Secondly, the large differences in Ulam distances between indicators based on the upper 10% or the upper 1% tail imply again that becoming even more selective with countries' performance gives rise to substantial re-rankings. Thirdly, in spite of the similarity of the rankings generated by additive procedures and *HV* indicators when they are based on the same categories, there are a few exceptions worth noting. In particular, there are important differences between the rankings provided by the *NFS6* and the *HVNFS6* procedures. This indicates that the choice of weights in the case of more than two categories can have important consequences.

Finally, the rankings provided by the *RCR* indicator and the *ANG k%* indicators are also very different from their corresponding *HV* indicators: their Ulam distances are around 20 (often above) for the *RCR* indicator, and well above 20 for the *ANG k%* indicators. Notice also that these indicators have another characteristic in common: they are potentially very sensitive to extreme observations. Therefore, they should be treated with caution.[12]

Next, we move to discuss cardinal differences between the units' index values. For all procedures in the four fields, Table 2 presents measures of outcome ***variability***: the range, the ratio Max/Min, and the coefficient of variation (CV).

Our discussion focuses on three questions.

[12] The case of *ANG 1%* in the field of Economics is particularly informative. While common wisdom in the profession points to the U.S. as the leading country by far, the U.S. would be third in the ranking according the *ANG 1%* (Table SM3 in the Supplementary Material) below Switzerland (first) and, very surprisingly, Mexico (second). Of course, the small country effect and international collaboration can partly explain this, but these two issues also potentially affect other indicators. Actually, Mexico goes up a lot in any ranking focusing on the upper 1% tail. But since those indicators are not so extremely sensitive to extreme observations, they deliver more sensible rankings.

1. Do the *HV* procedures exhibit a greater variability than the percentile procedures under the same category set? In other words, do we find that variability is ordered as follows:

$$C100 < HV100?$$

$$NSF6 < HVNSF6?$$

$$Top10\% < HV10\%?$$

$$Top1\% < HV1\%?$$

Consider first the dichotomous case, i.e. the last two rows. As expected, the answer to this key question is in the affirmative in the four fields (rows 7 and 8, and 11 and 12 in Table 2). Note that, as *k* decreases, the difference in variability between the *HV k%* and *Top k%* procedures becomes negligible, as expected (see Example 2 in the Appendix A).

On the other hand, when there are more than two categories, *HV* procedures clearly discriminate more than percentile-based procedures (rows 3 and 4, and 5 and 6 in Table 2).

2. Does the variability increase when we move to procedures based on indicators more and more focused on the upper tail of the citation distribution? In other words, do we find the variability ordered as follows:

$$C100 < NSF6 < Top10\% < Top1\%? \tag{7}$$

$$HV100 < HVNSF6 < HV10\% < HV1\%? \tag{8}$$

$$ANG10\% < ANG1\%? \tag{9}$$

In the first set of scoring rules (sequence 7), the *C100* and *NSF6* procedures are conceptually very different. We find that, contrary to what one may have expected, the variability of the *C100* procedure is greater than that of the *NSF6* procedure. This is due to the effect of the weighting system, which exhibits a much larger spread in the *C100* case (i.e. a ratio of 100/1 in the first case and a ratio of 6/1 in the second). The contrary happens if we compare the *C100* and the *Top 10%* (and the *Top 1%*) procedures in all fields. The increase in the variability, jointly with the relatively low rank correlation and the high Ulam distances discussed before, clearly confirms the existence of differences in this sequence (rows 3, 7, and 11 in Tables 1 and 2).

In the *HV* case (sequence 8), we reach a different result in all fields. Although the *HV100* and the *HVNSF6* rankings are rather similar, the *HVNSF6* procedure has a greater variability than the *HV100* procedure. Thus, sequence 8 is clearly established (rows 4, 6, 8, and 12 in Table 2). Finally, as far as sequence 9 is concerned, the variability also increases as we move towards the very upper tail of citation distributions in all fields (rows 10 and 14 in Table 2).

