Scholarly Publishing and Academic Resources Coalition
Advocating change in scholarly communications for the benefit of researchers and society

Better ways to evaluate research and researchers

A SPARC Europe BRIEFING PAPER

"We may say, by the way, that success is a hideous thing. Its counterfeit of merit deceives people […] Prosperity supposes capacity. Win in the lottery, and you are an able man."
— Victor Hugo[1]

Introduction

The most striking aspect of the recent series of Royal Society meetings on the Future of Scholarly Scientific Communication[2] was that almost every discussion returned to the same core issue: how researchers are evaluated for the purposes of recruitment, promotion, tenure and grants. Every problem that was discussed – the disproportionate influence of brand-name journals, failure to move to more efficient models of peer-review, sensationalism of reporting, lack of replicability, under-population of data repositories, prevalence of fraud – was traced back to the issue of how we assess works and their authors.

It is no exaggeration to say that improving assessment is literally the most important challenge facing academia. Everything else follows from it. As shown later in this paper, it is possible improve on the present state of the art.

Measure what you want to improve

The problems are caused by short-cuts used to assess the quality of research and researchers. For example, the impact factor of the journal where a study is published is often used as a proxy for the quality of the research and therefore of the researcher. Even if journal impact factor were a good proxy, this practice would be harmful because rational researchers optimise their behaviour according to the criteria of evaluation. For this reason, some workers can invest as much effort in chasing publication in high-impact-factor journals as they do on their actual research. From the perspective of the broader goal of research – improving society – this effort is literally wasted. How can we do better?

Ideally, we would evaluate each work on its own merits, taking into account expert opinions, and ignoring numeric metrics. These after all are only proxies for the things we really care about: rigour, correctness, replicability, honesty.

Ideally, we would evaluate each work on its merits, taking into account expert opinions, ignoring numeric metrics.

In practice, this is simply not possible. For logistical reasons, metrics *are* going to be used whether they are good for the

1 Hugo, Victor. 1862. *Les Miserables* volume 1, book 1, chapter 12, "The Solitude of Monseigneur Welcome". Translated by Isabel F. Hapgood.

2 The Royal Society. 2015. *The future of scholarly scientific communication Conference 2015*. https://royalsociety.org/~/media/events/2015/04/FSSC1/FSSC-Report.pdf

community or not. For example, consider a search committee with the best intentions of recruiting on the quality of research as assessed by expert reading and interview. Even in this case, metrics will surely be used in the early stages of the recruitment process when sifting a pool of 500 applicants down to a shortlist for interview.

This being so, it's crucial that the metrics we use are "honest signals" – good, robust proxies from the things that we really care about – so that researchers who concentrate on improving their own metrics are thereby incentivised to do genuinely good work. If we measure and reward statistical robustness, we will get statistically robust research. If we measure and reward the ability to present research sensationally, we will get sensationalism, and even fraud.

Why are the current metrics not up to the job?

At present, two metrics dominate evaluation: impact factor (IF) and H-index. Both have flaws and their widespread use has some serious negative consequences, especially when used incorrectly. Both measures were created by inventors who never imagined the ways their creations would be misused.

The Journal Impact factor (JIF)

The impact factor was intended as a measure of the impact of a journal. Created by Eugene Garfield in 1972[3], its intended purpose was to help librarians in deciding which journals to subscribe to. It is defined as the average number of citations over the preceding two years for each citable article published in the journal.

Even as a measure of journal impact, the JIF is badly flawed, as it uses a simple mean of citation counts. This results in skewed results when a very small proportion of papers are cited disproportionately often – as happened with *Acta Crystallographica Section A* in 2009, when its impact factor leapt from 2.051 to 49.93 due to a single highly-cited review article[4]. As a result, counter-intuitively, there is no statistically significant correlation between the citation count of a paper and the IF of the journal that it appears in[5]. Neither is there significant correlation between a journal's impact factor and the statistical power of the articles that appear in it[6]. On the other hand, there *is* a significant correlation between impact factor and retraction rate: articles appearing in high-IF journals are more likely to be retracted than those in regular journals[7]. This may be partly a consequence of the greater scrutiny that articles in high-IF journals are subjected to, but is also due to the pressure on authors to present their findings in the most sensational possible light in order to secure publication in these journals.

