



Co-citation Percentile Rank and JYUcite: a new network-standardized output-level citation influence metric and its implementation using Dimensions API

Janne-Tuomas Seppänen¹ · Hanna Värri¹ · Irene Ylönen¹

Received: 2 November 2021 / Accepted: 21 April 2022 / Published online: 21 May 2022
© The Author(s) 2022

Abstract

Judging value of scholarly outputs quantitatively remains a difficult but unavoidable challenge. Most of the proposed solutions suffer from three fundamental shortcomings: they involve (i) the concept of journal, in one way or another, (ii) calculating arithmetic averages from extremely skewed distributions, and (iii) binning data by calendar year. Here, we introduce a new metric Co-citation Percentile Rank (CPR), that relates the current citation rate of the target output taken at resolution of days since first citable, to the distribution of current citation rates of outputs in its co-citation set, as its percentile rank in that set. We explore some of its properties with an example dataset of all scholarly outputs from University of Jyväskylä spanning multiple years and disciplines. We also demonstrate how CPR can be efficiently implemented with Dimensions database API, and provide a publicly available web resource JYUcite, allowing anyone to retrieve CPR value for any output that has a DOI and is indexed in the Dimensions database. Finally, we discuss how CPR remedies failures of the Relative Citation Ratio (RCR), and remaining issues in situations where CPR too could potentially lead to biased judgement of value.

Keywords Article-level metrics · Citation analysis · Citation count normalization

Introduction

Judging value of a scholarly output¹ remains a difficult but unavoidable challenge. It is difficult because the ideals of judging scholarly value—thorough careful reading, long deliberation, debating in person or via published exchanges, peer review, replication—defeat their purpose in situations where the need for judgement in the first place arises from

¹ Note on terminology: we use “scholarly” and “output” here as umbrella terms to include all serious academic study and different formats and channels of its communication, some of which might be seen to be excluded by terms “scientific” and “article” which we use here only when referring to earlier work by others using those terms.

✉ Janne-Tuomas Seppänen
janne.t.seppanen@jyu.fi

¹ Open Science Centre, University of Jyväskylä, Jyväskylä, Finland

available time, attention and expertise resources being far exceed by the material. These situations, and hence judgement by less than ideal methods, are unavoidable: they arise in grant review panels and academic hiring committees, for science journalists, for evidence-based governance, for teachers, not to mention in daily attention allocation decisions of individual researchers.

Judging is also difficult because different aspects of value, such as *importance* and *influence*, are multifaceted concepts weighed by subjective considerations. Furthermore, influence and importance, however one measures them, are entangled but not interchangeable measures, further confounding the task. Appearing on a more prestigious platform obviously does make an output more influential, by virtue of being noticed more. Being cited more is both evidence of and path towards more influence (Wang et al., 2013).

Though simple to explain and understand, appearance on a prestigious platform and raw citation count are not meaningful proxies for importance:

Platform prestige is a largely intangible and informally propagated social construct of academic communities. In the case of journals, it is nonetheless influenced by and correlated with the Journal Impact Factor, JIF (Garfield, 1955, 2006), the annually published quantitative metric of citation performance of journals. But regardless of the degree or pretension of formality, platform prestige is always an inherently flawed proxy to judge value of individual outputs (Johnston, 2013; Pulverer, 2013; Seglen, 1997; Stringer et al., 2008). The prestige of a platform arises from high submission volumes coupled with extreme selection exclusivity, and importance of some past exceptional outputs—in other words, platform prestige mostly reflects lack of importance of a mass of material and extremity of importance of a very small number of outputs *other* than the individual output in question. It is therefore not surprising that platform prestige does not correlate positively with quality, reliability or importance of outputs published on it (Brembs, 2018).

Raw citation count of an output—basis of the widely used h-index (Hirsch, 2005; Norris & Oppenheim, 2010)—is first and foremost predicted by the size and citation traditions of the field and obviously the age of the output (Radicchi & Castellano, 2012), not the content of the output. Variation due to field of research is not limited to differences between entirely different disciplines such as Mathematics versus Medicine, but also includes significant variation between narrow sub-fields (Radicchi & Castellano, 2011) and even between study questions falling within the scope of a single topical journal (Ophthof & Leydesdorff, 2010; Radicchi et al., 2008). Annual global publication volumes and citation patterns such as typical length of reference lists and typical age of cited publications can be very different between, say, Theoretical Ecology versus Environmental Ecology, or between studies investigating genetic basis of behaviour in *Orcinus orca* versus in *Caenorhabditis elegans*, both published on some same Behavioural Ecology platform.

