COMMENTARY

Measures for measures

Are some ways of measuring scientific quality better than others? **Sune Lehmann**, **Andrew D. Jackson** and **Benny E. Lautrup** analyse the reliability of commonly used methods for comparing citation records.

Although quantifying the quality of individual scientists is difficult, the general view is that it is better to publish more than less and that the citation count of a paper (relative to citation habits in its field) is a useful measure of its quality. How citation counts are weighed and analysed in practice becomes important as publication records are increasingly used in funding, appointment and promotion decisions. Typically, a scientist's full citation record is summarized by simpler measures, such as average citations per paper, or the recently proposed Hirsch index[1], which is ever more being used as an indicator of scientific quality[2]. Despite their growing importance, there have been few attempts to discover which of the popular citation measures is best and whether any such measure is statistically reliable.

> "There have been few attempts to discover which of the popular citation measures is best and whether any are statistically reliable."

Measures of citation quality are of value only if they can be assigned to individual authors with high confidence. Previous bibliometric studies[2] have compared different measures of scientific quality, but just because two measures agree does not mean that either one is accurate or reliable. We will argue that some citation-based measures can provide useful information given data of sufficient quality, but others fail to meet minimum acceptable standards. This should concern every working scientist.

Unfair discrimination

Because citation practice differs markedly between disciplines and subfields, a homogeneous set of authors is essential for any statistical analysis of citations. Here we use data from the theory section of the SPIRES database in high-energy physics, which has the requisite homogeneity[3]. Within this database, the probability that a paper will receive k citations falls slowly with increasing k and is described by a power-law distribution, a/k^{γ} with $\gamma \approx 2.8$, for large k. This long-tailed distribution has a number of consequences. About 50% of all papers have two or fewer citations; the average number of citations is 12.6. The top 4.3% of papers produces 50% of all citations whereas the bottom 50% of papers yields just 2.1% of all citations. Measuring an author's mean or median citation count per paper probe different aspects of their full citation record: which is better? Fortunately, this question can be posed in a way that yields a statistically compelling answer.

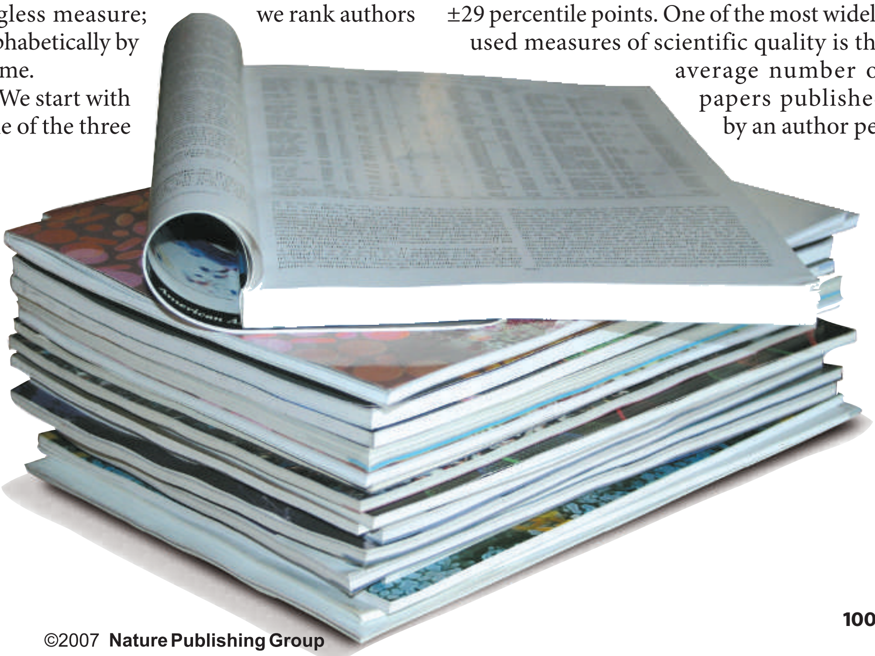
The purpose of comparing citation records is to discriminate between scientists. An author's citation record is a list of the number of citations of each of the author's publications. Until reduced to a single number, this list cannot provide a means of ranking scientists. But whatever the intrinsic merits of the chosen number, it will be of no practical use unless the uncertainty in assigning it to individual scientists is small. From this perspective, the 'best' measure will be that which minimizes uncertainty in the values assigned and hence maximizes discrimination between individuals. We analyse three measures of author quality: mean number of citations per paper, number of papers published per year, and the Hirsch index. A scientist is said to have Hirsch index h if h of their total, N, papers have at least h citations each, and the remaining $(N-h)$ papers have fewer than h citations[1]. For this study, we adopt Hirsch's assumption that h divided by N "should provide a useful yardstick". To calibrate our results, we also consider an obviously meaningless measure; we rank authors alphabetically by name.