As a measure of journal quality, then, the impact factor is not perfect; but at least in this context it's a bad measure of the right thing. Far worse is its prevalent use as a proxy measure of the quality of an article; and worst of all is the unhappily common situation where a researcher is evaluated by the impact factors of the journals in which her work has appeared: a measurement two steps removed from the thing we actually want to measure.

---

3 Garfield, Eugene. 1972. Citation analysis as a tool in journal evaluation. *Science* **178**:471–479.

4 Schwarzenbach, Dieter, Gernot Kostorz, Brian McMahon and Peter Strickland. 2010. SHELX makes an impact. International Union of Crystallography leading article for 12 July 2010. http://www.iucr.org/home/leading-article/2010/2010-07-12

5 Seglen, Per O. 1997. Why the impact factor of journals should not be used for evaluating research. *BMJ* **314(7079)**: 498–502. See figure 3.

6 Brembs, Björn, Katherine Button and Marcus Munafò. 2013. Deep impact: unintended consequences of journal rank. *Frontiers in Human Neuroscience* **7:291**. doi:10.3389/fnhum.2013.00291. See Figure 2.

7 Brembs, Björn. 2011. High impact factors are meant to represent strong citation rates, but these journal impact factors are more effective at predicting a paper's retraction rate. *LSE Impact of Social Sciences*. http://blogs.lse.ac.uk/impactofsocialsciences/2011/12/19/impact-factor-citations-retractions/

H-index

The H-index, or Hirsch-index, is named after its creator Jorge E. Hirsch, who proposed it in 2005 as a measure of how widely cited an author is[8]. It is defined as the largest integer *n* for which the author has published at least *n* papers, each with at least *n* citations. Perhaps because this definition is simple and intuitive, the H-index is now widely used – for example, Google Scholar tracks it for each indexed author – and is often factored into researcher evaluations.

Unfortunately, like many simplistic measures, the H-index suffers from several flaws – for example, it takes no account of the difference between a sole-authored paper and one on which the author appears third in a list of six. Most damningly, it has been shown that a researcher's H-index is strongly correlated with the square root of the total number of citations[9]. Thus a rational researcher wanting to increase her H-index will simply publish more papers in search of more citations – an outcome very much at odds with the prevailing wish for fewer and better papers.

Since these flaws in the H-index are well understood, numerous modified versions have been proposed[10]. None, however, has yet established a foothold.

What are the alternatives for measuring publishing impact?

More egregious than the inherent flaws of the Journal Impact Factor and H-index is their pervasive misuse. Each of them is, by design, a measure of only one thing: citations. Yet they are used as the sole factor in evaluations of journals and researchers, even though many other dimensions of evaluation are relevant.

Evaluation criteria for journals

Rather than evaluating and ranking journals purely on the very flawed measure of Impact Factor, the following criteria should be taken into account. All of these pertain to the clear, rapid and secure communication of good, reliable research – and therefore of service to authors, other researchers, and the broader community:

- absence of arbitrary limits on length and figures
- support for high-resolution colour illustrations, video and other media
- usability of submission system
- speed of editorial handling
- transparency of editorial handling (notification emails, etc.)
- rigour of the peer-review filter
- helpfulness of peer-review in improving manuscripts
- provision of editorial services such as copy-editing
- speed of production after acceptance
- page design
- openness of publication
- functionality of journal website
- indexing
- archiving in PubMed Central, LOCKSS, etc.
- provision of post-publication review facilities
- adherence to codes of practice (e.g. COPE, OASPA)

For journals, these include: absence of arbitrary limits on length and figures; support for high-resolution colour illustrations, video and other media; usability of submission system; speed of editorial handling; transparency of editorial handling (notification emails, etc.); rigour of the peer-review filter; helpfulness of peer-review in improving manuscripts; provision of editorial services such as copy-editing; speed of production after acceptance; page design; openness of publication; functionality of journal website; indexing; archiving in PubMed Central, LOCKSS, etc.; provision of post-publication review facilities; and adherence to codes of practice (e.g. COPE[11], OASPA[12]).

