Who is the 'Journal Grand Master'? A new ranking based on the Elo rating system

Robert Lehmann and Klaus Wohlrabe

8 March 2017

Who is the 'Journal Grand Master'?

A new ranking based on the Elo rating system

Robert Lehmann* Klaus Wohlrabe†

March 8, 2017

Abstract. In this paper we transfer the Elo rating system, which is widely accepted in chess, sports and other disciplines, to rank scientific journals. The advantage of the Elo system is the explicit consideration of the factor time and the history of a journal's ranking performance. Most other rankings that are commonly applied neglect this fact. The Elo ranking methodology can easily be applied to any metric, published on a regular basis, to rank journals. We illustrate the approach using the SNIP indicator based on citation data from Scopus. Our data set consists of more than 20 000 journals from many scientific fields for the period from 1999 to 2015. We show that the Elo approach produces similar but by no means identical rankings compared to other rankings based on the SNIP alone or the Tournament Method. Especially the rank order for rather 'middle-class' journals can tremendously change.

Keywords: Elo rating system, journal rankings, SNIP indicator

*Ifo Institute – Leibniz-Institute for Economic Research at the University of Munich e.V., Poschingerstr. 5, D-81679 Munich. Phone: +49 89/9224-1652. Email: lehmann@ifo.de.

†Corresponding author. Ifo Institute – Leibniz-Institute for Economic Research at the University of Munich e.V., Poschingerstr. 5, D-81679 Munich. Phone: +49 89/9224-1229. Email: wohlrabe@ifo.de.

1 Introduction

Measuring the 'quality' of scientific publishing has always been an important aspect for researchers, institutions, politics, and the public. Next to financial incentives for the publisher, publications in high-quality journals are necessary prerequisites for future job market signals of the scientists. What defines a journal as 'high-quality' mainly depends on the classification or ranking scheme that is applied. The question on how such a classification scheme should look like has entailed a heated debate in general, which is especially pronounced in several scientific disciplines such as economics. In this article we do not want to comment on the 'right' or the 'wrong' of existing rankings, but rather adopt an alternative system that was originally developed for chess: the Elo rating system.

One of the main criticisms which can be raised when it comes to rank journals is the largely time invariance of the classification scheme. Generally, many journal metrics are reported with respect to a given year. The prestige of a journal can be negatively affected in a given year if the corresponding metric significantly drops, although in the years before the performance was very good. This shortcoming becomes irrelevant with the Elo rating system, since the Elo ranking for a given year explicitly incorporates the complete trajectory of the journal's ranking performance until this specific year.

The rationale of the Elo rating system is the following. Each journal has an Elo number which is based on its impact. Every year, the journals compete with each other and earn Elo points which are based on the expected values for a win or a loss. After this competition, the Elo number is adjusted according to the result. In the upcoming years, the journals compete with each other based on their last available Elo numbers and therefore on expected values that change over time. The journals earn or lose Elo points that also vary over time, which generates a more dynamic ranking approach that not only decides between 'better' or 'worse'. In the end, the complete time path of the journal's ranking performance is relevant for the latest competition and therefore the latest ranking. The aim of this article is by no means an examination of the ranking's properties, but rather to present a new approach that is subsequently compared to rather standard rankings based on, for example, the latest SNIP (source normalized impact per paper) indicator or the Tournament Method. In the end we ask, whether the inclusion of the trajectory changes the current ranking of journals.

We base our analysis on a data set provided by Scopus for the period from 1999 to 2015, which contains more than 20 000 journals per year. In a first step, we build a balanced sample that only contains journals that have a SNIP available for the whole observation period. From this balanced set, we can state that the time line of a journal's ranking performance is very important for the most recent ranking. Our Elo approach produces a similar but by no means identical ranking compared to the Tournament Method, the average SNIP between 1999 to 2015 or the latest SNIP in 2015. With our approach, the top journals remain top-ranked. However, there are substantial differences observable for rather 'middle-

class' journals and not only for the top 30. We also show that a 'bad' year in terms of the SNIP does not necessarily lead to a large drop in the ranking position. In order to investigate a more realistic setting in the second step, we allow for entries and exits of journals. It turns out that the ranking of the journals from the balanced sample is preserved. Finally, we discuss the ranking results for different scientific categories and investigate the sensitivity of the latest ranking to a crucial parameter. The Elo rating system seems to be a promising alternative to rank scientific journals compared to existing ones. A further advantage is the possibility to apply the Elo ranking system to any journal metric, like the Journal Impact Factor or citation counts, that is published on a regular basis.

The paper is organized as follows. In Section 2 we elaborate on the data and the Elo ranking system. Section 3 presents and discusses the results. The last section offers some conclusions.

2 Data and Methodology

2.1 Data

General remarks. One aim of this paper is to present a new ranking approach for a wide range of journals from different scientific fields. Therefore, we need high quality and notably comparable data. Such high-quality data are available from Scopus at `http://www.journalmetrics.com`. The data, as of June 2016, are available for the period ranging from 1999 to 2015 and comprise 21 626 journals in 2015.

