Journal of Informetrics 13 (2019) 78–86

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

The F^3-index. Valuing reviewers for scholarly journals

Federico Bianchi [a], Francisco Grimaldo [b], Flaminio Squazzoni [a,*]

[a] Department of Social and Political Sciences, University of Milan, Italy
[b] Department of Computer Science, University of Valencia, Spain

ARTICLE INFO

Article history:
Received 5 April 2018
Received in revised form 7 November 2018
Accepted 18 November 2018
Available online 7 December 2018

ABSTRACT

This paper presents an index that measures reviewer contribution to editorial processes of scholarly journals. Following a metaphor of ranking algorithms in sports tournaments, we created an index that considers reviewers on different context-specific dimensions, i.e., report delivery time, the length of the report and the alignment of recommendations to editorial decisions. To test the index, we used a dataset of peer review in a multi-disciplinary journal, including 544 reviewers on 606 submissions in six years. Although limited by sample size, the test showed that the index identifies outstanding contributors and weak performing reviewers efficiently. Our index is flexible, contemplates extensions and could be incorporated into available scholarly journal management tools. It can assist editors in rewarding high performing reviewers and managing editorial turnover.

1. Introduction

Most academics consider reviewing for journals a fundamental part of their professional work. Not only is it rewarding when editors consider us as an authoritative expert in our field, sometimes reviewing is also accepted for moral reasons, i.e., to comply with social norms characterizing the scientific community (Lamont, 2012; Warne, 2016).

However, reality often works against good intentions. Indeed, reviewing is time consuming and there is a trade-off with other activities, such as preparing grant proposals, lecturing or preparing submissions of journal articles (Bianchi, Grimaldo, Bravo, & Squazzoni, 2018). The fact that reviewing is not recognized for grants and professional career advancement as well as being a voluntary effort which is potentially over-exploitable and sometimes not reciprocated by others, does not help (Tennant et al., 2017; Veríssimo & Roberts, 2013).

However, journal editors know that the reputation of their journal greatly depends on reviewers. On the one hand, the reliability of peer review increases journal legitimacy when authors often see their manuscripts evolving and improving after rounds of reviews (Casnici, Grimaldo, Gilbert, Dondio, & Squazzoni, 2017a,b; Rigby, Cox, & Julian, 2018). On the other, reviewers would benefit from the recognition of their effort either by journals, funding agencies or research institutions (Kachewar & Sankaye, 2013). In this respect, while reviewing is a fundamental step towards academic learning and responsible scholarship, receiving a generic e-mail of thanks or having our name included in the list of journal reviewers published in the latest issue has little real value either symbolically or materially. It is worth noting that recent research has suggested that reputational incentives for peer reviewer should be increased, as while this does not nurture short-term material interests it could have potential long-term material consequences in terms of academic recognition (Mulligan, Hall, & Raphael, 2013; Squazzoni, Bravo, & Takács, 2013; Warne, 2016).

* Corresponding author.
E-mail address: flaminio.squazzoni@unimi.it (F. Squazzoni).

https://doi.org/10.1016/j.joi.2018.11.007

Here, the problem is that recognizing reviewers requires systematic effort of quantifying and comparing a complex, context-specific, and widespread activity. Not surprisingly, these are rarely performed by journals (Margalida & Colomer, 2016; Siler, Lee, & Bero, 2015), while available recognition platforms, such as Elsevier Reviewer Recognition and Publons, value quantity (i.e., the number of reviews) over other dimensions. Although previous research has suggested the need for developing indices to quantify reviewer effort (Cantor & Gero, 2015; Kachewar & Sankaye, 2013; Veríssimo & Roberts, 2013), to our knowledge, there is no example of an index that looks at multiple dimensions or that has been tested on a journal dataset.

To fill this gap, we present here an index that considers reviewing as a measurable activity whose outcomes can be rated, cross-compared and ranked. Following the sports' analogy, we considered two or more reviewers assigned to the same manuscript as match players who compete to deliver a pertinent, informative and timely review (Langville & Meyer, 2012). Given that any submission is potentially different from another in length, depth and required effort, we compared reviewers in the same context so that their performance was relative and context-specific, in order to establish a level playing field. Indeed, we would never see Ivan Lendl, a top player in the 1980s playing against Roger Federer in a tennis tournament or compare the performance of Federer in a first round of a satellite tournament against the number 100 of the ATP ranking with his achievement in a Wimbledon gentlemen's singles final against Rafael Nadal, his most legendary opponent and former number 1 of the ATP ranking.

It is important to note that we are not interested in estimating the quality of reviewers. Considering quality would require defining what quality in peer review actually is, which is a difficult task (Cowley, 2015). Not only is quality a slippery concept hardly quantifiable and standardizable; peer review can also have multiple purposes for whoever is involved. For instance, besides being only a quality screening mechanism, it can also be viewed as a process that is expected to increase the knowledge value of manuscripts (Casnici, Grimaldo, Gilbert, Dondio, et al., 2017a,b; Casnici, Grimaldo, Gilbert, & Squazzoni, 2017). Here, we wanted to consider reviewer performance on certain factors valuable for editors and journals (and indirectly for authors) that can be fully quantified and compared.

For this purpose, we developed an algorithm that rates and ranks reviewers by jointly considering three factors: (1) report delivery time, (2) length of reports and (3) recommendation (Casnici, Grimaldo, Gilbert, Dondio, et al., 2017a,b; Hartonen & Alava, 2013; Laband, 1990). This does not mean that other factors or dimensions are irrelevant. We have proposed a tool that could be adapted to context-specific interests of editors and journals, as well as to the quality of available data.

Furthermore, we did not aim to identify a 'one-size-fits-all' methodology that could reflect context-specific journal characteristics and include varying dimensions. Note that we aggregated the scores of each reviewer in each dimension as in a sports tournament to generate a comprehensive ranking of reviewers' performance only as an illustrative example. We believe that such an exercise could stimulate extensions and implementations, thus enriching the depth, representativeness and flexibility of the index.

