Metrics to avoid – the impact factor

August 2020
bibliometrics@ucl.ac.uk

The UCL bibliometrics policy sets out some principles for the use of citation metrics in research assessment at UCL. As part of the overall guidance, this paper sets out some recommendations for metrics to avoid. Some very commonly used metrics are potentially very misleading, and we do not recommend using them in most circumstances. The two most common are the h-index and the impact factor.

What's wrong with the journal impact factor?
UCL's bibliometric policy places very few specific restrictions on metric use; instead it defers to the learned community's judgement under a general framework that guides flexible metric use among disciplines and purposes. However, one key aspect is eliminating the use of the journal impact factor (JIF) as an indicator for article quality.

To some researchers, this is initially controversial and challenging because the JIF has become engrained into many research cultures. But the JIF is an extraordinarily poorly used metric, which has led some commentators to suggest it may be the cornerstone of an unhealthy research culture with the potential to distort the scientific process. Misuse of the impact factor is perhaps the single largest force behind the drive for responsible metrics.

Because JIF is so commonly and fundamentally misused, and its misuse may damage the integrity of the research system, it is worth explaining why it is singled out in UCL's bibliometrics policy, and why discouraging its use is also one of the few shared principles among most external responsible metrics initiatives, such as DORA and the REF guidance.

So what's so wrong with one of the most ubiquitously used metrics?

1) **Aggregate does not equal individual.**
   JIF is an aggregate value calculated based on citations of the individual articles in a journal. But because citation counts are highly variable among articles, JIF cannot tell us anything about the quality of any given individual article.

2) **Skewed distribution**
   Some might argue that it is acceptable to use as a group mean, like JIF, as a rough indicator for citations of individual papers, because the citation counts of most papers will be around the mean, and on average variability will balance out. However, citation data of papers within a journal are almost always heavily skewed: most articles receive very few citations and a few articles are very highly cited. The mode, median, and mean for citation counts within a journal are usually significantly different values, and using the mean is unlikely to tell you anything about an individual paper. The JIF of most journals is dragged up by a few highly cited papers, and most articles receive relatively few citations regardless of where they are published.

3) **It's not calculated the way you think it is**
   The average number of citations per paper in a journal over two years sounds straightforward – except that's not *quite* how the JIF is calculated.

   Averages are usually calculated by dividing the sum of the values for a sample of observations (numerator) by the number of those observations (denominator). Note simple averages are usually symmetrical – which means the sum of the sample is based only on the observations counted in the denominator. But the JIF isn't calculated in this way.

   Instead, the numerator – sum of citations – is based on all of the citations received by items in a given journal. This includes articles and reviews, but also letters to the editor, comments, and other front matter that aren't primary research articles – even news and obituaries; while

the denominator is based not on the number of cited documents, but only articles and reviews. Hence, a journal's impact factor is driven not only by their research articles, but inflated by the other accompanying material in the journal. While this doesn't always get very heavily cited, it does usually add some extra citations.

4) **JIF is not even objective.**
   All quantitative and logic issues aside, at least the numbers behind JIF are objective? Except again – they're not. What constitutes citable material - whether something is primary research, front matter or another category – is negotiated by journals and the indexer of journals. These definitions are subjective and open to manipulation.

Skewed citation distributions and subjective calculation are enough to abandon JIF for inferring article quality. But if you're still not convinced, here's more problems which make it inappropriate.

1) **Short temporal window**
   The period of calculation is citations over two years. Such a small temporal window can lead to sensitive and highly variable metrics. For example, in a crystallography journal, a single massively-cited paper caused the journal's impact factor to leap from around 2 to around 50, and then drop back again two years later.

   Short temporal window is also inappropriate for most research fields because it takes much longer for a piece of work to be read, synthesized, appreciated and the built upon by the research community. Even in fast paced fields. For example, it takes 8 years for a paper to reach half of its eventual citations in the biomedical sciences.

   A short window also favours quick comment or turnaround of subsequent studies which may not be desirable in situations where it takes many years to research and design new work to examine the findings of a study to cite.

2) **Poor predictive tool**
   Building on the skewed distributions point, JIF is a poor predictive tool for future citations of an individual article. If you publish in a high impact factor journal, chances are you will get about as many citations as you would by publishing anywhere else. Indeed the most likely outcome remains that your paper will receive much fewer citations than the JIF – because the JIF is much higher than the median article citations.

3) **Matthew effect**
   Even though most papers are poorly cited, following journals because they've got a high JIF could lead to Matthew effects – the rich get richer – even though the quality of research is unchanged.

4) **Not transparent, nor reproducible.**
   The negotiations for what defines the citable material (and thus used to calculate JIF) is not public knowledge, and you can't reconstruct JIF from the data available. So we really don't know what exactly goes into any given JIF, or how reliable the calculation is.

5) **Other issues of metrics misuse present or exacerbated in JIF**
   As with many other metrics, it is important to remember that they only measure one thing, citations, and that those are a loose proxy of the abstract characteristic we're usually interested in – i.e., quality. There is strong variation in citation rates (and thus impact factors) among different fields and document types, which would mean that review journals or those in biomedical fields will tend to have higher impact factors. And when the metrics are incentivised and become targets in their own right, they lose meaning and is vulnerable to gaming lead to gaming

How did we get here? And where do we go?

In the pre-digital era, the unit of distribution for science was the physical journal volume. Libraries needed to make decisions on which journals to purchase and retain, and so the JIF was developed with no intention of reflecting research quality – but rather research readership and use. A journal with a high impact factor likely had a large number of potential readers, and the journal was likely to be

heavily used. This explains the numerator/denominator issue, because all citations in a journal is a better estimate of readership and journal use than only research articles or reviews. Further, bibliographic tools like Web of Science and Scopus did not exist, and citation counts per article were not accessible to most researchers.

In the absence of any more detailed information, an average number of citations over a defined period was a reasonable metric to determine the use of journals. However, the digital era has effectively unbundled the unit of research consumption – researchers can read and use a single article without any awareness of the articles within a journal.

Because the impact factor was for so long the *only* citation-based metric readily available, it became popular as a metric of quality – despite all the issues discussed above. But metrics are now easily attributed directly to the individual articles – we can count how many people are reading, downloading, and citing a journal article. This means that we no longer need to estimate the impact of papers when we can get that data directly, more informatively, and more accurately.

Further reading

https://arxiv.org/ftp/arxiv/papers/1801/1801.08992.pdf
https://www.sciencemag.org/news/2016/07/hate-journal-impact-factors-new-study-gives-you-one-more-reason
https://im2punt0.wordpress.com/2013/11/03/nine-reasons-why-impact-factors-fail-and-using-them-may-harm-science/
https://blogs.lse.ac.uk/impactofsocialsciences/2019/04/26/the-impact-of-the-journal-impact-factor-in-the-review-tenure-and-promotion-process/