A main challenge is the comparability of journals across different disciplines. To this end, we use the SNIP (source normalized impact per publication) indicator (Moed, 2010; Waltman *et al.*, 2013). The strength of the SNIP lies in its normalization of citations in order to make scientific fields comparable. It especially pays attention to different citation practices within and between subjects. According to Moed (2010), the SNIP is basically the ratio of the so called raw impact per paper (RIP) and the Relative Database Citation Potential (RDCP) in the journal's sub-field. Whereas the RIP is defined as the number of citations in year t for papers published in the journal in the time span $t-3$ to $t-1$, the RDCP explicitly uses the distribution of citations. For each journal in the list, one can calculate its database citation potential (DCP). Repeating this step for each journal, results in a distribution of DCPs for the whole data set. In order to gain the RDCP, each journal's DCP is divided by the median DCP of the whole distribution.

Scopus also provides information on the scientific journal classification. We use these so called 'top levels' to present, on the one hand, category-specific rankings. And on the other hand, to investigate whether the category-specific journal order is independent from the size of the data set. It is desirable that the rank order of two journals A and B which are categorized as Social Sciences is identical for the whole data set as well as if we would apply

our ranking only to the bunch of social science journals. Scopus distinguishes between five top levels and a general category: Life Sciences, Social Sciences, Physical Sciences, Health Sciences and General. In the results section, we discuss the ranking for the whole data set as well as the rankings for each of the five categories.

Data preparations. We treat our data set in two different ways. First, we build our analysis on a balanced set of journals. A couple of SNIP entries in the original data set are missing. Since we want to have a ranking in this first step that is conducted for journals that have an impact over the whole observation period, we only use these journals that have a SNIP greater-than-or-equal to zero for all years from 1999 to 2015. This leaves us with 8 246 journals for the whole observation period. If we split up this balanced sample into our five categories, we end up with the following number of observations: 2 106 journals categorized as Life Sciences, 2 407 as Social Sciences, 3 173 as Physical Sciences, 2 513 as Health Sciences and 29 as General. Please note that the sum of these five categories does not equal the total number presented above since some journals are classified more than once.

Second, we repeat our calculations for an unbalanced data set. An unbalanced set of journals has the advantage to better map real conditions. Since journals enter or exit the data set, newcomers, for example, should be taken into account for the ranking. Thus, our unbalanced data set varies in its number of journals over time. The ranking in 2015, which we present later on, is based on 23 731 journals. Another advantage of taking an unbalanced sample is to check whether our ranking reacts strongly to newcomers or journals that exit. How we deal with these issues is described in the next section by introducing the Elo rating system.

2.2 The Elo Rating System

Fundamentals. Originally developed to rate chess players, the Elo rating system is nowadays adopted by many other sports such as table tennis (see, for example, Glickman, 1995) or used to, for example, rank evolutionary algorithms (Veček *et al.*, 2014). The eponym for this rating system is Arpad Emrick Elo, who was a Hungarian-born American physicist and statistician. His main objective was to develop a rating system for the United States Chess Federation (USCF) that has a statistical foundation. Later on, the rating system was also adopted by the Fédération Internationale des Échecs (FIDE), the world chess federation.

The two main steps of the ranking comprise (i) calculating the expected score and (ii) updating the player's rating (see here and henceforth Glickman and Jones, 1999). Additionally, we refer to Elo (1978) for a very detailed description. Since the inherent strength of a player is unknown to outsiders, one has to approximate it by a rating. Thus, the match outcome between two players A and B can be approximated with the following formula:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \, . \tag{1}$$

E_A is the expected score for player A to win the game, based on the unknown strengths for both players (R_A and R_B). To illustrate the expected score for player A, we use the example by Glickman and Jones (1999). Imagine a game between two players with strengths $R_A = 1\ 500$ and $R_B = 1\ 700$, respectively. The expected long-run score of player A is $E_A = 0.24$. Thus, based on these hypothetical strengths, player A is expected to win the game or gain a draw in 24 of 100 cases. The opposite is true for player B, since his expected score is $E_B = 0.76$. As mentioned, these figures are long-run scores. However, a game score can only take three possible values: 1 for a victory of player A, 0.5 if the game ended in a draw or 0 if player A loses the match. Since the strengths of both players are unknown, they are replaced by their estimates, the so called Elo number or Elo rating (for player A it is R_A).