3. Do the incidence indicators have a smaller variability than the indicators capturing both the incidence and the intensity of the high-impact phenomenon? In other words, do we find that the variability is ordered as follows:

$$Top10\% \ < \ ANG10\%?$$

$$Top1\% \ < \ ANG1\%?$$

$$HV100 \ < \ HVmax?$$

$$HV10\% \ < \ HV10\%minmax?$$

$$HV1\% \ < \ HV1\%minmax?$$

We find that the variability is moderately but clearly higher in the *ANG k%* than the *Top k%* indicators. On the contrary, the variability of the *HV100* and *HV max* procedures is indistinguishable (rows 4 and 2 in Table 2), whereas in the two remaining cases the variability of the incidence procedures is even slightly greater than that of the *min-max* alternatives (rows 8 and 9, and 12 and 13 in Table 2). Discriminating between publications with different citations within percentile rank classes in the *HV100*, *HV 10%*, and *HV 1%* procedures has practically no consequences.

5. Discussion

Based on the ideas presented in Herrero & Villar (2013), in this paper we have introduced a new procedure for the evaluation of research units that can be regarded as an alternative to size-independent additive methods. The *HV* index evaluates the scientific influence of research units in terms of the likelihood of getting publications with higher citations. The key value judgement is the comparison of any two units in terms of the probability that a random extraction from one of them yields a publication with a higher citation level than one random extraction from the other. The index derives from a consistent application of this notion for the entire set of units participating in the evaluation problem.

A useful way of assessing the interest of a new procedure is to compare it with some relevant alternatives. Here we have selected three types of procedures that also evaluate citations using the relative citations distributions embedded in a set of categories. All those procedures can be described as weighted sums, where the weights correspond to the importance given to each category.

In the previous Section, we have compared from an empirical perspective seven additive procedures representing current practice and the corresponding *HV* alternatives under the same set of categories. Let us now comment from a conceptual viewpoint on the main features that additive and *HV* procedures have in common, and the key aspects that separate them.

All procedures rely on a previous decision −expressing our value judgements–, about the assignment of publications into categories in such a manner that each category includes all publications considered to have the same citation merit. As all publications within a category are indistinguishable, the more generic the category is, the less attention we pay to individual differences (and vice-versa). Our view on how many categories to distinguish and how inclusive they are decisively conditions the evaluation exercise. Changes in the definition of categories affect relative citations frequencies, and hence the final result.

Once the categories have been established, additive procedures and the *HV* evaluation follow different paths. Additive procedures attach a weight to each category, which expresses how important it is, and then obtain the evaluation as a weighted sum of the shares of citations into the different categories. Consequently, once the categories have been defined, the weights fully determine the evaluation. The problem with this approach is that the choice of weights can be rather arbitrary, in which case the same applies to the whole evaluation. To see this, notice that changing the weights of any additive procedure may substantially alter the evaluation (e.g., by continuity, any of the rules in the percentile approach can be approximated arbitrarily from the *C100* by conveniently adjusting the weights).

The *HV* index does not require any external weighting system. Using a different evaluation approach it provides a *relative assessment* of the different research units in terms of the likelihood of getting publications with higher citations (relative here in the sense that the value attached to each unit depends on all the units with which it is compared). The multilateral nature of these comparisons and the evaluation principle lead to a cardinal, complete and transitive ranking, as it is the case in the additive approach.

From the above discussion it follows that the valuation of a unit in the additive methods is independent of the valuation of any other unit, whereas the valuation of a unit by the *HV* method depends on all other units. This type of structural difference appears in different evaluation exercises. Think, for instance, of the way of ranking athletes in a decathlon competition at the Olympic Games, and the way of ordering soccer teams in the European national leagues. In the first case, each of the disciplines is graded separately and the athletes' scores depend on their individual performance. Overall individual scores are obtained by adding up the outcomes of the different disciplines, and the athletes are ordered according to the aggregate outcome. But if a certain athlete disappears from the competition, the score of the remaining athletes does not vary. In the soccer leagues all teams in the same division compete twice with each other and the final ranking takes into account the results of all pairwise matches. The evaluation of a team *vis-a-vis* another depends upon their performance against all the competitors, and upon the competitors' performance. Because of that, if after the league ends a certain team is eliminated for some external reason, the ranking of the remaining teams may vary.