The rest of the article is organized as follows. The following section presents our rating and ranking algorithm while Section 3 presents a test of our algorithm on a dataset of reviews for a scientific journal. Finally, in Section 4 we draw some conclusions and discuss limitations of our study.

2. The method

2.1. Computing reviewer ratings

In an influential contribution, Keener (1993) proposed a method to calculate numerical ratings of football teams with incomplete matching. His algorithm used non-negative statistics from contests to generate a *ranking* of teams based on ratings ordering, which included non-evenly matched pairings. The idea was that any contest participant could have assigned a *rating* measuring his/her *absolute strength* that depended on his/her *relative strength*, which in turn reflected the strength of each opponent with which he/she was matched. Therefore, each contest participant's rating depended on interaction outcome and the *strength* of participants. In his inspiration, these computational efforts were intended to question the validity of intuitions and subjective impressions, while showing the advantages of objective performance measures. It is worth noting that Keener's method has been applied to a variety of fields, including meteorology (Christy, Norris, Redmond, & Gallo, 2006), the study of animal social hierarchies (Bush, Quinn, Balreira, & Johnson, 2016), the ranking of web pages by popularity (Franceschet, 2011) and group-ranking aggregation in management science (Hochbaum & Levin, 2006).

Here, we adapted Keener's method to develop a rating and ranking algorithm for reviewers of scholarly journals. We considered reviewers as participants of a tournament in which each reviewer was matched over time with others as they were assigned to evaluate the same manuscripts. Similarly to participants in a tournament, reviewers' *strength* can be measured by calculating a *rating* based on their reviewing behaviour, by measuring it together with the behaviour of other reviewers assigned to the same manuscripts and their strength.

Coherently with Keener's idea, we define the relationship between *strength* and *rating* as follows:

1 the *strength* of a reviewer (s_i) can be assessed through *i*'s reviewing *behaviour* compared to the strength of other reviewers matched with *i*;
2 the *rating* of a reviewer r_i is uniformly proportional to its *strength* s_i, such that $\forall i$, $s_i = \lambda r_i$, with λ being constant for all reviewers.

The first assumption reflects the idea that manuscripts vary in terms of length and technical difficulties of the text so that different efforts are implied by reviewers. Returning to the sports analogy, today it is harder to play against Roger Federer than against Ivan Lendl who would soon be turning 60. Similarly, comparing the performance of a serve and volley tennis player in a competition on a grass court with a red-clay court tournament is not very informative. Therefore, we estimated the *strength* of each reviewer by considering the context, thus comparing reviewers of the same manuscript. Let us assume that an editor assigns two different manuscripts to reviewers (i, j) and (y, z) respectively. Let us also assume that the first manuscript is more complicated to review than the latter. Then, the *strength* of reviewer i can be best assessed by comparing his/her behaviour to j's rather than to y or z, who are assigned a manuscript that is less difficult to review.

Following Keener's method, we propose to assess s_i by measuring one or more attributes of i's behaviour when evaluating the same manuscript as j (e.g., the report length). In order to measure an attribute a_{ij}, we calculate the value of a non-negative statistic S_{ij}, i.e., a statistical value representing an attribute of i's behaviour in the evaluation of the same manuscripts reviewed by j (e.g., the number of words contained in the report text). Following a consolidate practice (Langville & Meyer, 2012), in case of multiple matchings of the same reviewers i and j, S_{ij} is equal to the sum of all values across various manuscripts.

Given that manuscripts can vary considerably in certain structural characteristics (e.g., text length, employed methods, etc.,), the distribution of statistics for the selected attributes could vary considerably between different reviewers. In order to control for this, we calculate attribute measurements by standardizing raw statistics. Then, following a recommendation by Keener (1993), we transform standardized statistics by applying Laplace's Rule of Succession (1995 [1825]) in order to avoid 'winner-takes-all' effects, as follows:

$$a_{ij} = \frac{S_{ij} + 1}{S_{ij} + S_{ji} + 2}. \tag{1}$$

For instance, let us assume that i reviewed two manuscripts which were also assigned to j. Then, suppose that i delivered two reports consisting of 500 and 650 words each, while j's review reports consisted of 200 and 250 words, respectively. In this case, $S_{ij} = 500 + 650$ and $S_{ji} = 200 + 250$, hence

$$a_{ij} = \frac{(500 + 650) + 1}{(500 + 650) + (200 + 250) + 2} \approx 0.718.$$

Let us now assume that i delivered his/her report after 10 and 15 days each, while j took 11 and 25 days. As quicker responses are preferable here, we use the numbers of the opponent as a convenient statistic. Then, $S_{ij} = 11 + 25$ and $S_{ji} = 10 + 15$, and

$$a_{ij} = \frac{(11 + 25) + 1}{(11 + 25) + (10 + 15) + 2} \approx 0.587.$$

We then distribute the obtained attribute values into a square matrix

$$\mathbf{A} = [a_{ij}]_{m \times m}, \tag{2}$$

where m is the number of reviewers of the considered journal.

Considering r_j as the numerical *rating* of reviewer j and building on Assumption 1, we assume that

$$s_{ij} = a_{ij}\, r_j, \tag{3}$$

where the *relative strength* of reviewer i with respect to j (s_{ij}) is yielded by measuring the attribute of i's behaviour as reviewer of the same manuscripts reviewed by j, weighted by the rating of j.