The second step comprises the update of a player's strength. This is done by the following equation, again from player A's perspective:

$$R_{A,t+1} = R_{A,t} + k(S_A - E_A) \, . \tag{2}$$

The new Elo rating of player A ($R_{A,t+1}$) is based on his or her old rating ($R_{A,t}$) plus the difference from the game score S_A and the expected long-run score E_A, which is weighted by the factor k to allow how fast a rating can evolve. In chess, this factor is either based on the number of games played, the age of the player or the strength. Suppose that the Elo ratings of two players are $R_{A,t} = 1\ 500$ and $R_{B,t} = 1\ 700$ before they play a match. We set the adjustment parameter $k = 32$, which is mainly used in chess for weaker players. Three possible match outcomes can emerge and thus resulting ratings:

- A **wins:** $R_{A,t} = 1\ 500$, $S_{A,t} = 1$, $E_{A,t} = 0.24$, $R_{A,t+1} = 1\ 524$, $R_{B,t+1} = 1\ 676$,
- **Draw:** $R_{A,t} = 1\ 500$, $S_{A,t} = 0.5$, $E_{A,t} = 0.24$, $R_{A,t+1} = 1\ 508$, $R_{B,t+1} = 1\ 692$,
- A **loses:** $R_{A,t} = 1\ 500$, $S_{A,t} = 0$, $E_{A,t} = 0.24$, $R_{A,t+1} = 1\ 492$, $R_{B,t+1} = 1\ 708$.

As one can see, player A's rating either increases by winning the game or by gaining a draw since the expected long-run score of player A lies below the score for a draw ($0.24 < 0.50$). In the next match, the expected score is calculated based on the new Elo ratings. For the mathematics of such pairwise comparisons, for which the Elo rating system is a special case, we refer to Joe (1991).

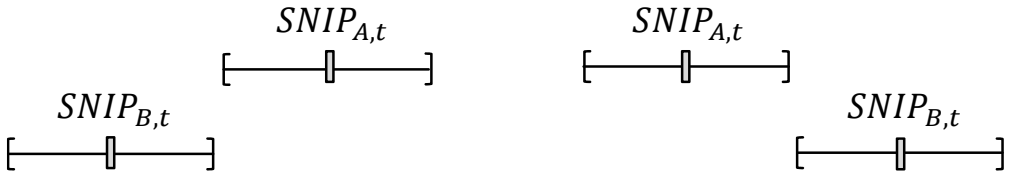
Application to rank journals. After the discussion of the fundamentals, it is the aim in the following to present how we apply the Elo rating system to rank journals. Therefore, we need to introduce parameter values: $R_{A,0}$, S_A and k. Each journal A is treated as a single

'player' at any point t in time. As for each sports or any other competition, the score S_A can take three values: 1 if journal A has a higher SNIP in year t compared to journal B, 0.5 if they are not statistically different from each other and 0 in the case of $\text{SNIP}_{A,t} < \text{SNIP}_{B,t}$.

The SNIP indicator is provided as three-digit number, which makes a perfect draw rather unlikely. Furthermore, it suggests an accuracy and clear-cut journal ranking that might not reflect the reality. Leydesdorff and Opthof (2010), Moed *et al.* (2012), and Vanclay (2012) called for confidence intervals to be provided for journal metrics. Such uncertainty measures can be found, for instance, in Schubert and Glänzel (1983), Nieuwenhuysen and Rousseau (1988), Opthof (1997), Greenwood (2007), Stern (2013), and Chen *et al.* (2014). Therefore, we decided to base our decision for a draw on official stability intervals, provided by CWTS Journal Indicators.[1] Basically, these stability intervals are based on bootstrapping and can be interpreted as 95% confidence bands, representing a range the SNIP fluctuates in. Thus, we can observe a lower bound for the SNIP of journal A at time t ($LB_{A,t}$), the SNIP itself and an upper bound ($UB_{A,t}$). Figure 1 presents our decision on a win or a loss of journal A against its competitor B.

Figure 1: Schematic representation to the decision of a win or a loss

(a) A wins against B (b) A loses against B



Journal A wins (loses) against journal B in a given year t if $LB_{A,t} > UB_{B,t}$ ($UB_{A,t} < LB_{B,t}$) holds. This is displayed in panel (a) in Figure 1 (or in panel (b) in case of a loss). If the stability intervals of two journals overlap, the match ends up in a draw. Additionally, we have to decide on two extreme events that can emerge by evaluating on a win or a loss. First, if the condition $SNIP_{A,t} = SNIP_{B,t} = 0$ is met, then no game is played between the two journals. And second, if the condition $LB_{A,t} = SNIP_{B,t} = 0$ holds, then also no game is played between the two journals. By including journals with a SNIP equaling zero would result in an inflationary effect of draws between these journal that have no impact at all. If a bulk of journals with no impact exist, all these journals would gain Elo points from draws by competing against each other. However, they should get no points since they have no impact at all.

We set the adjustment parameter to $k = 1$ in order to apply the same 'catch-up speed' for each journal from each scientific category. The main reason for this parameter value is the usage of the SNIP. Since this indicator is comparable between scientific categories as well as sub-categories of a single profession, we do not need to control for different citation patterns

[1]For a description of these stability intervals see `http://www.journalindicators.com/methodology`.

or anything similar that makes categories not comparable. We, however, will elaborate more on this point in Section 3.