Indeed, the *HV* method solves the evaluation of all research units *simultaneously* instead of unit by unit. One may argue that this makes the evaluation less transparent and harder to compute. This is not the case. On the one hand, the evaluation turns out to be obtained as the dominant eigenvector of a suitable stochastic matrix, so that the solution is well defined and conceptually well grounded. On the other hand, there is a free online algorithm that immediately solves the problem. This algorithm allows interested parties to perform a sensitivity analysis regarding alternative specifications of the categories at no cost.

Needless to say, obtaining relative evaluations in terms of eigenvectors is far from new. This technique appears in a variety of problems, such asthe ranking of income distributions (Yanoletzky, 2012), the analysis of segregation and discrimination (Echenique and Fryer, 2007; Grannis 2002), the social theory of voting (Chabotarev & Shamis, 1998; Laslier, 1997), the evaluation of scientific influence of journals (Bergstrom, 2007; Laband & Piette, 1994; Liebowitz & Palmer, 1984; Palacios-Huerta & Volij, 2004; Pinski & Narin, 1976; West, Bergstrom & Bergstrom, 2010), the analysis of network structures (Bergstrom, 2007; DeGroot, 1974; Golub & Jackson, 2000; Newman, 2003; Rosvall and Bergstrom, 2007; West et al., 2010), the allocation of scores in tournaments (Daniels, 1969; Keener, 1993; Moon & Pullman, 1970; Slutzki and Volij, 2005), or even the Google page-rank (Page, Brin, Motwani, & Winograd, 1998; Slutzki and Volij, 2006). Each of those problems requires building a particular matrix that properly captures the relevant relationships between the outcome distributions. Tailoring the procedure to the problem is what makes of this general approach a suitable evaluation protocol. The *HV* formula fits perfectly well the evaluation of the scientific influence of research units as the stochastic matrix that yields the evaluation is made of the probabilities that a publication of a given unit dominates that of any other.

It might be tempting to think of the *HV* method as providing an endogenous way of attaching weights to the different categories or citations levels, so that the result is actually a blurred weighted average. This is not the case. Example 1 in the Appendix A shows that it might be impossible to recover the HV evaluation as a weighted sum.

Let us point out that there is no best ranking procedure for the evaluation of citation impact (for recent discussions, see Bouyssou and Marchant, 2014, and Waltman, 2016). Different ways of selecting categories reflect different conceptions of the relevance of citations. Once this has been established, the key question is whether there is a rationale for the choice of weights, or at least a general agreement on why those weights are the proper rates of substitution between categories.

When this is the case, additive procedures provide simple and transparent formulae to make the evaluation. Yet many disagreements concerning the ranking of institutions derive, precisely, from the different visions about how publications in different percentiles should be pondered. When there are insufficient reasons to choose a particular weighting system, or when people disagree about the choice of weights, the *HV* method provides a sound alternative that avoids having to decide on how to ponder different categories.

Note that the skewness of citation distributions in all sciences has recently led practitioners towards methods that focus on the upper tail of such distributions. We have confirmed that this focus leads to substantial changes in the evaluation of 40 countries. Yet, what the CCL or the *k%* should be in order to represent excellence in citation impact is not obvious at all. In this situation, the Leiden Ranking has temporarily found an interesting solution: compute *Top k%* indicators for *k%* = *10%* and *1%*. One may well consider reasonable to define three categories consisting of the bottom 90 percentiles, the next nine percentiles, and the top last percentile.[13] In this case and, generally, whenever a user believes that based on fundamental value judgments it is best to have more than two categories, we may conclude that the *HV* procedure is a good choice without any *a priori* weighting scheme.