We therefore define the *absolute strength* of reviewer i as the sum of i's relative strengths compared to all other reviewers, as follows:

$$s_i = \sum_{j=1}^{m} s_{ij} = \sum_{j=1}^{m} a_{ij} r_j. \tag{4}$$

Then, we define a *strength* vector for all reviewers as:

$$\mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix} = \begin{pmatrix} \sum_j a_{1j} r_j \\ \sum_j a_{2j} r_j \\ \vdots \\ \sum_j a_{mj} r_j \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix} = \mathbf{A}\mathbf{r} \tag{5}$$

By considering Assumption 2 and Eq. (5), it follows that

$$\mathbf{A}\mathbf{r} = \lambda \mathbf{r}. \tag{6}$$

Table 1
Parameters of the F^3-index: reviewers' attributes and related statistics.

	Attributes (a)	Statistics (S)
l	Report length	Number of words contained in reviewers' reports
t	Delivery time	Number of days between acceptance to review and report delivery
e	Alignment of editorial decision to reviewer recommendation	Number of events in which the editorial decision was equal to a reviewer's recommendation

Therefore, by calculating the $\mathbf{r}$ vector, we derive the rating of each reviewer on attribute a. It is worth noting that these computations can be performed by applying a power algorithm, as in Algorithm 1, on condition that $\mathbf{A}$ complies with the principles of non-negativity, irreducibility, and primitivity (see Meyer, 2000).

Algorithm 1. Pseudocode to calculate the rating vectors

Input: Number of reviewers (m), Attribute-value matrix ($\mathbf{A}$), Convergence criterion (ϵ)
Output: Ratings vector ($\mathbf{r}$)

$\mathbf{r}_{-1} \longleftarrow 0$
$\mathbf{r} \longleftarrow 1/m \cdot \mathbf{e}^T$
while $max(|\mathbf{r} - \mathbf{r}_{-1}|) > \epsilon$ **do**
  $\mathbf{r}_{-1} \longleftarrow \mathbf{r}$
  $\mathbf{r} \longleftarrow \mathbf{Ar} / \sum_{i=1}^{m} (\mathbf{Ar})_i$
end while

2.2. Computing the F^3-index

Following this method, various *attributes* can be used to rate the performance of journal reviewers, by calculating appropriate *statistics*. Here, we followed previous attempts at measuring important characteristics of peer review reports and concentrated on three dimensions (Ellison, 2002; Hartonen & Alava, 2013; Laband, 1990).

First, we consider the report length as a proxy of the amount of information provided by the reviewer to editors and authors (Bravo, Farjam, Grimaldo, Birukou, & Squazzoni, 2018; Casnici, Grimaldo, Gilbert, Dondio, et al., 2017; Laband, 1990). We assume that a detailed report is not only expected to include more valuable information that helps editors to judge the quality, rigour and validity of a manuscript; it can also provide insights for authors to improve their work. Secondly, we consider the time taken by the reviewer to return the report. Indeed, minimizing delivery time is key to reduce time wasted by editors to chase reviewers and secure supplementary reports, while benefiting authors by reducing publication delay (Chetty, Saez, & Sandor, 2014). Finally, we look at the alignment of editorial decisions to reviewer recommendations. We consider this as a proxy of the influence of the reviewer on the editor's opinion (Kravitz et al., 2010). In this respect, it is worth noting that not only could editors profit from understanding more systematically which reviewers were more influential on their own decisions, certain editors may want to select reviewers with different opinions and attitudes. Balancing reviewers with different attitudes and exploiting their diversity to deliver a sound justified report to authors are one of the trade secrets of any good editor but could be supported more efficiently by systematic use of internal journal data.

More specifically, we create parameters for our index on three attributes: *report length*, *delivery time*, and *alignment of editorial decision to reviewer recommendation*. Each attribute is measured through appropriate statistics, as reported in Table 1:

By choosing a relevant set of statistics, the algorithm can be parameterised against different dimensions, so that different rating vectors are computed, one for each attribute, as specified in Section 2.1. The F^3-index is generated by aggregating each reviewer's different attribute-specific rates into a synthetic rating value.

Of the varying methods used to aggregate different rankings (Lin, 2010), heuristic algorithms are the most appropriate for a small number of long attribute-specific rating lists. They range from simple arithmetic ranking averages to Markov chains. The latter have been used to generate right-skewed ratings focused on the top of the rank, but are less accurate at assigning positions at the bottom. Given that the aim of the F^3-index was to consider both aspects, we aggregate the different attribute-specific ratings by standardizing them and calculating the mean of the obtained values for each reviewer.

Therefore, being $n = 3$ the number of different attribute-specific ratings ($\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$), the F^3-Index of each reviewer can be calculated as follows:

$$F_i^3 = \frac{1}{n} \sum_{k=1}^{n} \frac{r_{k,i} - \mu_{\mathrm{r_k}}}{3\sigma_{\mathrm{r_k}}} \tag{7}$$

3. An application

We tested the F^3-index by applying it to a dataset of reviewers of the *Journal of Artificial Societies and Social Simulation* (from now on, *JASSS*). JASSS is an open-access, online interdisciplinary journal which publishes applications of computer simulation to the understanding of social dynamics and processes. The multi-disciplinary character of *JASSS* was instrumental to provide a robust test of our algorithm, which is useful especially to compare reviewers with different backgrounds and skills (see

