Research Trends

Volume 1 | Issue 26 Article 3

1-1-2012

F1000 Journal Rankings: an alternative way to evaluate the scientific impact of scholarly communications

Sarah Huggett
Elsevier

Section 2: Value of Bibliometrics

F1000 Journal Rankings: an alternative way to evaluate the scientific impact of scholarly communications

Sarah Huggett, MPhil

In recent years the bibliometrics world has been booming with new metrics such as the h-index, EigenFactor, SJR, and SNIP. This expansion of the bibliometrics toolkit has been driven by the continued growth of scholarly content, combined with computational advances, growing global requirements for science to be measured and evaluated, and the problems of information overload and filter failure.

Before bibliometrics became widespread the evaluation of science was mainly performed through peer review. This more traditional approach was given a new breath of life when the Faculty of 1000 (F1000) was launched in 2002 to evaluate the quality of biomedical scientific articles based on the opinion of scientific experts. Initially, the papers were evaluated by 1,000 international Faculty members; now F1000 boasts more than 10,000 evaluators spread across 44 subject-specific Faculties. It is worth noting, however, that not all of the Faculty members are active or active to the same extent: **Research Trends** randomly checked the F1000 records of 20 members of the Reproductive Endocrinology Faculty, and although 4 have contributed more than 10 reviews, half have contributed only 1 or 2 evaluations yet, and a quarter have not made any recommendation yet. Jane Hunter, Managing Director at F1000, commented:

"Unsurprisingly, some Faculty Members (FMs) are more active than others and activity levels vary also depending on what other obligations FMs have on a month-by-month basis. Evaluation submission rates drop during congresses and rise immediately after (also unsurprising). Our most productive FMs select and evaluate 10–20 papers a year; our least productive may pick 1 or 2. F1000 has selected and evaluated 91,000 articles to date and these articles have attracted nearly 116,000 evaluations."

On average, 1,500 new articles are reviewed every month, which according to the F1000 website corresponds to about 2% of all published articles in the biological and medical sciences. This has led to some criticism of the journal rankings derived from the reviews, as overall they are based on very limited coverage of journal content[1,2]. For instance, according to Phil Davis, independent researcher and frequent blogger at the Scholarly Kitchen: "because of its limited scope of coverage, the real value of F1000 is not what the aggregate data can tell us about individual journals, but in what experts can tell us about individual articles." Looking at the 2010 provisional journal rankings, only 5 titles had more than 50% of their papers evaluated, less than 8% of journals had more than 10% of their articles reviewed, and more than 85% had less than 5% of their papers evaluated. According to Jane Hunter, however, this is not a problematic issue:

"F1000's purpose is to select and evaluate only the top papers in biology and medicine, so it follows that a relatively small percentage of papers from most journals will be included in our database. F1000's whole system is based on selectivity. This doesn't invalidate our Journal Rankings. Journals that publish relatively few papers judged as 'top' by our Faculty will have a lower FFj (F1000 Journal Factor) in our system and journals that publish a lot of top papers will have a higher FFj."

One of the unique aspects of the system at the time of the launch was that the ratings were not based on bibliometric data at the journal level, but on expert evaluation at level of individual articles. However, in 2011, in what has been labeled "a 180-degree turn"[1] F1000 started a new journal ranking system, including global journal rankings as well as rankings by subject area.

How does it compare to citations?

Citations are usually accepted as a measure of intellectual debt, and although there are negative citations the vast majority of citations are neutral or positive. This can be seen as roughly similar to the F1000 system, in which Faculty members can assign papers to one of three positive quality levels: Exceptional, Must Read, and Recommended. (Interestingly, there is no option to submit negative recommendations.) However, the similarity ends here: while citations are relatively easy to make (scientific papers routinely include dozens of references), reviews are more time-consuming to produce, and are therefore less numerous. Consequently, it can be argued that F1000 reviews have more weight (there are fewer of them) but also more bias (they can only be positive). However, Jane Hunter disagrees that the absence of negative evaluations introduces bias to the system:

"Negative reviews are simply not what we do. F1000 is a guide to what's best in science, not a thumbs up/thumbs down review service. There are plenty of comprehensive subject-area reviews published by other companies and we don't think the world needs another one from us. The fact that we only publish positive reviews doesn't

introduce bias into our system — it is our system. Our subscribers rely on us to tell them what they need to read and not what they need to avoid, so we will never publish negative evaluations. That said, we do publish dissents; if one of our FMs disagrees with another's article selection or with some aspect of an evaluation, he or she can submit a dissenting opinion, which is then published alongside the article's evaluation/s on our site. And we also allow registered subscribers to comment on evaluations or dissents, so if they have something to add we invite and encourage them to do so."

How does it work?