The next parameter value we have to choose is the initial Elo number of each journal ($R_{A,0}$). It becomes immediately obvious that this number cannot be estimated from the data, thus, we decided to attribute each journal the same initial number in the case of a balanced panel: $R_{A,0} = 10\ 000$. Our resulting ranking is, however, independent from this initial value as we treat the time before 1999 as non-existing and let the journals be established in this year. Afterwards, the Elo numbers develop from this constant starting value. Choosing a different initial value that is also identical for all journals does not influence the ranking that results at the end of our data set. However, it should be a sufficiently large number to avoid negative Elo ratings. In the case of the unbalanced panel, we proceed in a different way. If a journal enters the competition as a newcomer, we cannot simply attribute a fixed value to it. Thus, we rather place it in the distribution of Elo numbers in the following way. Imagine a journal enters the competition in year t. It then plays a 'pre-tournament' against all journals that competed against each other in year $t-1$ in a first step, based on all rules mentioned before. Afterwards, we count the journal's fictive wins and draws and relate this value to the total number of matches played. We use this fictive share of wins and draws of the new competitor to calculate its position in the distribution of existing Elo numbers in $t-1$. The resulting number than serves as the Elo score of the new competitor in year t. In the case of observing a missing in the SNIP time series of a journal, the lastly observed Elo number of that journal is put forward. Thus, this journal is excluded from the competition in year t, but competes in $t+1$ with its Elo number of $t-1$. It holds: $SNIP_{A,t} = . \rightarrow R_{A,t} = R_{A,t-1}$.

Applying our notation to Equation (1) and (2), the expected long-run score of journal A to beat journal B and the corresponding update of journal A's Elo number transform into:

$$E_{A,t} = \frac{1}{1 + 10^{(R_{B,t} - R_{A,t})/400}} , \tag{3}$$

$$R_{A,t} = R_{A,t-1} + (S_{A,t} - E_{A,t}) . \tag{4}$$

Since our balanced data set comprises 8 246 journals, we have to calculate 8 245 pairwise comparisons for each journal and each year. As the number of journals varies over time in the case of the unbalanced data set, also does the number of pairwise comparisons. So the natural question to raise is: How does the Elo number develop between these pairwise comparisons? The answer can also be found in the chess system. The Elo rating is adjusted only once a year, after a journal has 'played' against all the other journals. Thus, the final Elo rating of a journal at the end of year t is: $R_{A,t} = R_{A,t-1} + \sum_{B \neq A} (S_{A,t} - E_{A,t})$. Each pairwise result is summed up and added to the previous Elo number at the end of all comparisons.

Another important point which we have to take care of is to set a maximum for the rating difference between two journals ($R_{B,t} - R_{A,t}$). If this difference is not restricted, the resulting

ranking becomes very volatile over time. We follow the official procedure by FIDE and allow the absolute value of this difference to be maximal 400, so as to hold: $|R_{B,t} - R_{A,t}| \leq 400$.[2]

Based on the Elo ratings in 2015 ($R_{A,2015}$), we calculate the overall ranking of all journals. At this point, our main contribution of the paper sets in: the Elo rating in 2015 incorporates the complete trajectory or history of the journal's ranking performance and thus produces a more realistic ranking. The Elo ranking is also unique since it is based on the continuous defined SNIP indicator. Thus, in the case of a win, the following must hold for a given year t: if $\text{SNIP}_{A,t} > \text{SNIP}_{B,t} > \text{SNIP}_{C,t}$, then '$A$ wins against B, C' $\wedge$ 'B wins against C'.[3] This is one main difference compared to sports, where it is also possible that 'C wins against A'.

2.3 An Alternative: the Tournament Method

An alternative approach using pairwise comparisons is the so called Tournament Method, which has been applied to rank economics journals by Kóczy and Strobel (2010). In the 'tournament', the journals compete in 'citation games' against each other. Thus, the ranking is based on cross-comparisons of citations between the journals.

In terms of our notation, the score $\sigma_{A,t}$ of journal A for a given year t is simply the share of games it wins against competitors or matches that end in a draw:

$$\sigma_{A,t} = \frac{|\{SNIP_{A,t} > SNIP_{B,t}\}| + \frac{1}{2}|\{SNIP_{A,t} = SNIP_{B,t} > 0\}|}{|\{SNIP_{A,t} + SNIP_{B,t} > 0\}|} \,. \tag{5}$$

A victory of journal A is defined as $SNIP_{A,t} > SNIP_{B,t}$. However, many different possibilities exist to identify the winner of a tournament in general (see Laslier, 1997). The main difference to the Elo rating system is that the relative position of a journal does not matter. In the tournament, a win of a 'bad' journal against a 'good' gives the same score as a win against a 'less good' journal. This issue is varied in the Elo system by introducing the expected value $E_{A,t}$, which is a continuous value between 0 and 1.

Kóczy and Strobel (2010) propose to account for the ranking's time line by applying a geometric decay function to calculate the total score of journal A:

$$S_{A,T} = \frac{1-\delta}{1-\delta^T} \sum_{t=1}^{T} \delta^{T-t} \sigma_{A,t} \,. \tag{6}$$

To be in line with Kóczy and Strobel (2010), we choose $\delta = 0.5$ in our application.