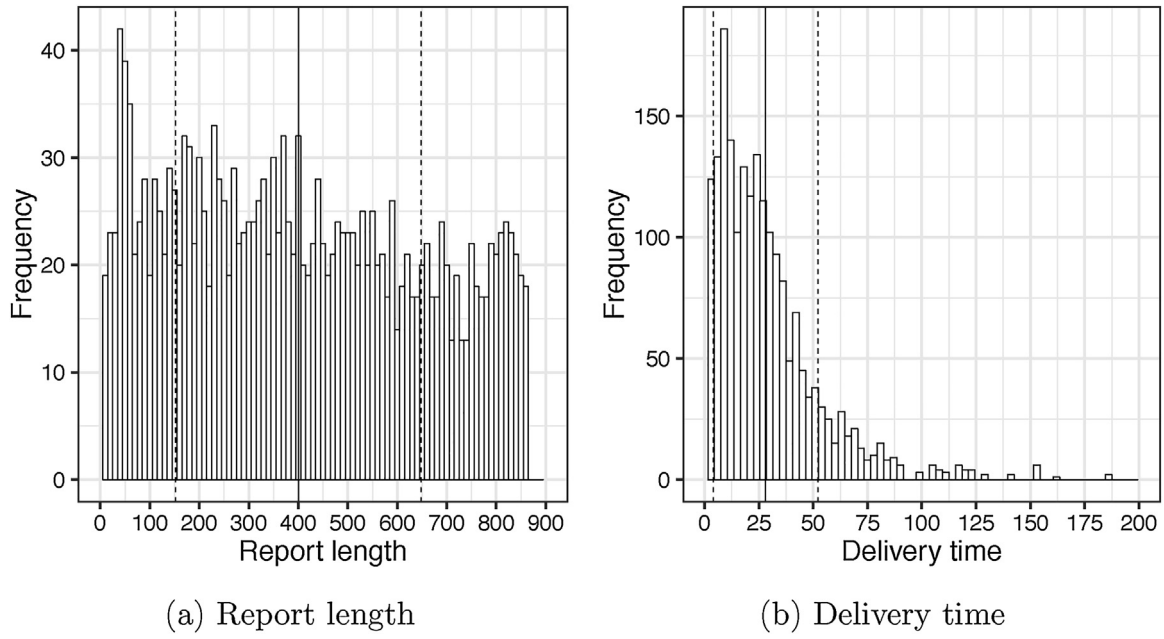

Fig. 1. Distributions of *Report length* and *delivery time* with mean (vertical solid lines) and standard deviation (vertical dashed lines) values.

Table 2
Reviewers' recommendations and editorial decisions: relative frequencies.

Level	Reviewer's recommendation	Editorial decision
Accept	6.68%	0.00%
Minor revision	26.03%	26.94%
Revise and resubmit	40.65%	46.94%
Reject	26.63%	26.13%

Casnici, Grimaldo, Gilbert, & Squazzoni, 2017; Hauke, Lorscheid, & Meyer, 2017; Squazzoni & Casnici, 2013 and Squazzoni and Casnici, 2013 for an overview on the journal and its reviewers).

The dataset included information on the first round of evaluation of 606 submitted manuscripts which passed the editorial rejection desk between 2005 and 2011. Considering that the journal strictly follows double-blind peer review, but each reviewer also receives all review reports included in the editorial notification letter to the authors, restricting data to the first round of reviews permitted us to avoid any learning or imitation effect between reviewers (Bravo et al., 2018). We removed all review guidelines and introductory comments from each review report and extracted the text from the journal template, which included instructions on how to perform reports. Each manuscript was assigned to at least two reviewers, with a total number of $m = 544$ reviewers and $k = 995$ matched pairs. Note that if a paper was assigned to more than two reviewers, we used all different reviewer pairs to calculate the ratings. Note that we excluded manuscripts reviewed directly by the journal editor from the dataset so to prevent bias in the *alignment* rating.

The computation of the F^3-index was possible because the dataset complied with the required conditions mentioned in Section 2.1. More specifically, 501 out of 544 reviewers (92.1%) were indirectly connected with each other by a chain of co-reviewed manuscripts. This allowed us to build sufficiently irreducible non-negative matrices $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ for the three parameters. Moreover, 69.21% of all possible chains of manuscripts connecting reviewers had a length between 4 and 6. This meant that most reviewers were connected by a similar number of co-reviewed manuscripts and that the primitivity condition was sufficiently met to calculate the rating vectors as presented in Section 2.1.

Submitted reports had an average length of approximately 400.1 words ($SD = 248.56$) and distribution of the *report length* (l) was approximately uniform for reviewers (see Fig. 1a). The mean *delivery time* (t) between reviewers' acceptance notification and report submission was 28.08 days ($SD = 24.11$). Unlike the *report length*, the distribution of t was right-skewed, ranging up to a maximum of 187 days (see Fig. 1b).

Along with the report text, reviewers were required to give a recommendation concerning the final publication decision, which could be in four categories: "Accept", "Minor revision", "Revise and resubmit" (or "Major revision"), or "Reject", which were the same used by the editor for authors' notifications. Table 2 reports the distribution of reviewer recommendations to the editor and the final editorial decisions. The two distributions were positively correlated ($V = 0.47$, 95% CI: [0.44, 0.50]). "Revise and resubmit" (from here on, "R&R") was prevalent among both reviewer recommendations and editorial decisions (46.94% of manuscripts). It is worth noting that manuscripts for which reviewers recommended the same evaluation

Table 3
Anonymized results of the application of the F^3-index on the *JASSS* dataset. r_l, r_t and r_e report the rating of the reviewers with respect to *report length*, *delivery time*, *and alignment with the editorial decision*. $\#_l$, $\#_t$ and $\#_e$ report the grouped ranking of the reviewers on each of the previous attributes, *delivery time*, and *alignment with the editorial decision*. k_i reports the number of pair matches between the reviewer and another reviewer. $\bar{l}_i$ and $\bar{t}_i$, report the reviewers' mean values of *report length* and *delivery time*. $\bar{e}_i$ reports reviewers' proportions of alignment of their recommendations with the editorial decision.