Finally, we address three possible directions for further research. Firstly, we have studied a large dataset but a relatively small number of research units. Before we can give full credit to our empirical conclusions, we must experiment with other large datasets and a more extensive list of research units. Secondly, we know that additive indicators might be dominated, in some cases, by a few highly cited publications[14] (even if this depends on the convexity of the indicator). Instead, percentiles or percentile rank classes are scarcely affected by extreme observations (see *inter alia* Bornmann and Marx, 2013). It should be noted that *HV* ranking procedures would behave as the corresponding percentile based indicators in this respect. In any case, the robustness of indicators to extreme observations is an empirical matter beyond the scope of this paper. Last, but not least, in this paper we have limited ourselves to the analysis of fields in isolation. The theoretical problem of aggregating fields to provide a joint evaluation of research units (countries) is solved in the case of additive rules, but is still an open question in the case of *HV* indicators. This is also left for further research.

Author contributions

Pedro Albarrán: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Carmen Herrero: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Javier Ruiz-Castillo: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Antonio Villar: Conceived and designed the analysis, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Acknowledgements

The dataset has been acquired thanks to a grant by Santander Universities Global Division of *Banco Santander*. P. Albarrán acknowledges financial support from the Spanish MEC through grant ECO2014-58434-P. C. Herrero acknowledges financial support from Generalitat Valenciana under PROMETEO 2013/037, and from the Spanish MEC under project ECO2015-65820-P. J. Ruiz-Castillo acknowledges financial support from the Spanish MEC through grant ECO2014-55953-P, as well as grant MDM 2014-0431 to his Departamento de Economía, and grant MadEco-CM (S2015/HUM-3444) from the Comunidad Autónoma de Madrid. A. Villar acknowledges financial support from the Spanish MEC under project ECO2015-65408-R (MINECO/FEDER). Thanks are due to two anonymous referees for very helpful comments. All remaining shortcomings are the authors' sole responsibility.

Appendix A.

Example 1. Let us consider three units, each with 18 papers, and three categories. The first unit has 9 papers in category 1 and 9 papers in category 3; Unit 2 has all the papers in category 2, and unit 3 has 6 papers in each category. According to the HV procedure, the three units have the same value (the reason is that, in this example, $p_{ij} = p_{ji}$ for all $i, j = 1, 2, 3$).Thus, if we weight the categories with scores 0, α, β we obtain that $\beta = 2\alpha$. Now, take another problem in which again we have three units *a*, *b*, and *c*, and the same three categories as before. The three units have ten papers each. Unit *a* has five papers

[13] Another interesting example leading to three categories is Rodriguez-Navarro (2011) who looks for an indicator defined on the very upper tail of citation distributions capable of predicting the number of Nobel Prizes in Chemistry, Physics, and Physiology/Medicine. This author proposes a scoring rule with three categories of percentiles, namely, [0.99), [99,99.9) and [99.9, 100] for which, using regression analysis, he suggests weights 0, 1 and 15, respectively.

[14] The case of the University of Göttingen, which was ranked second in the 2011/2012 edition of the Leiden Ranking according to the *MNCS* indicator on the strength of a single highly cited publication, is cited as a good example of this problem (Waltman et al., 2012). For the influence of extreme observations see *inter alia* Li & Ruiz-Castillo (2014).

in category 1 and five papers in category 3, unit *b* has 3 papers in category 1 and seven papers in category 2, and unit *c* has ten papers in category 2. According to the *HV* evaluation, unit *c* goes first, the next is unit *a*, and the last is unit *b*. But for this to happen with the scores previously obtained, 2α > β, contradicting the previous finding.