The F1000 Article Factor (FFa) can be calculated from one or several reviews, depending how many are available. If there are several recommendations for one article, the FFa is calculated from the highest rating, which bears a value of 10 for Exceptional, 8 for Must Read, and 6 for Recommended. An incremental value is then added for each of the other ratings (3 for Exceptional, 2 for Must Read, 1 for Recommended).

Research Trends was unable to find publicly available explanations for this methodology, and found it difficult to understand why these particular weights were chosen for initial and incremental values, but Jane Hunter was happy to explain:

"The values we assigned to our Recommended, Must Read and Exceptional ratings (6, 8 and 10) are arbitrary, but in essence reflect above-average scores on a 1–10 scale. The rationale for our calculation of total FFa for articles evaluated more than once is also arbitrary – and utilitarian – it made sense to us and seems to work."

This methodology however raises some concerns about the consistency of the FFa metrics – see example in text box. Furthermore, the FFa calculation gives more weight to the first highest rating and less weight to the following ratings, which has implications for the F1000 Journal Factor (FFj) derived from the FFas: more influence is given to articles with one recommendation compared to articles with several evaluations. As a consequence the FFj appears to be sensitive to enthusiastic reviewers rating numerous papers in small journals.[1]

Jane Hunter acknowledged this fact, but countered:

"This is not related to our weighting in favor of the highest score a paper receives from us or because we bias our system in favor of number of articles selected over number of evaluations (though we do, intentionally). It's because at the very specialist end of the scale where there are few journals and we have selected relatively few papers, a small number of additional reviews from a single journal can have a disproportionate impact on a journal's rank […] For future reference, we will be highlighting articles that have a declared competing interest on our main rankings journal pages in an upgrade planned for later this year. One important feature that sets us apart is complete transparency; our subscribers can easily see how each paper in F1000 was judged, by named experts, and review their reasoning. If there is a competing interest, it is clearly stated."

Consistency issue: let's look at some examples

Article A with two Exceptional scores would get an FFa of 13 (10 for the first Exceptional score + 3 for the second Exceptional score). Article B with three Must Read scores and one Recommended score would also get an FFa of 13 (8 for the first Must Read score, 2 for each of the other two Must Read scores, and 1 for the Recommended score), and so would article C with 8 Recommended scores (6 for the first Recommended score + 1 (×7) for the other Recommended scores).

So all three articles would get the same FFa of 13. Let's imagine now that each article receives one supplementary review (highlighted in red in below table), with an Exceptional score. This would result in article A getting an FFa of 16 (10 for the first Exceptional score and 6 (2 × 3) for the other two Exceptional scores, article B getting an FFa of 17 (10 for the Exceptional score + 6 (3 × 2) for the three Must Read scores + 1 for the Recommended score), and article C getting an FFa of 18 (10 for the Exceptional score + 8 (8 × 1) for the Recommended scores).

So while all articles initially had the same FFa, adding one same rating to each article causes differences in their ranking.

Article A									
Rating	Exc	Exc							FFa
Score	10	3							13
Article B									
Rating	MR	MR	MR	Rec					FFa
Score	8	2	2	1					13
Article C									
Rating	Rec	Rec	Rec	Rec	Rec	Rec	Rec	Rec	FFa
Score	6	1	1	1	1	1	1	1	13

Article A										
Rating	Exc	Exc	Exc						FFa	
Score	10	3	3						16	
Article B										
Rating	Exc	MR	MR	MR	Rec				FFa	
Score	10	2	2	2	1				17	
Article C										
Rating	Exc	Rec	Rec	Rec	Rec	Rec	Rec	Rec	Rec	FFa
Score	10	1	1	1	1	1	1	1	1	18

The FFj is calculated from the individual article ratings for a given journal, normalized according to the proportion of eligible scientific articles reviewed by the Faculty. The formula is as follows:

$FFj = \log_{10}\{(\text{Sum of Article Factors}) \times (\text{Normalization Factor}) + 1\} \times 10$

For each journal, the FFa scores are added to obtain the Sum of Article Factors. This sum is then normalized by the Normalization Factor, which is the percentage of articles evaluated by Faculty members compared to all scholarly articles published in the journal according to PubMed. Most bibliometrics indicators normalize for journal size using the number of articles published, but FFj's normalization is different: going back to our previous bibliometrics analogy, it is similar to multiplying the Impact Factor numerator by the percentage of cited papers rather than dividing it by the number of scholarly papers. This means that FFj's normalization does not actually account for journal size, but for journal coverage by F1000. For Jane Hunter, this is not a drawback but a benefit:

"Our normalization factor (number of articles selected by F1000/total number of eligible articles) introduces a variable representing journal coverage – or a journal's F1000 success rate – into our metric. The multiplier accounts for journal size, but it also rewards journals that have had relatively more articles selected by F1000. This is intentional. We want lots of evaluated papers to have a larger positive per-journal effect than a few very highly regarded ones. We believe publishing a lot of good articles is a more reliable indicator of a journal's value than its ability to publish the occasional megastar."