#	Name	F^3	$\#_l$	r_l	$\#_t$	r_t	$\#_e$	r_e	k_i	$\bar{l}_i$	$\bar{t}_i$	$\bar{e}_i$
1	Lorem Ipsum	0.857	4	0.458	1	1.388	3	0.725	12	480	9.7	0.67
2	Fusce At	0.840	1	1.699	3	0.823	6	-0.001	16	661	19.4	0.75
3	Vestibulum Sit	0.794	5	0.250	1	1.609	4	0.524	12	509	3.1	0.92
4	Aenean Eget	0.757	4	0.425	2	1.241	4	0.605	8	518	5.1	1
5	Vivamus Ac	0.537	4	0.425	1	1.893	8	-0.706	30	469	13.4	0.60
6	Praesent Vitae	0.528	2	1.154	5	0.228	5	0.202	9	714	28.4	0.67
7	Fusce Quis	0.527	5	0.287	3	0.892	4	0.402	9	478	11.1	0.78
8	Donec Egestas	0.525	2	1.191	6	-0.017	4	0.401	9	586	24.8	0.67
9	Aliquam Ac	0.479	4	0.409	5	0.223	3	0.806	4	395	26.3	1
10	Vivamus Viverra	0.459	3	0.773	4	0.603	6	0	4	649	6.5	0.75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
535	Morbi Lectus	−0.436	5	0.004	8	−0.706	7	−0.605	8	491	61.9	0.25
536	Morbi pulvinar	−0.451	8	−0.678	8	−0.673	6	0	4	151	50.5	0.50
537	Quisque Congue	−0.461	5	0.225	7	−0.600	9	−1.007	6	596	38.2	0.17
358	Quisque Iaculis	−0.473	7	−0.601	8	−0.817	6	−0.001	8	254	35.6	0.50
539	Vivamus Auctor	−0.474	6	−0.330	6	−0.285	8	−0.806	4	180	19.0	0
540	Mauris Id	−0.477	9	−1.042	7	−0.589	5	0.201	7	140	95.3	0.71
541	Nulla Eu	−0.522	7	−0.472	6	−0.287	8	−0.806	8	290	35.1	0.25
542	Nullam Vel	−0.558	5	0.131	8	−0.798	9	−1.007	6	485	70.3	0
543	Ut Scelerisque	−0.600	9	−1.287	8	−0.715	5	0.202	6	80	37.3	0.67
544	Cras Mattis	−0.733	7	−0.443	8	−0.751	9	−1.007	9	246	40.8	0.11

accounted for 37.89% of the total, most frequently suggesting "R&R". In case of equal recommendation by two reviewers, the editor followed their recommendations 89.39% of times (33.87% of the total). Finally, 59.30% of editorial decisions corresponded to at least one reviewer recommendation.

In order to check whether *report length*, *delivery time*, and *alignment* were actually measuring three different dimensions of reviewer behaviour, we checked possible correlations between them. We found no evidence of any linear correlation between *report length* and *delivery time* ($r = -0.04$, 95% CI: [−0.08, 0.00]). Furthermore, analysis of variance reported very small effects of reviewers' recommendations on *report length* ($\eta^2 = 0.014$, 95% CI: [0.005, 0.025]) and *delivery time* ($\eta^2 = 0.004$, 95% CI: [0.000, 0.010]).

Table 3 reports the F^3-indices of the 10 top and least rated reviewers. The ranking was obtained by simply sorting the F^3-index distribution by decreasing order and ratings were calculated using Algorithm 1 with a convergence criterion of $\epsilon = 10^{-16}$. Results were anonymized in order to protect the reviewers' identity. Due to the mathematical properties of Keener's method on which our algorithm is based, the distance between two adjacent ranking positions are not uniform across the whole distribution. This implies that the difference between the F^3-index of two adjacently ranked reviewers tends to decrease as long as the pair approaches the mean value of the distribution. For instance, the distance between reviewers number 1 and 2 will be higher than that of reviewers 200 and 201, as the normal distribution of the F^3-index is mapped on a linear vector in order to generate a ranking.

Fig. 2a shows the distribution of the F^3-index of reviewers. The index is approximately normally distributed with zero mean and a standard deviation of $\sigma_{F^3} = 0.192$. As a consequence of Eq. (1), each newly added reviewer was located around the mean of the distribution to move towards the head or the tail when matched with other reviewers. Fig. 2b shows the standardized attribute ratings that were averaged to obtain the F^3-index. For illustrative purposes only, we followed Huntsberger's (1962) rule to discretize ratings into 10 intervals[1] of similar rankings based on multiples of the standard deviation. Following this partition, columns $\#_l$, $\#_t$ and $\#_e$ in Table 3 show the grouped ranking of the reported reviewers for the three attribute-specific ratings: *report length*, *delivery time*, and *alignment with the editorial decision*, respectively. It is worth noting that while this permits an easy outlook, our ranking was sufficiently flexible to add or delete any attribute that is considered relevant, as well as to look at individual attributes separately.

Table 3 shows that our algorithm generated rating scores which did not overlap the raw statistics of attributes used to parameterize the algorithm, although they did partially correlate. The correlation was moderate with reviewers' average report length $\bar{l}_i$ ($r = 0.34$, 95% CI: [0.27, 0.41]) and average delivery time $\bar{t}_i$ ($\rho = -0.37$, 95% CI: [−0.43, −0.29]), while it was relatively strong with the average alignment to editorial decision $\bar{e}_i$ ($\rho = 0.50$, 95% CI: [0.44, 0.56]). We also did not find any correlation with reviewers' pair matches k_i ($\rho = 0.01$, 95% CI: [−0.08, 0.10]).

The F^3-index is sufficiently robust to generate stable orderings in the presence of low and even moderate variability in the attributes that map reviewer behaviour. Fig. 3 shows the Spearman's rank correlation coefficients under two scenarios,

[1] The number of intervals to discretize a continuous variable ($m = 544$ reviewer ratings) can be estimated using the rule of Huntsberger as $1 + 3.3\, log_{10} \approx 10$.