[2]The official statement can be found in the handbook on FIDE Rating Regulations effective from 1 July 2014 at: `https://www.fide.com/fide/handbook.html?id=172&view=article`.

[3]This statements holds if we abstract from draws and confidence bounds on journal metrics.

3 Results

3.1 Balanced Panel

In the following, we present our results for the balanced panel in two steps. First, we show and discuss the top 30 ranked journals based on the Elo numbers in 2015. And second, we close this section with some statements on how the different rankings for 2015 (based on either the Elo numbers, the Tournament Method or simply the SNIP) are correlated.

Let us start with the presentation of the top journals. Table 1 shows the top 30 ranked journals in ascending order, based on the Elo rating system as of 2015.[4] For reasons of comparison, we also include the ranks resulting from the Tournament Method in 2015, a ranking based on the average SNIP for the years 1999 to 2015 and the ranking of the latest available SNIP for 2015. The top 3 journals are: *New England Journal of Medicine*, *Reviews of Modern Physics* and *Chemical Reviews*. Only the *Reviews of Modern Physics* is among the top 3 by calculating rankings that are based on other indicators than the Elo number. Especially *Chemical Reviews* is only ranked sixth, if we base our decision on the latest SNIP indicator. However, the time dependency of the journal ranking performance and therefore the recent available ranking become much clearer by investigating much more volatile journals in Table 1: *The Lancet*, *CA – A Cancer Journal for Clinicians* and *Annals of Internal Medicine*. *The Lancet* is ranked third or fourth by either applying the SNIP of 2015 or the Tournament Method. If we, however, take the whole performance of the journal between the years 1999 to 2015 into account, *The Lancet* is only ranked on 12th place, based on its latest Elo number. More impressive are the differences between the rankings for the journal *CA – A Cancer Journal for Clinicians*. Whereas this journal is always top 3 ranked based on the Tournament Method or the SNIP indicator, it only reaches 24th place based on the Elo number of 2015. This result can be described by one fact: the allowance of draws in the competition. Since the SNIP of the *CA – A Cancer Journal for Clinicians* has very large stability intervals, it gains a lot of draws against other top journals, thus, its pure SNIP of 2015 would suggest the journal ranks on first place, but the Elo system more or less 'downgrades' its performance. The last example is *Annals of Internal Medicine*. This journal is ranked top 30 based on the Elo rating system. It, however, is not ranked that high by looking at the rankings that are based on the other three indicators. The SNIP 2015 would suggest to rank this journal on 67th place; the average SNIP between 1999 and 2015 would suggest 41st place. So here we can see what the Elo system does. Since the Elo 2015 ranking is based on the latest Elo number, which is by definition a function of the Elo numbers between 1999 to 2015, our ranking incorporates the whole trajectory of the journal's ranking performance over time. For most of the top journals, the position across different rankings show rather similar results. However, the different methodologies produce results

[4]The full ranking is available from the authors upon request.

Table 1: Top 30 ranked journals in 2015 for the balanced data set

Journal	Elo 2015	Tournament Method	Average SNIP (1999-2015)	SNIP 2015
New England Journal of Medicine	1	2	2	4
Reviews of Modern Physics	2	1	3	2
Chemical Reviews	3	5	4	6
Physiological Reviews	4	11	6	14
Annual Review of Immunology	5	14	5	16
JAMA – Journal of the American Medical Association	6	10	8	15
Nature	7	9	11	19
Science	8	13	13	25
Journal of Economic Literature	9	21	7	22
Annual Review of Biochemistry	10	26	9	33
Clinical Microbiology Reviews	11	18	12	24
The Lancet	12	4	19	3
IEEE Transactions on Pattern Analysis and Machine Intelligence	13	15	16	18
Endocrine Reviews	14	40	15	41
Annual Review of Plant Biology	15	24	17	38
Psychological Bulletin	16	22	22	29
Nature Genetics	17	19	28	31
Quarterly Journal of Economics	18	23	20	27
Cell	19	29	27	50
Progress in Energy and Combustion Science	20	8	10	12
Pharmacological Reviews	21	35	18	44
Physics Reports	22	12	23	11
Nature Medicine	23	31	30	43
CA – A Cancer Journal for Clinicians	24	3	1	1
Progress in Polymer Science	25	7	24	17
Advances in Physics	26	20	14	10
Annals of Internal Medicine	27	42	41	67
Accounts of Chemical Research	28	36	42	64
Chemical Society Reviews	29	17	37	23
Proceedings of the IEEE	30	39	33	69

Note: The journals are ordered according to the Elo ranking for 2015. *Source:* Data are taken from Scopus and are available at `http://www.journalmetrics.com`.

that are by no means identical. Our main criticism deals with the missing consideration of the time line or the history of a journal in most of the common rankings. The evidence from Table 1 strengthens our hypothesis that timely variation is very important for the ranking outcome. Even for the top 30 journals in our data set, we observe a certain degree of ranking heterogeneity.