Example 2. Consider three units so that for *k* = *10%*, present the following distributions:

$$D_{10\%}^{1} = (0.5, 0.5)\,;\, D_{10\%}^{2} = (0.9, 0.1)\,;\, D_{10\%}^{3} = (0.99, 0.001)\,.$$

And for *k* = *1%*, the citation distributions are as follows:

$$D_{1\%}^{1} = (0.95, 0.05)\,;\, D_{1\%}^{2} = (0.99, 0.01)\,;\, D_{1\%}^{3} = (0.999, 0.001)$$

In both cases, the values of the *Top k%* relative to the third unit are $d_1 = 50$; $d_2 = 10$; $d_3 = 1$. Nonetheless, the *HV* values are as follows:

$$\upsilon_1^{10\%} = \frac{0.99}{0.5}50; \upsilon_2^{10\%} = \frac{0.99}{0.9}10; \upsilon_3^{10\%} = 1$$
$$\upsilon_1^{1\%} = \frac{0.999}{0.95}50; \upsilon_2^{1\%} = \frac{0.999}{0.99}10; \upsilon_3^{10\%} = 1.$$

Example 3. Suppose that a field citation distribution is given by

$$\begin{aligned}
&Q(1) = Q(3) = Q(4) = Q(5) = Q(14) = Q(15) = Q(20) = Q(25) = Q(40) = Q(50) = Q(100) = 1;\\
&Q(2) = Q(6) = Q(7) = Q(8) = 2;\\
&Q(0) = 4,\\
&Q(r) = 0\text{otherwise}.
\end{aligned}$$

Next, consider the dichotomous case for *k* = 25% with the following two categories:

$$\theta_1 = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 14, 15\};$$

$$\theta_2 = \{20, 25, 40, 45, 50, 100\}.$$

Assume that there are three units (*1, 2, 3*), with the following citations distributions

$$\begin{aligned}
&q^1(4) = q^1(6) = q^1(8) = q^1(15) = q^1(20) = q^1(40) = q^1(50) = 1;\\
&q^2(0) = 2, q^2(2) = q^2(6) = q^2(7) = q^2(14) = q^2(45) = q^2(50) = 1;\\
&q^3(0) = 2, q^3(1) = q^3(2) = q^3(3) = q^3(5) = q^3(8) = q^3(100) = 1;\\
&q^j(r) = 0\text{otherwise}, j = 1, 2, 3.
\end{aligned}$$

Panel A in the following table reports the index values (relative to the field index value) for six indicators, grouped in three pairs consisting of one additive rule and the corresponding *HV* indicator. In particular, we report the *Top 25%* and the *HV 25%* (columns 1 and 2), the *RCR* and the *HV max* (columns 3 and 4), and finally the *ANG 25%* and the *HV 25% min-max* (columns 5 and 6).[15] As indicated in the Introduction, we measure the dispersion of the different procedures by means of three statistics: the range, the ratio Max/Min, and the CV. The results are in Panel B of the following table.

	Top 25%	HV 25%	RCR	HV max	ANG 25%	HV 25% min-max
	(1)	(2)	(3)	(4)	(5)	(6)
A. Index values						
Unit 1	1.50	1.80	1.02	1.87	0.60	1.33
Unit 2	1.00	1.00	1.01	0.99	1.03	1.12
Unit 3	0.50	0.43	0.97	0.53	1.38	0.56
B. Variability						
Range	1.00	1.40	0.05	1.34	0.78	0.77
Max/Min	3.00	4.19	1.05	3.55	2.31	2.40
CV	0.50	0.64	0.03	0.60	0.39	0.40

Some comments are in order. Firstly, as we know, the ranking corresponding to the *Top 25%* and the *HV 25%* procedures is the same, but the dispersion of the index values is different. As expected, we observe that the dispersion of the *HV 25%* procedure is greater than that of the *Top 25%* procedure. Finally, let us mention the following interesting observation. Some might believe that for any *z* the corresponding *ANG k%* indicator always has a greater dispersion than the *Top k%* indicator. However, this example provides a counter-example to this conjecture. The reason is that the assignment of the most cited publication to unit *3* offsets the greater share of excellent publications enjoyed by unit 1. This generates a complete re-ranking of the units. The same phenomenon is even more pronounced for *HV* procedures.