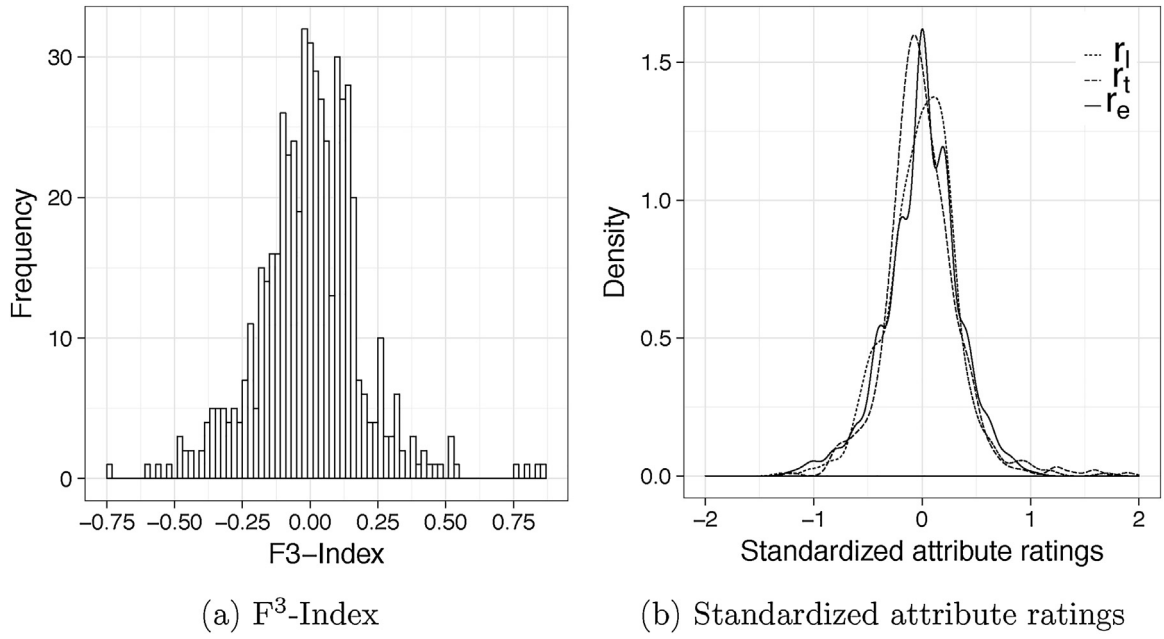

Fig. 2. Distribution of the F^3-index and of the standardized attribute ratings.

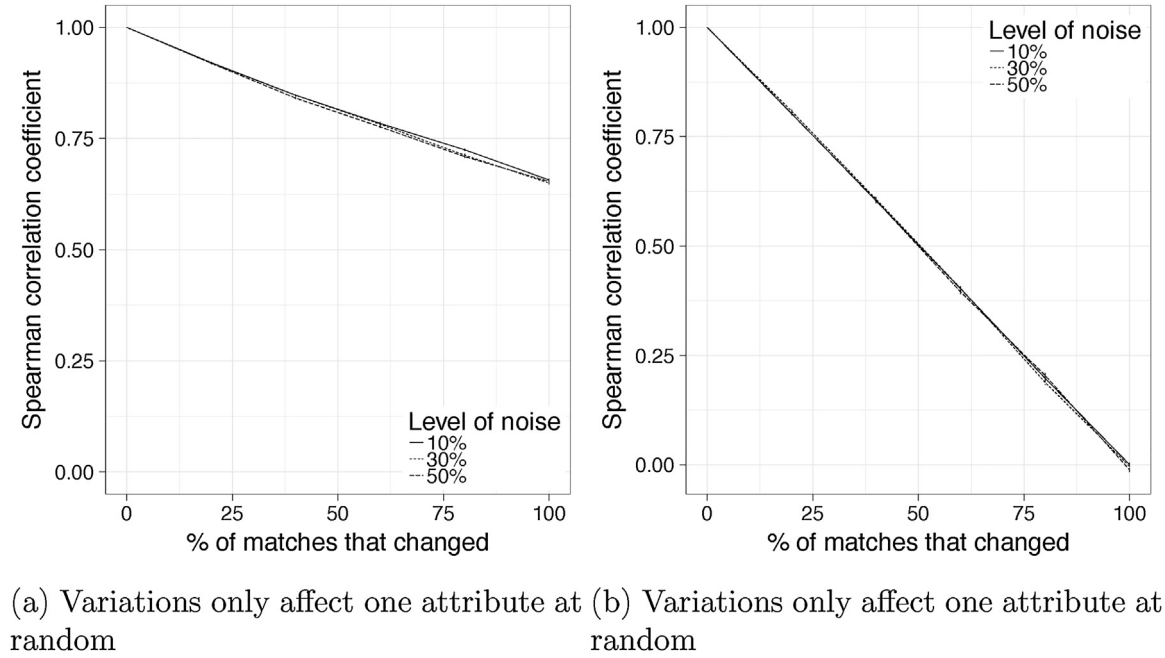

Fig. 3. Robustness check for the F^3-index. Error bars are plotted for 100 runs of each configuration.

i.e., when variations affected only one random attribute or all attributes (i.e. *l*, *t* or *e*). The small error bars show the stability of the results for 100 runs of each experiment. Fig. 3a shows that correlation values were strong even for a high percentage of variation and noise level (expressed as a percentage of the maximum value for each attribute). This would corroborate the robustness of the F^3-index, as it combines a set of attribute-specific ratings that overcome local variations. Furthermore, when variation affects all attributes (see Fig. 3b), correlation values decreased with the percentage of variation, while the level of noise increased.

4. Conclusions

Thanks to the digitalization of scientific records, rankings have been increasingly used in assessment of research performance at various levels and for different purposes (Bonaccorsi & Cicero, 2016; Dunaiski, Geldenhuys, & Visser, 2018). Here, we presented an index that ranks and rates reviewers in academic journals and used internal journal data to test it. We wanted to compare the relative performance strength of reviewers on three dimensions, i.e., the report delivery time, the length of the report and the alignment of their recommendations to editorial decisions. Rather than measuring only the number of reviews, our index considered these multiple dimensions either separately or jointly.