These issues have been recognised and debated for a long time. Proposed solutions include radical reassessment of the value of “value” in scholarly work, ranging from fundamental sociological and linguistic changes to abandon “fetishisation of excellence” (Moore et al., 2017), to replacing research grant review panels with random lotteries (Fang & Casadevall, 2016) as was recently done by The Health Research Council of New Zealand (Liu et al., 2020). These approaches merit serious consideration and real-world testing and adoption, but are outside the scope of this article.

Others recognise the usefulness of bibliometrics, or at least their unavoidability, as long as they are better than the obviously maligned things of yore. These solutions tend to be new metrics, and there are hundreds of them. Wildgaard (2014) and Waltmann (2016) provide recent reviews. A non-exhaustive list includes: normalizing citation counts by averages of articles appearing in a single journal or in journal categories assigned by publisher

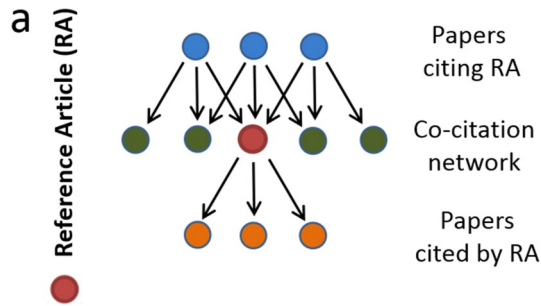


Fig. 1 Adopted from Hutchins et al. (2016a, 2016b) “a. Schematic of a co-citation network. The reference article (RA) (red, middle row) cites previous papers from the literature (orange, bottom row); subsequent papers cite the RA (blue, top row). The co-citation network is the set of papers that appear alongside the article in the subsequent citing papers (green, middle row). The field citation rate is calculated as the mean of the latter articles’ journal citation rates.” Permission: Originally published under Creative Commons CC0 public domain dedication

or database vendor (Bornmann & Leydesdorff, 2013; Opthof & Leydesdorff, 2010; Waltman, van Eck, et al., 2011; Waltman, Yan, et al., 2011); normalizing citation counts by average number of cited references in citing articles (Moed, 2010; Zitt & Small, 2008); counting number of articles in the “top 10%” of most cited articles but normalizing by researcher’s (academic) age (Bornmann & Marx, 2013); citation network centrality measures such as the journal Eigenfactor and the derivative Article Influence which is journal Eigenfactor divided by the number of articles on that platform (as Bergström 2007 admits, this is essentially judging “importance of an article by the company it keeps” as thus no better than the misuse of JIF to judge individual articles); dealing with collaborative authorship by fractioning the credit among individual researchers (e.g. Schreiber, 2009), their institutions (Perianes-Rodríguez & Ruiz-Castillo, 2015) or countries (Aksnes et al., 2012).

Most of the proposed solutions suffer from three fundamental shortcomings:

- (1) They involve the concept of journal, in one way or another.
- (2) They involve calculating arithmetic averages from extremely skewed distributions.
- (3) They involve binning data by calendar year.

All of these probably first emerged historically from only aggregate and annually compiled data being easily available in the pre-internet era, and later became staple traditions seen to hold some value in themselves. But today they just bring unnecessary inaccuracy, bias, restricted samples, delays and lack of generality to measures seeking to judge value of individual scholarly outputs and researchers.

Methods

Co-citation Percentile Rank (CPR)

Here, we describe Co-citation Percentile Rank (CPR), a metric that relates the current citation rate of the target output to the distribution of current citation rates of outputs in its

co-citation set (Fig. 1), as its percentile rank in that set. Here, *citation rate* is defined by number of citations received by the target output to date, divided by the number of days the output has been available for citation. The *co-citation set* is defined identically to "co-citation network" in Hutchins et al., (2016a, 2016b), see Fig. 1. and Discussion.