[15] The 75th percentile value turns out to be $z = 16.25$.

Table A1
Distribution of the number of articles by country in the all-sciences case, Clinical Medicine, Physics, Engineering, and Economics & Business.

				Percentage of articles in:			
Ranking according to (1)		Total number of articles	%	Clinical Medicine	Physics	Engineering	Economics & Business
		(1)	(2)	(3)	(4)	(5)	(6)
31	ARGENTINA	25,939	0.5	0.3	0.5	0.2	0.2
12	AUSTRALIA	126,072	2.3	2.4	1.2	2	3.3
25	AUSTRIA	43,009	0.8	1.2	0.8	0.6	0.6
21	BELGIUM	60,038	1.1	1.3	1	1	1.3
18	BRAZIL	66,556	1.2	0.9	1.7	1	0.3
8	CANADA	195,938	3.6	3.6	1.9	3.6	5
7	CHINA	197,462	3.6	1.3	5.6	5.3	1.8
30	CZECH REPUBLIC	26,542	0.5	0.2	0.7	0.4	0.7
23	DENMARK	45,908	0.8	1	0.7	0.5	0.9
24	FINLAND	43,769	0.8	1.1	0.6	0.7	0.7
5	FRANCE	282,729	5.2	5.1	6.3	4.4	3.2
4	GERMANY	390,873	7.2	7.8	9	5.8	3.6
27	GREECE	30,917	0.6	0.7	0.6	0.9	0.5
34	HUNGARY	24,398	0.4	0.3	0.5	0.4	0.1
14	INDIA	107,025	2	0.8	2.2	2.3	0.5
39	IRAN	9717	0.2	0.1	0.2	0.3	0
38	IRELAND	16,005	0.3	0.3	0.2	0.3	0.4
22	ISRAEL	55,837	1	1.2	1.2	0.9	1.2
9	ITALY	190,078	3.5	4.1	4.1	3.7	1.9
2	JAPAN	431,828	7.9	8.3	10.1	8.6	1.5
40	LUXEMBOURG	584	0	0	0	0	0
28	MEXICO	29,858	0.5	0.3	0.8	0.5	0.2
13	NETHERLANDS	111,959	2.1	2.7	1.5	1.6	3.1
32	NEW ZEALAND	25,437	0.5	0.4	0.2	0.3	0.7
29	NORWAY	29,511	0.5	0.7	0.3	0.4	0.8
20	POLAND	61,172	1.1	0.4	2.1	1.1	0.2
37	PORTUGAL	20,173	0.4	0.2	0.4	0.5	0.3
6	RW	216,949	4	3.2	4	4.5	2.1
10	RUSSIA	157,349	2.9	0.6	7.2	3.4	0.3
35	SINGAPORE	22,834	0.4	0.3	0.5	1.3	0.6
36	SOUTH AFRICA	21,994	0.4	0.4	0.2	0.3	0.4
16	SOUTH KOREA	89,445	1.6	0.9	2.4	2.9	0.8
11	SPAIN	135,317	2.5	2.3	2.3	2	2
15	SWEDEN	89,902	1.6	2.2	1.3	1.2	1.5
17	SWITZERLAND	80,669	1.5	1.7	1.8	1.2	1
19	TAIWAN	62,928	1.2	1	1.3	2.8	0.8
26	TURKEY	40,018	0.7	1.3	0.4	0.9	0.4
3	UK	397,488	7.3	8.5	5.3	6.8	12.2
33	UKRAINE	24,631	0.5	0	1.2	0.7	0
1	UNITED STATES	1,463,587	26.8	30.8	17.9	24.9	44.6
	TOTAL	5,452,445	100	100	100	100	100

Note: Total number of articles in each field: Clinical Medicine = 1,102,367; Physics = 626,304; Engineering = 421,332, and Economics & Business = 75,687

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2017.04.008.