The choice of dimensions and parameters reflected previous attempts at quantifying reviewer performance (Casnici, Grimaldo, Gilbert, Dondio, et al., 2017; Hartonen & Alava, 2013; Laband, 1990). However, it must be noted that our index is flexible and easily extendable to other dimensions while parameters could be weighted differently depending on the journal context and available data. For instance, following recent innovations by Publons and expected technological progress in editorial journal management systems, it would be feasible to improve the index by including author and/or editor ratings of reviewer reports, as well as applying and sharing it in a bunch of linked/sister journals. Considering the increasing number of publications and journals and the imbalance of the reviewing distribution effort, any improvement of reviewer recognition services is essential – in our opinion – to improve the sustainability of peer review as a whole (Dean & Forray, 2018; Kovanis, Porcher, Ravaud, & Trinquart, 2016; Tennant et al., 2017).

Regarding its potential use by journal editors, our index could identify and recognize the most active reviewers, who could be awarded by symbolic incentives. For example, rather than a generic e-mail of thanks, a list publishing top reviewers in the journal website could be more valuable for reviewers' reputation, especially when they are not editorial board members. In this respect, the index could be also used to manage a journal's editorial turnover, by co-opting new members based on performance. Here, by working with multiple dimensions and using internal statistics, journals could avoid nurturing scholars' competitive spirits and gaming temptations, which are often intrinsic to any formalized ranking and quantification (Seeber, Cattaneo, Meoli, & Malighetti, 2017).

Furthermore, once updated regularly, the index could inform editors on reviewer selection, by supporting a more appropriate match between reviewers and task. For instance, considering submissions by prestigious authors, an editor could be interested in maximizing the editorial quality of the process by selecting reviewers who are more prolific and detailed in their report and frequently aligned with their decisions, while at the same time neglecting turnaround time. In other circumstances, for instance when submission authors are early career researchers, an editor could be also interested to reduce turnaround time, while at the same time ensuring informative reports that could help authors improve their manuscript. Obviously, this would require constant updates, which could be problematic for certain journals.

It should be noted that our work has also important limitations. First, although our index is flexible and adaptable, there are intrinsic limits in quantifying certain qualitative dimensions of peer review (Cowley, 2015). For example, our index should not be used to examine the quality of peer review comprehensively, as this would require discussing what quality is and for what purpose it is relevant for. As already mentioned, the index is only useful to rank performance on quantifiable factors and identify outstanding cases, while careful attention must be paid on conceptual frameworks to inform parameter choices (Subochev, Aleskerov, & Pislyakov, 2018) and communication means in order to avoid any misuse either by editors or reviewers.

Secondly, the index can neither substitute editorial intuitions and tacit knowledge on reviewer selection, nor can it fully automatize reviewer selection or recognition. We considered it as a means to complement editorial judgement and help editors to monitor reviewer performance more systematically. At the same time, there is no doubt that the management of journals could dramatically benefit from more robust, internal data-driven analytics. Here, given the prominent role of publications for academic career and promotion and considering the influence and the role of journals, adopting more robust indicators to manage editorial processes should be considered as part of a journal's responsibility and accountability.

Author contributions

Federico Bianchi: Conceived and designed the analysis; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Francisco Grimaldo: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Flaminio Squazzoni: Conceived and designed the analysis; Wrote the paper.

Acknowledgements

This work was partially supported by the COST Action TD1306 – New Frontiers of Peer Review (2014–2018). A preliminary version of this manuscript has been presented to the PEERE International Conference on Peer Review, CNR, Rome, 7–9 March 2018.

References

Bianchi, F., Grimaldo, F., Bravo, G., & Squazzoni, F. (2018). The peer review game: An agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics*, *116*(3), 1401–1420. http://dx.doi.org/10.1007/s11192-018-2825-4

Bonaccorsi, A., & Cicero, T. (2016). Nondeterministic ranking of university departments. *Journal of Informetrics*, *10*(1), 224–237. http://dx.doi.org/10.1016/j.joi.2016.01.007

Bravo, G., Farjam, M., Grimaldo, F., Birukou, A., & Squazzoni, F. (2018). Hidden connections: network effects on editorial decisions in four computer science journals. *Journal of Informetrics*, *12*(1), 101–112. http://dx.doi.org/10.1016/j.joi.2017.12.002

Bush, J. M., Quinn, M. M., Balreira, E. C., & Johnson, M. A. (2016). How do lizards determine dominance? Applying ranking algorithms to animal social behaviour. *Animal Behaviour*, *118*, 65–74. http://dx.doi.org/10.1016/j.anbehav.2016.04.026

Cantor, M., & Gero, S. (2015). The missing metric: Quantifying contributions of reviewers. *Royal Society Open Science*, *2*(2) http://dx.doi.org/10.1098/rsos.140540

Casnici, N., Grimaldo, F., Gilbert, N., Dondio, P., & Squazzoni, F. (2017). Assessing peer review by gauging the fate of rejected manuscripts: The case of the Journal of Artificial Societies and Social Simulation. *Scientometrics*, *113*(1), 533–546. http://dx.doi.org/10.1007/s11192-017-2241-1

Casnici, N., Grimaldo, F., Gilbert, N., & Squazzoni, F. (2017). Attitudes of referees in a multidisciplinary journal: An empirical analysis. *Journal of the Association for Information Science and Technology*, *68*(7), 1763–1771. http://dx.doi.org/10.1002/asi.23665