The top 10 are dominated by journals from Physical Sciences, with the very general interest journals *Nature* and *Science* among those. The first journal from Social Sciences, the *Journal of Economic Literature*, is ranked 9, followed by the *Psychological Bulletin* on 16th place. From Table 1 we can also state that many different (sub)disciplines are part of the top 30. For instance, astronomy, economics, health sciences and physics are on the list. We will

elaborate more on the rankings of different scientific categories in the next section.

The last step we want to undertake is a formal statement on the relation between the different rankings. Therefore, we first calculate Spearman rank correlations for ranking-pairs. The outcome is shown in Table 2. We find the highest rank correlation of 0.996 between our Elo ranking and the Average SNIP 1999 to 2015. This is straightforward and underpins our idea of including the time path of a journal's performance. By comparing our ranking to the latest available SNIP, the correlation drops to 0.869, which is also the lowest rank correlation in Table 2. Thus, the rankings are by no means identical. This finding also supports our main criticism that the complete history of a journal has to taken into account for the latest ranking, which is underpinned by a low rank correlation for the pair 'Average SNIP - SNIP 2015'.

Table 2: Spearman rank correlation between different rankings

	Elo 2015	Tournament Method	Average SNIP (1999-2015)	SNIP 2015
Elo 2015	1.000			
Tournament Method	0.879	1.000		
Average SNIP 1999-2015	0.996	0.888	1.000	
SNIP 2015	0.869	0.872	0.872	1.000

Note: All rankings are based on the balanced sample. *Source:* Data are taken from Scopus and are available at `http://www.journalmetrics.com`.

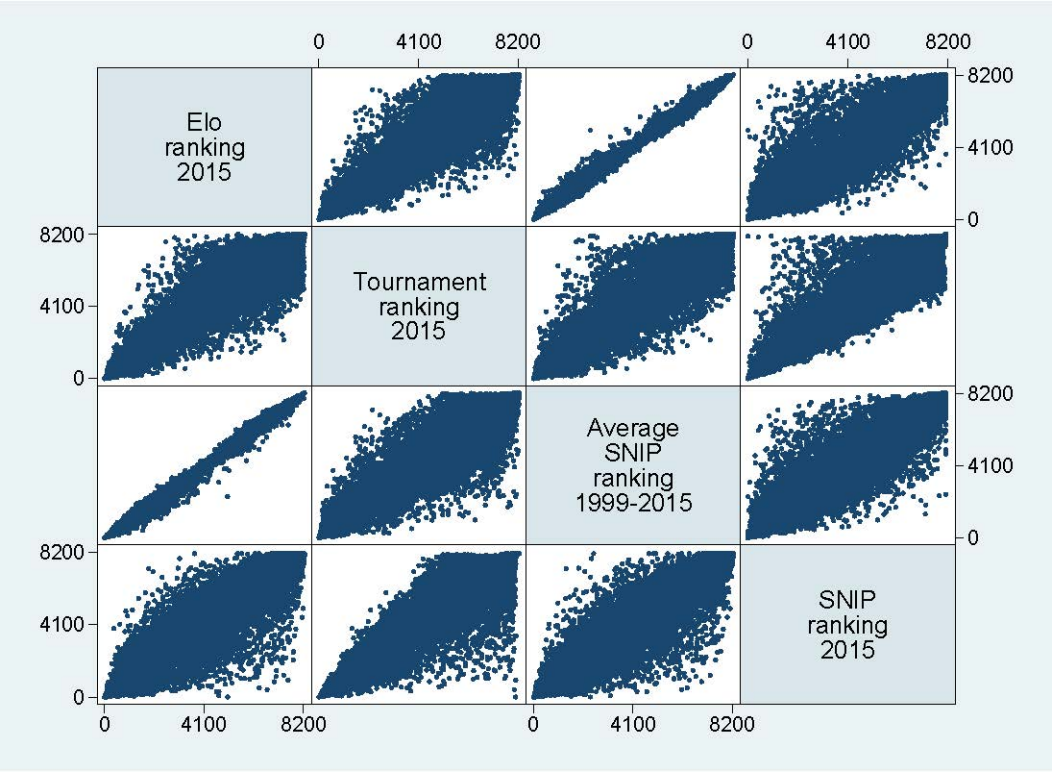
Figure 2 shows the relationships between the rankings in a graphical way. As suggested by the correlations, the rankings show a distinct linear relationship. However, we also observe a large mass of journals, and here especially in the middle, for which the methodologies deliver different ranking signals. Next to the introduction of draws that influence the journal's rank, also the performance of a journal over time that is incorporated in our Elo number heavily influences the latest calculable journal ranking.

3.2 Discussion

In this section we discuss the findings from the previous one. The discussion comprises three steps. First, we present the rankings for the five categories introduced in Section 2 and discuss differences and similarities in the results compared to the whole balanced data set. Second, we describe what happens to the ranking after allowing for entries and exits, thus, ranking the journals in an unbalanced data set. And third, we discuss a potential parameter sensitivity of our ranking.