References

Abramo, G., & D'Angelo, C. A. (2016). A farewell to the MNCS and like size-independent indicators. *Journal of Informetrics*, *10*, 646–651.

Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011a). The measurement of low- and high-impact in citation distributions: technical results. *Journal of Informetrics*, *5*, 48–63.

Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011b). High- and low-impact citation measures: Empirical applications. *Journal of Informetrics*, *5*, 122–145.

Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011c). Average-based versus high- and low-impact indicators for the evaluation of scientific distributions. *Research Evaluation*, *20*, 325–339.

Albarrán, P., Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2015). Differences in citation impact across countries. *Journal of the American Society for Information Science and Technology*, *66*, 512–525.

Altman, A., & Tennenholtz, M. (2005). Ranking systems: the PageRank axioms. In *Proceedings of the 6th ACM conference on electronic commerce (EC-05)* (pp. 1–8).

Bergstrom, C. T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, *68*, 314–316.

Bornmann, L., & Marx, W. (2013). How good is research really? *EMBO Reports*, *14*, 226–230.

Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, *98*, 487–509.

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics*, *5*, 228–230.

Bornmann, L., & Williams. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics*, *7*, 562–574.

Bornmann, L., Mutz, R., Nehaus, C., & Daniel, H.-D. (2008). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, *8*, 93–102.

Bornmann, L., De Moya Anegón, F., & Leydesdorff, L. (2012). The new excellence indicator in the World Report of the SCImago Institutions rankings 2011. *Journal of Informetrics*, *6*(2), 333–335.

Bornmann, L., Bowman, B., Bauer, J., Marx, W., Schier, H., & Palzenberg, M. (2014). Bibliometric standards for evaluating research institutes in the natural sciences. In B. Cronin, & C. Sugimoto (Eds.),. Cambridge: MIT Press.

Bouyssou, D., & Marchant, T. (2011). Ranking scientists and departments in a consistent manner. *Journal of the American Society for Information Science and Technology*, *62*, 1761–1769.

Bouyssou, D., & Marchant, T. (2014). An axiomatic approach to bibliometric rankings and indices. *Journal of Informetrics*, *8*, 449–477.

Carayol, N., & Lahatte, A. (2014). *Dominance relations and ranking when quality and quality both matter: Applications to U.S. universities and econ.* Departments worldwide. *Cahiers du GRETha 2014-13*, Groupe de Recherche en Economis Théorique et Appliqué.

Chabotarev, & Shamis. (1998). Characterization of scoring methods for preference aggregation. *Annals of Operations Research*, *80*, 299–332.

Daniels, H. (1969). Round-robin tournaments scores. *Biometrika*, *56*, 295–299.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121.

Echenique, F., & Fryer, R. G. (2007). A measure of segregation based on social interactions. *Quarterly Journal of Economics*, *122*(2), 441–485.

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics*, *9*(4), 895–906.

Foster, J. E., Greeer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, *52*, 761–766.

Glänzel, W., Thijs, B., & Debackere, K. (2014). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, *101*(2), 939–952.

Golub, B., & Jackson, M. (2000). Naïve learning in social networks and the wisdom of crowds. *American Economic Journa: Microeconomics*, *2*(1), 112–149.

Gordon, A. D. (1979). A measure of the agreement between rankings. *Biometrika*, *66*(1), 7–15.

Grannis, R. (2002). Segregation indices and their functional inputs. *Sociological Methodology*, *32*, 68–84.

Herranz, N., & Ruiz-Castillo, J. (2012). Sub-field normalization in the multiplicative case: high- and low-impact indicators. *Research Evaluation*, *21*, 113–125 [2012]

Herranz, N., & Ruiz-Castillo, J. (2013). The end of the 'European paradox'. *Scientometrics*, *95*, 453–464.

Herrero, C., & Villar, A. (2013). On the comparison of group performance with categorical data. *PLoS ONE*, *8*(12), e84784.