Chetty, R., Saez, E., & Sandor, L. (2014). What policies increase prosocial behavior? An experiment with referees at the Journal of Public Economics. *Journal of Economic Perspectives*, *28*(3), 169–188. http://dx.doi.org/10.1257/jep.28.3.169

Christy, J. R., Norris, W. B., Redmond, K., & Gallo, K. P. (2006). Methodology and results of calculating central California surface temperature trends: Evidence of human-induced climate change? *Journal of Climate*, *19*(4), 548–563. http://dx.doi.org/10.1175/JCLI3627.1

Cowley, S. J. (2015). How peer-review constrains cognition: On the frontline in the knowledge sector. *Frontiers in Psychology*, *6*, 1706. http://dx.doi.org/10.3389/fpsyg.2015.01706

Dean, K. L., & Forray, J. M. (2018). The long goodbye: Can academic citizenship sustain academic scholarship? *Journal of Management Inquiry*, *27*(2), 164–168. http://dx.doi.org/10.1177/1056492617726480

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). Author ranking evaluation at scale. *Journal of Informetrics*, *12*(3), 679–702. http://dx.doi.org/10.1016/j.joi.2018.06.004

Ellison, G. (2002). The slowdown of the economics publishing process. *Journal of Political Economy*, *110*(5), 947–993. http://dx.doi.org/10.1086/341868

Franceschet, M. (2011). PageRank: Standing on the shoulders of giants. *Communications of the ACM*, *54*(6), 92–101. http://dx.doi.org/10.1145/1953122.1953146

Hartonen, T., & Alava, M. J. (2013). How important tasks are performed: Peer review. *Scientific Reports*, *3*, 1679. http://dx.doi.org/10.1038/srep01679

Hauke, J., Lorscheid, I., & Meyer, M. (2017). Recent development of social simulation as reflected in JASSS between 2008 and 2014: A citation and co-citation analysis. *Journal of Artificial Societies and Social Simulation*, *20*(1), 5. http://dx.doi.org/10.18564/jasss.3238

Hochbaum, D. S., & Levin, A. (2006). Methodologies and algorithms for group-rankings decision. *Management Science*, *52*(9), 1394–1408. http://dx.doi.org/10.1287/mnsc.1060.0540

Huntsberger, D. V. (1962). *Elements of statistical inference*. Prentice-Hall.

Kachewar, S. G., & Sankaye, S. B. (2013). Reviewer index: A new proposal of rewarding the reviewer. *Mens Sana Monographs*, *11*(1), 274–284. http://dx.doi.org/10.4103/0973-1229.109347

Keener, J. P. (1993). The Perron–Frobenius theorem and the ranking of football teams. *SIAM Review*, *35*(1), 80–93. http://dx.doi.org/10.1137/1035004

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLOS ONE*, *11*, 1–14. http://dx.doi.org/10.1371/journal.pone.0166387

Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., & Tierney, W. M. (2010). Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLOS ONE*, *5*(April), 5. http://dx.doi.org/10.1371/journal.pone.0010072

Laband, D. N. (1990). Is there value-added from the review process in economics? Preliminary evidence from authors. *The Quarterly Journal of Economics*, *105*(2), 341–352.

Lamont, M. (2012). *How professors think. Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press.

Langville, A. N., & Meyer, C. D. (2012). *Who's # 1?: The science of rating and ranking*. Princeton University Press.

Laplace, P.-S. (1995). *Philosophical Essay on Probabilities [1825]*. Springer.

Lin, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(5), 555–570. http://dx.doi.org/10.1002/wics.111

Margalida, A., & Colomer, M. A. (2016). Improving the peer-review process and editorial quality: Key errors escaping the review and editorial process in top scientific journals. *PeerJ*, *4*, e1670. http://dx.doi.org/10.7717/peerj.1670

Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*. Philadelphia, PA: SIAM.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*, *64*(1), 132–161. http://dx.doi.org/10.1002/asi.22798

Rigby, J., Cox, D., & Julian, K. (2018). Journal peer review: A bar or bridge? An analysis of a paper's revision history and turnaround time, and the effect on citation. *Scientometrics*, *114*(3), 1087–1105. http://dx.doi.org/10.1007/s11192-017-2630-b

Seeber, M., Cattaneo, M., Meoli, M., & Malighetti, P. (2017). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, *12* http://dx.doi.org/10.1016/j.respol.2017.12.004 (In press)

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, *112*(2), 360–365. http://dx.doi.org/10.1073/pnas.1418218112

Squazzoni, F., Bravo, G., & Takács, K. (2013). Does incentive provision increase the quality of peer review? An experimental study. *Research Policy*, *42*(1), 287–294. http://dx.doi.org/10.1016/j.respol.2012.04.014

Squazzoni, F., & Casnici, N. (2013). Is social simulation a social science outstation? A bibliometric analysis of the impact of JASSS. *Journal of Artificial Societies and Social Simulation*, *16*(1), 10. http://dx.doi.org/10.18564/jasss.2192

Subochev, A., Aleskerov, F., & Pislyakov, V. (2018). Ranking journals using social choice theory methods: A novel approach in bibliometrics. *Journal of Informetrics*, *12*(2), 416–429. http://dx.doi.org/10.1016/j.joi.2018.03.001

Tennant, J., Dugan, J., Graziotin, D., Jacques, D. C., Waldner, F., Mietchen, D., et al. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review [version 3; referees: 2 approved]. *F1000 Research*, *6*(1151) http://dx.doi.org/10.12688/f1000research.12037.3

Veríssimo, D., & Roberts, D. L. (2013). The academic welfare state: Making peer-review count. *Trends in Ecology and Evolution*, *28*(11), 623–624. http://dx.doi.org/10.1016/j.tree.2013.07.003

Warne, V. (2016). Rewarding reviewers – Sense or sensibility? A Wiley study explained. *Learned Publishing*, *29*(1), 41–50. http://dx.doi.org/10.1002/leap.1002