To retrieve the co-citation set and its bibliometric metadata, we use Dimensions Search Language (DSL) on their API:

- (1) get Dimensions **ID** and other data about the target, by **DOI** (or multiple targets: up to 400 can be queried in a single call, then looped through the queries below):

```
search publications where doi in [DOI] return publications [basics+extras+date_inserted]
```
- (2) get data about outputs where the target appears in their list of references, i.e. the citing outputs, and in particular the reference_ids of outputs they cite, i.e. the co-citation set of the target:

```
search publications where reference_ids="ID" return publications [title+id+reference_ids]
```
- (3) Collect the reference_ids from the result array, remove duplicates, concatenate the unique ids into comma-separated strings chunked to max 400 ids per string.
- (4) get metadata about the co-citation set, by **reference_ids** string (looping through if multiple chunks):

```
search publications where id in ["reference_ids "] return publications [title+doi+times_cited+date+date_inserted+id]
```

Once the citation counts and days since the output became citable have been assembled, these are converted to citation rate = $\frac{\text{citation count}}{\text{days since first citable}}$. When used in graphs and tables, that per-day citation rate is rescaled to "*citations/365 days since becoming citable*". We do this to aid interpretation and comparisons, but we emphasize that 365 days is not a calendar year, and that the citation rate is defined also for outputs that are less than one year old.

With data as extremely skewed as citation counts, it should be obvious that arithmetic averages are not appropriate denominators. Leydesdorff et al. (2011) make this case explicitly for bibliometrics, and argue that newly proposed metrics should use percentile ranks instead of ratios to arithmetic averages of reference sets.

The co-citation set is then ordered by citation rate, and then

$$\text{CPR} = \text{percentage of citation rates in the co-citation set that are less than or equal to the citation rate of the target output}$$

CPR = percentage of citation rates in the co-citation set that are less than or equal to the citation rate of the target output

CPR thus ranges from 0 (=all outputs in the co-citation set are cited more frequently than the target) to 100 (=none of the outputs in the co-citation set are cited more frequently than the target). CPR is undefined for targets that have not been cited, or for which the algorithm is unable to find co-citation set metadata. The quartiles and average of the citation rates, and the size of the co-citation set are also calculated and saved for illustrative purpose.

Note on the day count used in the algorithm: when an output first becomes citable, typically as an “early online” article, Dimensions lists that date in the metadata. However, for platforms that later assign outputs to issues (some even distribute those arbitrary collections of outputs as anachronistic physical representations created on cellulose-based perishable material), that date eventually gets replaced in Dimensions metadata by a later so-called “publication date”. The original, true date when the work first became citable, is not currently retained in Dimensions in these cases. For some platforms the resulting error may be just few tens of days, but in others the error may be many hundreds of days. In most cases the true date would be recoverable from metadata in CrossRef or the landing page of the output DOI, but this would impose prohibitive time performance cost on the algorithm. We hope this unfortunate and unnecessary source of error gets remedied in the future. For now, we resort to retrieving also the **date_inserted** -value (i.e. the date when the Dimensions record was first created) alongside the **date** (i.e. publication date) value from Dimensions API, and choose the earlier of these two dates, acknowledging some outputs may still enter the algorithm with unknown error in their citation rate, and that this error is more common and likely larger for some disciplines, such as Economics or Humanities. When the date value included only year and month, we assumed the day to be the first of that month.

The source code of the algorithm is publicly available available at https://gitlab.jyu.fi/jyucite/published_cpr and published in Seppänen (2020).

Example dataset

Metadata for total of 41 713 outputs from JYU current research system published between 2007 and 2019 (all kinds, including non-peer-reviewed outputs) were assembled and evaluated, out of which 13 337 (i) were discoverable in Dimensions by either DOI or title and (ii) had at least one citer.

That dataset of metadata for 13 337 outputs is published alongside this article. It gives DOI, Dimensions id, title, type of output (as defined by Ministry of Education and Culture in Finland (2019), and assigned by information specialists at research organizations), university home department(s) at JYU, publication date, number of days available and number of citations accrued up to date of analysis, calculated citation rate per 365 days, co-citation set size, number of co-citations for which times_cited metadata was present, co-citation median citation rate, quartiles and average, the CPR metric, and the solve time (seconds it took to retrieve the co-citation set and calculate the metrics) as well as UNIX timestamp of the time when the calculation was done.

Out of those, 13 170 had at least 10 co-citations for which metadata could be found, and these were retained for beta regression analysis.