8 Hirsch, Jorge E. 2005. An index to quantify an individual's scientific research output. *PNAS* **102(46)**:16569–16572. doi:10.1073/pnas.0507655102.

9 Yong, Alexander. 2014. Critique of Hirsch's citation index: a combinatorial Fermi problem. *Notices of the AMS* **61(9)**:1040–1050.

10 Alonso, S., F. J. Cabrerizo, E. Herrera-Viedma, F. Herrera. 2009ff. H-index and variants. http://sci2s.ugr.es/hindex

11 Committee on Publication Ethics, http://publicationethics.org/

12 Open Access Scholarly Publishers Association, http://oaspa.org/

Even when considering only activity related to publishing research, evaluation of researchers should include: significance of the subject investigated; clarity of writing; rigour of experimental design; replicability of methods; reproducibility of results; statistical strength; validity of conclusions; adherence to ethical codes; openness of publications; service on editorial boards; and participation in peer-review. If we do not measure these aspects of scholarly behaviour, we will not reward them; and so researchers will have a reduced incentive for such desirable behaviours. (This is already apparent in the increasing difficulty of soliciting peer-reviews, as reviewing takes time and brings little career reward.)

If we do not measure these aspects of behaviour, we will not reward them, and so researchers will have a reduced incentive for such desirable behaviours.

Ways of measuring impact not related to publishing

Beyond revising and expanding our measurement of researchers' publishing activity, it is surely also desirable to measure (and so incentivise) other aspects of scholarly behaviour such as: mentoring early-career researchers; collaborating fruitfully with peers; serving scholarly societies; and public engagement.

Evaluation criteria for researchers

Apart from the quality of their research, as assessed by the criteria in callout 2, researchers should also be evaluated according to the following additional criteria:

- service on editorial boards
- participation in peer-review
- mentoring early-career researchers
- collaborating fruitfully with peers
- serving scholarly societies
- public engagement

The motivation of the "altmetrics" (alternative metrics) movement is to capture more of these aspects of quality – since in practice measuring is the first step towards rewarding. At present, the organisations recording and promoting altmetrics – Altmetric[13] Impact Story[14], Plum Analytics[15], etc.– are concentrating primarily on measures of articles rather than of journals or researchers. As a result, their work does not *directly* address the problem of improving on either impact factor or H-index, but there are obvious methods of aggregating article-level scores across either the journals that publish those articles or the researchers who write them.

Which metrics to use

At present, altmetrics efforts are largely focussed on collecting all available data more or less indiscriminately. However, as noted above, some metrics are inherently misleading, and should be dropped or modified. For example, if something along the lines of an impact factor is to be taken into account, then it should probably be modified to use the median citation count rather than the mean, giving a truer representation of how frequently a typical article in a journal is cited.

It is also desirable to measure (and so incentivise) other aspects of scholarly behaviour such as mentoring early-career researchers; collaborating fruitfully with peers; serving scholarly societies; and public engagement.

[13] https://www.altmetric.com/
[14] https://impactstory.org/
[15] http://www.plumanalytics.com/

Similarly, we should beware of metrics that depend largely on luck. For example, we may be tempted to count how many patents a researcher's work generates. But if two researchers run equally replicable tests of similar rigour and statistical power on two sets of compounds, but one of them happens to have in her batch a compound that turns out to have useful properties, should her work be credited more highly than the similar work of her colleague?