Different scientific categories. As described previously, it is a preferable property that the journal's relative position in a ranking is not influenced by the size of the underlying data set.

Figure 2: Ranking cross-plots for 2015 between different rankings



Note: All rankings are based on the balanced sample.

We thus repeat our Elo competition for the five different categories separately: Life Sciences, Social Sciences, Physical Sciences, Health Sciences, and General. The methodology of the Elo rating is equivalent to the total balanced data set. We just take subsets and calculate Elo numbers. For example, the ranking for the journals classified as Social Sciences is just based on a competition between the journals of this category. In order to compare the ranks between the subsamples and the total data set, we rescale the ranks for the category-specific journals in the latter one. Table 3 shows the category-specific rank of the journal in the first column and the respective ranking in the total sample (last column).

The top 5 journals of a specific category are also, with few exceptions, ranked top 5 in the total sample. Taking Life Sciences as the example, the order of the first two journals changes between the total sample and the subsample. Whereas *Physiological Reviews* is the top Life Sciences journal in the total sample, it is replaced by the *Annual Review of Immunology* in the subsample. However, the Spearman rank correlation coefficients between the categorical rankings and the one for the total sample are all larger than 0.970. Thus, the rankings are de facto identical.

Unbalanced panel. Relying on a balanced sample may not be the best way to describe reality, since journals enter or exit the data set. So we explicitly have to take care of newcomers or leavers. To visualize the relationship between the 2015 ranking resulting either from the balanced or unbalanced sample, we again draw a cross-plot that is displayed

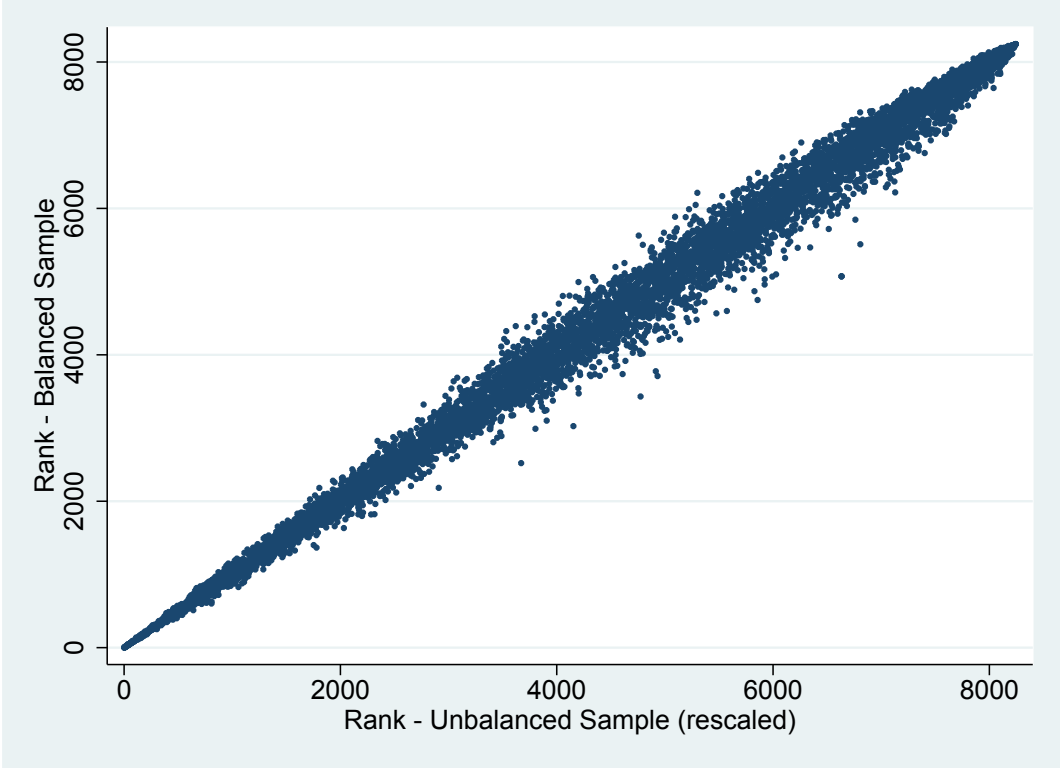
Table 3: Top 5 journals in 2015 for the five categories

Rank category	Journal	Rank total sample
	Life Sciences	
1	Annual Review of Immunology	2
2	Physiological Reviews	1
3	Nature	3
4	Annual Review of Biochemistry	4
5	Clinical Microbiology Reviews	5
	Social Sciences	
1	Nature	1
2	Journal of Economic Literature	2
3	Psychological Bulletin	3
4	Quarterly Journal of Economics	4
5	Academy of Management Review	6
	Physical Sciences	
1	Reviews of Modern Physics	1
2	Chemical Reviews	2
3	Nature	3
4	IEEE Transactions on Pattern Analysis and Machine Intelligence	4
5	Progress in Energy and Combustion Science	5
	Health Sciences	
1	New England Journal of Medicine	1
2	Annual Review of Immunology	3
3	Physiological Reviews	2
4	JAMA – Journal of the American Medical Association	4
5	Nature	5
	General	
1	Nature	1
2	Science	2
3	Proceedings of the National Academy of Sciences of the United States of America	3
4	International Journal of Bifurcation and Chaos in Applied Sciences and Engineering	4
5	Current Science	6