We start with one of the three measures we had chosen and sort the SPIRES authors into decile bins. We use the full citation records for all authors in a given bin, n, to calculate the conditional probability that a paper written by an author in bin n will have k citations. From these conditional probabilities, we use Bayes' theorem to determine the average probability that an author initially assigned to bin n should instead be assigned to bin m. (To do this, we calculate the probability that the full publication record of each author in bin n was drawn, at random, on the conditional probability appropriate for bin m; see Supplementary Information.) Because the m assignment is based on an author's full citation record, it is more reliable than the n assignment. This process is repeated for each decile bin and for each measure considered.

Quality testing

A perfect measure of author quality would place all weight in the diagonal entries of a plot of m versus n (Fig. 1, overleaf). The better the measure, the more weight will be found in the diagonal boxes. Figure 1 reveals that both accuracy and certainty are sensitive to the choice of indicator.

An alphabetical ranking of authors contains no information regarding scientific quality, and so every author is assigned to every decile with equal probability (Fig. 1a). The resulting root-mean-square (*rms*) uncertainty in author assignment thus has the maximum value of ±29 percentile points. One of the most widely used measures of scientific quality is the average number of papers published by an author per



A. MACDONALD

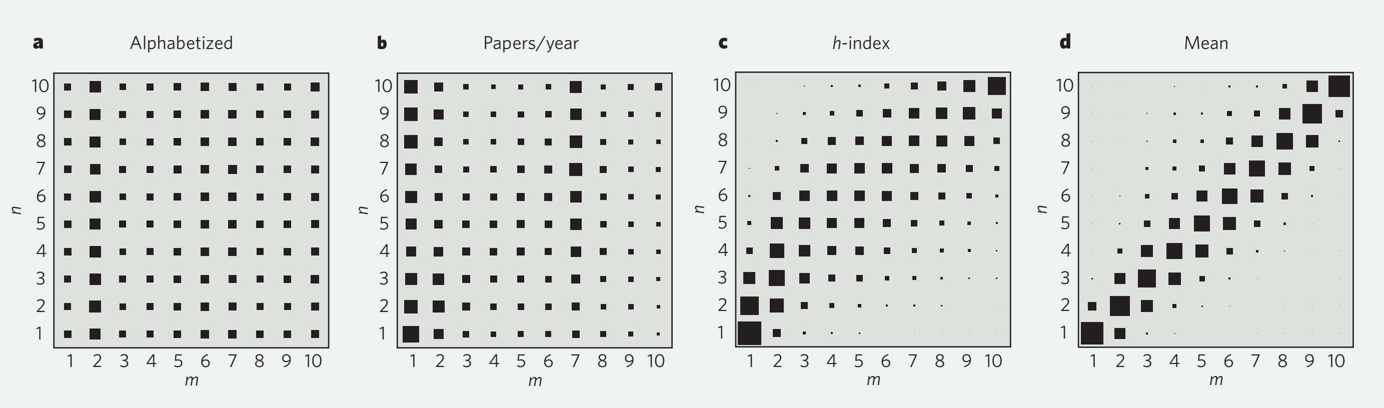


Figure 1 | The probabilities for four different measures. a–d, Each horizontal row, indexed by *n*, shows the average probabilities that authors initially assigned to a given decile bin *n* are predicted to lie in a different decile bin *m*. The probabilities are proportional to the areas of the corresponding squares.

year[1,4]. This measure has a similar *rms* variation to alphabetization (Fig. 1b). Publication frequency would be more useful if all papers were cited equally but, as noted above, this is not the case. The best that can be said of publication frequency is that it measures industry rather than ability.

Impact factors are widely used to introduce a citation measure into calculations of publication frequency. The impact factor for each journal, as defined by Thomson Scientific/ISI[5], is the average number of citations acquired during the past two years for papers published over the same period. But weighting each paper by the journal's current impact factor is unlikely to improve the situation, especially when estimating scientific quality across an author's entire career. The impact factor for reputable journals is determined by a small fraction of highly cited papers, so the citation rate for individual papers is largely uncorrelated to the impact factor of the journal in which it was published[6]. The widespread use of publication frequency — with or without an impact factor — is disturbing and requires further study.