We should also seek metrics that assess researchers over long periods where possible, to avoid penalising those whose research is longer-term in nature or women who, due to maternity leave, publish no papers in a given year. Short-termism in evaluation will inevitably result in researchers optimising their short-term outcomes at the expense of long-term progress.

Aggregation of multiple metrics

Although evaluation of journals and articles is important, the way researchers are evaluated is of far greater importance, because researchers will respond with changes in behaviour to optimise the measurements in use. Therefore, we will concentrate in the remainder of this paper on the problem of assessing researchers.

Evaluation criteria for papers

Evaluating papers by the journal in which they appear is not only misleading but also results in perverse incentives. Instead, published research papers should be evaluated according to criteria that directly contribute to the quality and reliability of the research:

- significance of the subject investigated
- clarity of writing
- rigour of experimental design
- replicability of methods
- reproducibility of results
- statistical strength
- validity of conclusions
- adherence to ethical codes
- openness of publication

The Altmetrics Manifesto[16] envisages no single replacement for any of the metrics presently in use, but instead a palette of different metrics laid out together. Administrators are invited to consider all of them in concert. For example, in evaluating a researcher for tenure, one might consider H-index alongside other metrics such as number of trials registered, number of manuscripts handled as an editor, number of peer-reviews submitted, invited conference presentations, total hit-count of posts on academic blogs, number of Twitter followers and Facebook friends, and potentially many other dimensions.

In practice, it may be inevitable that overworked administrators will seek the simplicity of a single metric that summarises all of these. Given a range of metrics x_1, x_2 … x_n, there will be a temptation to simple add them all up to yield a "super-metric", $x_1 + x_2 + \ldots + x_n$. Such a simply derived value will certainly be misleading: no-one would want a candidate with 5,000 Twitter followers and no publications to appear a hundred times stronger than one with an H-index of 50 and no Twitter account.

A first step towards refinement, then, would weight each of the individual metrics using a set of constant parameters k_1, k_2 … k_n to be determined by judgement and experiment. This yields another metric, $k_1 \cdot x_1 + k_2 \cdot x_2 + \ldots + k_n \cdot x_n$. It allows the down-weighting of less important metrics and the up-weighting of more important ones.

The Altmetrics Manifesto envisages no single replacement for any of the metrics presently in use, but instead a palette of difference metrics laid out together.

However, even with well-chosen k_i parameters, this better metric has problems. Is it really a hundred times as good to have 10,000 Twitter followers than 100? Perhaps we might decide that it's only ten times as good

[16] Priem, Jason, Dario Taraborelli, Paul Groth and Cameron Neylon. 2010. *Altmetrics: a manifesto*. http://altmetrics.org/manifesto

– that the value of a Twitter following scales with the square root of the count.

Conversely, in some contexts at least, an H-index of 40 might be more than twice as good as one of 20. In a search for a candidate for a senior role, one might decide that the value of an H-index scales with the square of the value; or perhaps it scales somewhere between linearly and quadratically – with H-index$^{1.5}$, say. So for full generality, the calculation of the "Less Wrong Metric", or LWM for short, would be configured by *two* sets of parameters: factors k_1, k_2 … k_n, and exponents e_1, e_2 … e_n.

Then the formula would be:
$LWM = k_1 \cdot x_1^{e1} + k_2 \cdot x_2^{e2} + \ldots + k_n \cdot x_n^{e}n$

Choosing the parameters for the Less Wrong Metric

How should the parameters for this general formula be chosen? One approach would be to start with subjective assessments of the scores of a body of researchers – perhaps derived from the faculty of a university confidentially assessing each other. Given a good-sized set of such assessments, together with the known values of the metrics x_1, x_2 … x_n for each researcher, techniques such as simulated annealing can be used to derive the values of the parameters k_1, k_2 … k_n and e_1, e_2 … e_n that yield an LWM formula best matching the subjective assessments.