Note: The journals are ordered according to the Elo ranking for 2015 of the respective category. *Source:* Data are taken from Scopus and are available at `http://www.journalmetrics.com`.

in Figure 3. The x-axis contains the rescaled journal ranks from the unbalanced sample; the y-axis shows the ranks from the balanced sample. Both rankings are highly positively correlated (Spearman rank correlation: 0.996), thus, the difference between both is rather small. For the top 10 journals, the ranks are identical. The largest variation can be found for rather 'middle-class' journals, a result that confirms the findings from the previous section.

Figure 3: Ranking cross-plot for 2015 between the balanced and the unbalanced sample



Changing parameter k**.** One crucial parameter that could affect the ranking is k, which regulates the 'catch-up speed' between the journals. Our basic ranking for the balanced sample was derived by setting $k = 1$. Here, we want to show how the variation of this parameter influences the final Elo ranking in 2015. Therefore, we let k take values in the following range: $k \in \{2, 3, \cdots, 100, 10\ 000\}$. It turns out that the Spearman rank correlation between the pairwise-compared rankings is close to one, thus, they are almost identical. So in our case, setting $k = 1$ is no problem at all. We hypothesize that the 'catch-up effect' of k is treated towards zero by introducing the stability intervals and therefore the draws. Since such questions are beyond the scope of this paper, we leave such issues for future research.

4 Conclusion

Most of the commonly applied rankings for scientific journals mainly neglect the time line of a journal's ranking performance. This paper explicitly accounts for this shortcoming by transferring a concept that is widely accepted in chess, sports and other disciplines to the field of publishing: the Elo rating system. The data set on which we base our analysis comprises more than 20 000 journals from all possible scientific categories for the period from 1999 to 2015. In order to make the journals comparable, we use the source normalized impact per publication (SNIP) index. It turns out that the time line is very important for the latest ranking since the Elo rating system produces similar but by no means identical rankings

compared to outcomes based on either the Tournament Method, the average SNIP for 1999 to 2015 or the latest SNIP from 2015. Since the Elo ranking is very easy to compute and widely accepted in other fields, it seems a promising alternative to already existing ranking approaches.

References

Chen, K.-M., Jen, T.-H. and Wu, M. (2014). Estimating the accuracies of journal impact factor through bootstrap. *Journal of Informetrics*, **8** (1), 181–196.

Elo, A. E. (1978). *The Rating of Chessplayers, Past & Present*. Arco Publishing, New York NY.

Glickman, M. E. (1995). A Comprehensive Guide to Chess Ratings. *American Chess Journal*, **3**, 59–102.

— and Jones, A. C. (1999). Rating the chess rating system. *Chance*, **12** (2), 21–28.

Greenwood, D. C. (2007). Reliability of journal impact factor rankings. *BMC Medical Research Methodology*, **7** (1), 48.

Joe, H. (1991). Rating systems based on paired comparison models. *Statistics & Probability Letters*, **11** (4), 343–347.

Kóczy, L. A. and Strobel, M. (2010). *The World Cup of Economics Journals: A Ranking by a Tournament Method*. IEHAS Discussion Papers No. MT-DP – 2010/18.

Laslier, J.-F. (1997). *Tournament Solutions and Majority Voting*. Studies in Economic Theory 7, Springer-Verlag Berlin Heidelberg.

Leydesdorff, L. and Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology*, **61** (11), 2365–2369.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, **4** (3), 265–277.

—, Colledge, L., Reedijk, J., Moya-Anegon, F., Guerrero-Bote, V., Plume, A. and Amin, M. (2012). Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way. *Scientometrics*, **92** (2), 367–376.

Nieuwenhuysen, P. and Rousseau, R. (1988). A quick and easy method to estimate the random effect on citation measures. *Scientometrics*, **13** (1-2), 45–52.

Opthof, T. (1997). Sense and nonsense about the impact factor. *Cardiovascular Research*, **33** (1), 1–7.

Schubert, A. and Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, **5** (1), 59–73.

Stern, D. I. (2013). Uncertainty Measures for Economics Journal Impact Factors. *Journal of Economic Literature*, **51** (1), 173–189.

Vanclay, J. K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, **92** (2), 211–238.

Veček, N., Črepinšek, M., Mernik, M. and Hrnčič, D. (2014). A Comparison between Different Chess Rating Systems for Ranking Evolutionary Algorithms. In *Proceedings of the 2014 Federated Conference on Computer Science and Information System (FedCSIS)*, pp. 511–518.

Waltman, L., van Eck, N. J., van Leeuwen, T. N. and Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, **7** (2), 272–285.