Herrero, C., & Villar, A. (2017). *The Balanced Worth: A procedure to evaluate performance in terms of ordered attributes, mimeo.*

Hirsh, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*, 16569–16572.

Kóczy, L., & Nichifor, A. (2013). The intellectual influence of economic journals: Quality versus quantity. *Economic Theory*, *52*, 863–884.

Keener, J. P. (1993). The Perron-Frobenius Theorem and the ranking of football teams. *SIAM Review*, *35*(1), 80–93.

Laband, D., & Piette, M. J. (1994). The relative impacts of economic journals 1970–1990. *Journal of Economic Literature*, *32*(2), 640–666.

Laslier, J. (1997). *Tournament solutions and majority voting*. New York: Springer-Verlag.

Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators (I3) compared with impact factors (Ifs): An alternative research design with policy implications. *Journal of the American Society for Information Science and Technology*, *62*, 2133–2146.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, *62*, 1370–1381.

Leydesdorff, L. (2012). Alternatives to the journal impact factor: i3 and the top-10% (or top-25%?:) of the most-highly cited papers. *Scientometrics*, *92*, 355–365.

Li, Y., & Ruiz-Castillo, J. (2014). The impact of extreme observations in citation distributions. *Research Evaluation*, *23*, 174–182.

Liebowitz, S., & Palmer, J. (1984). Assesing the relative impact of economic journals. *Journal of Economic Literature*, *22*, 77–88.

Marchant, T. (2009). Score-based bibliometric rankings of authors. *Journal of the American Society for Information Science and Technology*, *60*, 1132–1137.

Moon, J. W., & Pullman, N. (1970). On generalized tournament matrices. *SIAM Review*, *12*, 384–399.

National Science Board. (2012). *Science and engineering indicators 2012*. Arlington, VA: National Science Foundation.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The pagerank citation ranking: bringing order to the web. Technical report, Stanford DigitaI Library Technologies Project*. http://ilpubs.stanford.edu:8090/422

Palacios-Huerta, I., & Volij, O. (2004). The measurement of intellectual influence. *Econometrica*, *72*, 963–977.

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2016). A comparison of two ways of evaluating research units working in different scientific fields. *Scientometrics*, *106*, 539–561.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications-theory with application to literature of physics. *Information Processing & Management*, *12*(5), 297–312.

Rodriguez-Navarro, A. (2011). A simple index for the high-citation tail of citation distributions to quantify research performance in countries and institutions. *PLoS ONE*, *6*(5), e201510.

Rousseau. (2012). Basic properties of both percentile rank scores and the I3 indicator. *Journal of the American Society for Information Science*, *63*, 416–420.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*, 102–117.

Schreiber, M. (2012). Inconsistencies of recently proposed citation impact indicators and how to avoid them. *Journal of the American Society for Information Science and Technology*, *63*, 2062–2073.

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publications output and citation impact. *Scientometrics*, *9*, 281–291.

Slutzki, G., & Volij, O. (2005). Ranking participants in generalized tournaments. *International Journal of Game Theory*, *33*, 255–270.

Slutzki, G., & Volij, O. (2006). Scoring of web pages and tournaments: axiomatizations. *Social Choice and Welfare*, *26*, 75–92.

Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*, 157–177.

Wagner, C., & Leydesdorff, L. (2012). An integrated impact indicator (I3): A new definition of impact with policy relevance. *Research Evaluation*, *21*, 183–188.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *64*, 372–379.

Waltman, L., & Van Eck, N. J. (2015). Field normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, *9*, 872–894.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J., et al. (2012). The leiden ranking 2011/2012: data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*, 2419–2432.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*, 365–391.

West, J. D., Bergstrom, C., & Bergstrom, T. (2010). The eigenfactor metrics^{TM}: A network approach to assessing scholarly Journals: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, *71*, 236–244.

Yanoletzky, G. (2012). A dissimilarity index ofmultidimensional inequality of opportunity. *Journal of Economic Inequality*, *10*(3), 343–373.