The values produced span over several orders of magnitude, so a log scale is applied, and this number is then multiplied by 10 to increase the readability of the final FFj.

Expert Opinion: Ludo Waltman comments

Research Trends spoke to Doctor Ludo Waltman, Bibliometrics Researcher at the Centre for Science and Technology Study at the University of Leiden, about the FFj's calculation:

"It seems that the developers of the F1000 system wanted to reduce the effect a single publication can have on the overall score of a journal. I guess this is why incremental recommendations have less weight than the initial recommendation. I understand this objective of avoiding 'outliers', but I think there are better ways to achieve this. For instance, the distinction between the initial recommendation and incremental recommendations could be abandoned, giving equal weight to all recommendations of the same type (e.g., all exceptional recommendations have a value of 10, including the incremental ones). To avoid outliers, the final score obtained by adding together the scores obtained from all recommendations a publication has received could be transformed – for instance, by using a square root or logarithmic function. This would also reduce the effect of a single publication with a lot of recommendations, but it has the advantage that consistency of the measurements is maintained. I also have some doubts about the normalization factor used in the calculation of the journal indicators. For instance, suppose we have two journals that each have 100 publications, and in each 50 publications have a single exceptional recommendation and 50 publications do not have any recommendation. This yields a journal score of (50 × 10) × (50%) = 250 for each of the two journals. (For simplicity, I skip the logarithmic transformation performed at the end of the calculations.) Suppose that the two journals are now merged. We then have a single journal with 200 publications, half of them with a single exceptional recommendation and half of them without recommendations. So the score of the merged journal becomes (100 × 10) × (50%) = 500. In other words, journals can increase their score by merging. This means that what is measured by the F1000 journal indicator is first of all the size of a journal (in terms of its number of publications). To obtain a high score, a journal must not only publish high quality articles (i.e., articles that receive recommendations), but it must also publish a large volume of articles. This is different from almost all citation-based journal indicators, such as Impact Factor, SNIP, and SJR (but not Eigenfactor), and most people probably will not be aware of this size-dependence of the F1000 journal indicator."

What type of rankings does F1000 compute?

Currently, there are three different journal rankings available:

- Current Journal Rankings: computed on the first day of each month, these are the most up-to-date as they include all evaluations over the previous 12 months, regardless of the publication date of the articles. For instance, February 2012 Current Journal Rankings take into account all recommendations made between 1 February 2011 and 30 January 2012.
- Provisional Annual Journal Rankings: calculated at the beginning of July, these are based on ratings of articles published in the preceding full calendar year. For instance, 2010 Provisional Annual Journal Rankings take into account evaluations made in 2010 and the first half of 2011 to articles published in 2010; 15 percent of evaluations are received 3 months after an article is published or later: as this adds an extra 3 months for ratings to accumulate, the disadvantage to articles published later in a year is decreased.
- Final Annual Journal Rankings: also computed at the beginning of July, these take into account evaluations of articles that were published in the last but one full calendar year, enabling the inclusion of 99 percent of potential evaluations for an article regardless of its publication date within a year. For instance, 2010 Final Annual Journal Rankings take into account evaluations made in 2010, 2011, and the first half of 2012 to articles published in 2010.

How does it compare to traditional bibliometrics indicators?

To see how FFj compares with traditional bibliometrics indicators, Research Trends ran a correlation analysis of 2010 Impact Factors versus 2010 provisional FFj for 768 journals mostly of biomedical scope (see Figure 1), in which the proportion of evaluated papers is denoted by the size of the bubble.

The correlation between the two metrics is rather weak overall (correlation coefficient of 0.54), and unsurprisingly at its weakest where only a small proportion of journal content has been evaluated. Yet this correlation does not systematically increase for journals where a high proportion of content has been reviewed. Some of the most noticeable outliers are also some of the journals with the highest Impact Factors (labeled in Figure 1). The analysis was replicated for EigenFactor (correlation coefficient of 0.55), SJR (correlation coefficient of 0.57), and SNIP (correlation coefficient of 0.51). The results presented similar patterns, indicating that bibliometrics indicators and F1000 journal rankings show a different picture of the research landscape: expert ratings seem to measure an alternative dimension to citations. This may be linked to the skewness of the citation distribution in any given journal.

Jane Hunter was not surprised by the results of the analysis: "We wouldn't expect F1000's FFjs to directly correlate with bibliometrics indicators – in fact if they did our rankings would be a lot less interesting […] Our metric is based entirely on positive evaluations of science, paper by paper, by panels of experts who read and select articles based solely on their intrinsic – and subjectively judged – importance. Another basic difference between F1000's metrics and the Impact Factor is that we exclude reviews […] Because of this, journals like Nature Reviews Drug Discovery […] will rank relatively low on F1000, as will any other journal whose Impact Factor is significantly affected by review articles."