Word of caution

Hirsch's *h*-index attempts to strike a balance between productivity and quality and to avoid the heavy weight that power-law distributions place on a relatively small number of highly cited papers. Hirsch's measure is obtained by ranking papers in order of decreasing citations with paper *i* having $C(i)$ citations and solving the equation $h = C(h)$. This is the simplest version of $h = AC(h)^K$. Hirsch's choice of $A = K = 1$ is unsupported by any data. Nevertheless, Fig. 1 indicates that this measure does better than publication frequency, because the *h*-index depends on the entire citation record.

Hirsch's measure overestimates the initial *n*-assignments by some 8 percentile points as indicated by higher densities above the diagonal (Fig. 1c). Moreover, the *rms* uncertainty in the assignment of *h* is ±16 percentile points, which is only a factor of two better than alphabetization. Although capturing certain aspects of quality, Hirsch's index cannot make decile assignments with confidence.

Compared with the *h*-index, the mean number of citations per paper is a superior indicator of scientific quality, in terms of both accuracy and precision. The average assignment of each *n*-bin is in error by 1.8 percentile points with an associated *rms* uncertainty of ±9. Similar calculations based on authors' median citation give an accuracy of 1.5 and an uncertainty of only ±7 percentile points, suggesting that the median copes better with long-tailed distributions.

Simple scaling arguments[4] show that the *rms* uncertainty for any measure decreases rapidly (exponentially) as the total number of papers increases. Thus, for example, no more than 50 papers are required to assign a typical author to deciles 2–3 or 8–9 with 90% confidence when using the mean citation rate as a measure. Fewer papers suffice for deciles 1 and 10. Any attempt to assess the quality of authors using substantially fewer publications must be treated with caution.

> "Institutions have a misguided sense of the fairness of decisions reached by algorithm; unable to measure what they want to maximize (quality), they will maximize what they can measure."

Data access

The methods used here are not specific to high-energy physics. Given suitably homogeneous data sets, they can be applied to any scientific field and permit a meaningful (probabilistic) comparison of scientists working in different fields by assuming the equality of scientists in the same percentile of their respective peer groups. Similarly, probabilities can be combined to make meaningful quality assignments to authors with publications in several disjointed subfields.

There are strong indications that an author's initial publications are drawn on the same probability distribution as their remaining papers[7]. Therefore, with sufficient numbers of publications to draw meaningful conclusions (50 or more) the mean or median citation counts can be a useful factor in the academic appointment process.

Unfortunately, the potential benefits of careful citation analyses are overshadowed by their harmful misuse. Institutions have a misguided sense of the fairness of decisions reached by algorithm, and unable to measure what they want to maximize (quality), institutions will maximize what they can measure. Decisions will continue to be made using measures of quality that either ignore citation data entirely (such as frequency of publication) or rely on data sets of insufficient quality.

Access to the full citation distribution for an entire subfield is essential to our analysis. Existing databases such as the ISI can therefore actively help to improve the situation by compiling field-specific homogeneous data sets similar to what we have generated for SPIRES. This would allow institutions and scientists alike to evaluate the quality of any citation record using all available information. For their part, scientists should insist that their institutions disclose their uses of citation data, making both data and the methods used for data analysis available for scrutiny. In the meantime, we shall have to continue to do things the old-fashioned way and actually read the papers. ■

Sune Lehmann is at the Department of Informatics and Mathematical Modeling, Technical University of Denmark, DK-2800, Lyngby, Denmark.
Andrew D. Jackson and Benny E. Lautrup are at The Niels Bohr Institute, Blegdamsvej 17, DK-2100, Copenhagen, Denmark.

1. Hirsch, J. E. *Proc. Natl Acad. Sci. USA* **102,** 16569 (2005).
2. van Raan, A. F. J. *Scientometrics* **67,** 491 (2006).
3. Lehmann, S., Lautrup, B. E. & Jackson, A. D. *Phys. Rev. E* **68,** 026113 (2003).
4. van Raan, A. F. J. *J. Am. Soc. Inf. Sci.* **57,** 408 (2005).
5. Thomson Scientific/ISI http://www.isinet.com/
6. Seglen, P. O. *J. Am. Soc. Inf. Sci.* **45,** 1 (1994).
7. Lehmann, S. Thesis, The Niels Bohr Institute (2003).

Supplementary information accompanies this Commentary on *Nature*'s website.