Where the results of such an exercise yield a formula whose results seem subjectively wrong, this might flag a need to add new metrics to the LWM formula: for example, a researcher might be more highly regarded than her LWM score indicates because of her fine record of supervising doctoral students who go on to do well, indicating that some measure of this quality should be included in the LWM calculation.

Summary

Most of the problems afflicting research (short-termism, the disproportionate influence of brand-name journals, failure to move to more efficient models of peer-review, sensationalism of reporting, lack of replicability, under-population of data repositories, prevalence of fraud) are traceable directly to the perverse incentives offered by the way researchers are evaluated. The dependence on impact factors and H-indexes exemplifies this problem, but it is far more pervasive than these two measures alone.

To change researchers' behaviour to favour the outcomes we care about, it is necessary to change how they are assessed, and ensure they are rewarded for qualities that benefit the research community and wider society. To this end, we propose a schema for a "Less Wrong Metric" or $LWM = k_1 \cdot x_1^{e1} + k_2 \cdot x_2^{e2} + \ldots + k_n \cdot x_n^{e}n$, which takes into account many measurements, weighted by means of factors and exponents that can be optimised experimentally.

The goal is that rational researchers who wish to optimise their LWM score will do so by writing good papers, publishing them openly, helping colleagues, engaging with the media – in short, by becoming better researchers.

Alternative metrics currently in use

The following alternative metrics are currently in use by the three best-known altmetrics vendors. Many others are possible, related to the qualities listed in the first three callouts. This table is limited to metrics used for papers, omitting for example the "SlideShare downloads" metric that Impact Story tracks for slide decks. It also omits journal-specific metrics such as PLOS page views and downloads.

	Altmetric	Impact Story	Plum Analytics
#mentions from medics	✓		
#mentions from reporters	✓		
#mentions from researchers	✓		
#mentions from the public	✓		
Abstract views			✓
Blog citations	✓	✓	✓
Citation saves/exports			✓
Citations according to CrossRef			✓
Citations according to PubMed		✓	
Citations according to PubMed Central			✓
Citations according to Scopus		✓	✓
CiteULike users	✓	✓	
Connotea users	✓		
Delicious users	✓	✓	✓
Downloads (across many services)			✓
Dryad data downloads		✓	
Dryad package views		✓	
External peer-reviews	✓		
Facebook comments			✓
Facebook likes			✓
Facebook posts	✓	✓	
Facebook shares			✓
FigShare downloads		✓	
FigShare shares		✓	
FigShare views		✓	
Figure views			✓
Forum citations	✓		
Full-text views			✓
GitHub collaborators			✓
GitHub forks		✓	✓
GitHub stars		✓	
GitHub watchers			✓
Goodreads reviews			✓
Goodreads users			✓
Google+ +1s			✓
Google+ posts	✓	✓	
Libraries with holdings			✓
LinkedIn users	✓		
Mendeley groups		✓	
Mendeley users	✓	✓	
Mendeley users by career stage		✓	
Mendeley users by country		✓	
Mendeley users by discipline			
News reports	✓		

Opaque aggregate score	✓		
Patents citing work			✓
Pinterest posts	✓		
PubMed editorial citations		✓	
PubMed review citations		✓	
PubMed reviews (on F1000)		✓	
Reddit comments			✓
Reddit posts	✓		
Reddit score (upvotes – downvotes)			
StackExchange mentions	✓		✓
Supporting-data views			✓
Twitter mentions	✓	✓	✓
Wikipedia pages	✓	✓	✓
YouTube videos	✓		

This briefing paper was written for SPARC Europe by
Dr Michael P. Taylor
Department of Earth Sciences, University of Bristol, UK
dino@miketaylor.org.uk

SPARC Europe
98 Watermanstraat, Apeldoorn, The Netherlands

www.sparceurope.org