At article level though, there are more similarities: indeed, Allen et al. found a "strong positive association between expert assessment and impact as measured by number of citations and F1000 rating". They, however, acknowledged that "despite the significant positive correlations between assessments of importance and citations overall, at the individual paper level the analysis showed that there are exceptions; papers that were highly rated by expert reviewers were not always the most highly cited, and vice versa. Additionally, what was highly rated by one set of expert reviewers may not be so by another set; only three of the six 'landmark' papers identified by our expert reviewers are currently recommended on the F1000 databases."[3]

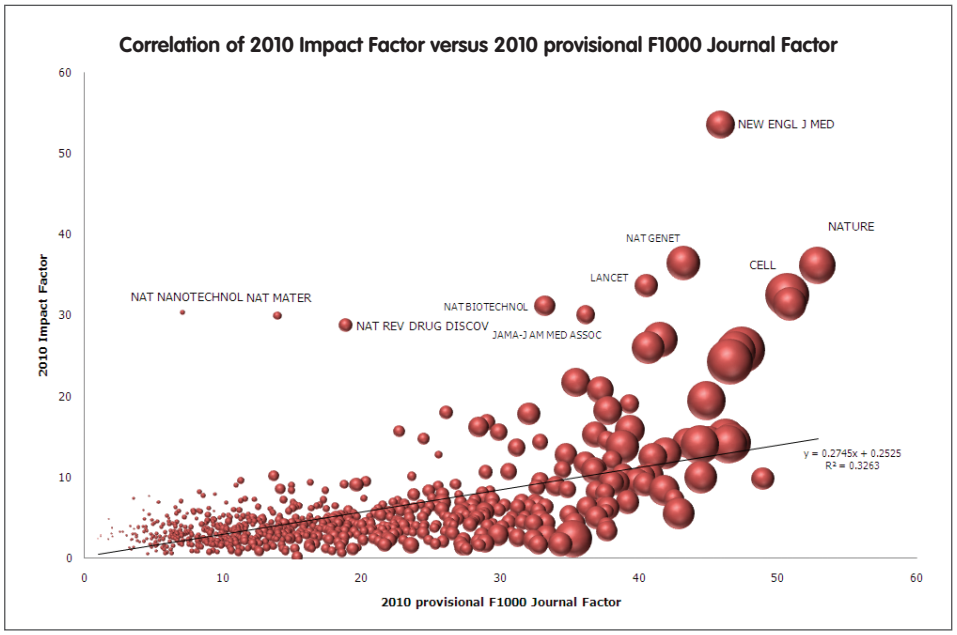

Figure 1 – comparison of 2010 Impact Factor versus 2010 provisional F1000 Journal Factor. Sources: 2011 Journal Citation Reports (© Thomson Reuters); F1000 2010 journal rankings.

Where do we go from here?

Jane Hunter offered some concluding remarks:

"We hope that the F1000 Journal Rankings will offer an alternate way of looking at and evaluating scientific success. The strengths and weaknesses of the various ranking systems may balance each other out and ultimately enable scientists to construct a truer picture of where to publish and what to read […] We know there are many ways in which the data generated by F1000 could be used and viewed. Our Article and Journal Factors represent just one way of crunching the individual article ratings allocated by Faculty Members and interpreting the results. The basic data are completely transparent and available on our site, and we're happy to consider other approaches. The numbers are the numbers, we think they're interesting, and we know they have other stories to tell." Further analyses are needed to help us understand the reasons behind our findings: in particular, it would be very interesting to see how FFjs relate to the distribution of article ratings for each journal. Doing some preliminary research for the article, **Research Trends** was actually surprised by the apparent lack of studies on the subject, and would therefore like to open a call for papers to the bibliometrics community: we'd love to see more research on F1000 FFa and/or FFj, in particular about their methodologies, or looking at comparison with other metrics. If you're up for it and would like to publish in **Research Trends**, just get in touch!

References:

1. Davis, P. F1000 Journal Rankings – The Map Is Not the Territory. Scholarly Kitchen blog post

2. Butler, D. (2011). Experts question rankings of journals. Nature 478, Vol. 20 doi:10.1038/478020a

3. Allen, L., Jones, C., Dolby, K., Lynn, D. & Walport, M. (2009). Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. PLoS ONE 4, e5910. doi:10.1371/journal.pone.0005910.

Links of interest:

- F1000 website http://f1000.com/
- Wikipedia entry http://en.wikipedia.org/wiki/Faculty_of_1000
- Scientist paper http://classic.the-scientist.com/article/display/57586/;jsessionid=0979DD1558B8EA321D99A115FCEECB66