Whether 10, or some other minimum number, of comparison articles is an adequate limit to meaningful comparisons cannot at present be quantitatively defined. Further research on temporal dynamics of CPR, and its correlations to other metrics of work’s scientific value, are required to begin to understand how little is too little. For example, if a work was published just one week ago and has been cited once, it ostensibly has very high present citation rate ($1/7 \text{ days} = 51/365 \text{ days}$), and a small set of co-citations to compare it with, so its CPR is not very meaningful. But alternatively, if a work was published exactly 10 years ago, and has been cited only once, it has a very small citation rate ($1/3650 \text{ days} = 0.1/365$), and though it too has a small set of co-citations to compare it to, many of them are very likely to be cited more frequently and CPR ends up being very low,

which clearly is meaningful in saying that the work has low citation impact. In addition to the size of the comparison set, future analysis of CPR's temporal dynamics may well help in defining a time-window in a work's early existence, where CPR is too volatile to be meaningful—in other words, when a publication is very new, we likely cannot confidently estimate its citation impact. Before such analyses are available, we consider it is prudent to keep broad validity thresholds to co-citation sets, only excluding extremely small sets.

The example dataset is publicly available at https://gitlab.jyu.fi/jyucite/published_cpr and published in Seppänen (2020).

Statistical analysis

Descriptive analyses on the effect of number of citers on CPR solving time and size of the co-citation set were done with quantile regression (Koenker, 2020). Quantile regression model fits are estimated using the R1 statistic (Koenker & Machado, 1999; Long, 2020).

Because the distribution of CPR data is by definition bound to unit interval (0,100), and, as is typical for any citation metric, also heteroscedastic and asymmetric, we model its response to predictors using beta regression (Cribari-Neto & Zeileis, 2010). Analysis incorporating precision parameter to account for heteroscedasticity failed to converge with raw data for any link function, so a preliminary analysis was run without precision parameter to identify outliers. Model with log–log link function converged so that is used in all subsequent analyses. Twenty extremely highly cited outputs showed outlier impact (Cook's distance > 1.0) in preliminary regression analysis and were removed from subsequent analysis. Beta regression including a precision parameter was then run. Cases where CPR was zero (405 cases) continued to have disproportionate generalized leverage on the fit, and were also removed. The final sample where both extreme highs and lows had thus been removed contained 12,740 observations.

After fitting the overall beta regression of CPR to citation rate, we next explore the differences in that relationship between academic disciplines. Each output in the database is affiliated with one or more academic departments at University of Jyväskylä (JYU). For purposes of this analysis, we simplify the data by merging data from units falling under same discipline (e.g. Institute for Education Research is merged with Department of Education). For outputs still having multiple affiliations after the mergers, the record is replicated so that an output occurs once for each affiliated department. Departments having fewer than 100 outputs were excluded from the final analysis. The derived expanded dataset contains 13,871 observations from 16 different academic departments (Table 1).

We then model CPR as a function of citation rate and academic department, including interaction terms and a precision parameter accounting for heteroscedasticity along the citation rate range.

The R code to replicate the analyses and figures presented here is publicly available at https://gitlab.jyu.fi/jyucite/published_cpr and published in Seppänen (2020).

Results

With our computing setup (a single RHEL7.7 Server, 1 × 2.60 GHz CPU, 2 GB RAM, PHP 5.4.16, in university's fiber-optic network), solving CPR for an output was within reasonable time performance and solve time was linearly scaled with number of citers, though variance was considerable (Fig. 2).

Table 1 Departments and output counts included in the analysis

| Department | No. outputs |
|--------------------------------|-------------|
| Physics | 2469 |
| Sport and Health | 2117 |
| Bio-environmental | 1912 |
| Information Technology | 1475 |
| Chemistry | 1415 |
| Psychology | 1047 |
| Education | 586 |
| Mathematics and Statistics | 585 |
| Business and Economics | 499 |
| Social Sciences and Philosophy | 442 |
| Language and Communication | 349 |
| Teacher Education | 312 |
| Music, Arts and Culture | 273 |
| Agora Center* | 185 |
| History and Ethnology | 104 |
| Applied Linguistics | 101 |

*Multidisciplinary institute on humanities, social sciences and technology, active 2012–2017

Co-citation set size expanded rapidly as number of citers grew: each new citer typically brought ca. 40 new entries to output’s co-citation set, and occasional review, book or reference work citer could bring hundreds or even thousands of entries (Fig. 3).

Overall beta regression of CPR as a function of citation rate illustrates that though CPR increases with increasing citation rate, the relationship has considerable variation, as should be expected if CPR captures sub-field specific citation influence. Some outputs achieve CPR around 50 (i.e. are cited at least as frequently as half of their co-citations) when they get cited 1–2 times per 365 days, while other outputs need 9–10 citations per 365 days to achieve similar CPR (Fig. 4).

With the expanded derived data, likelihood ratio test shows that including departments in the regression model gives significantly better fit ($\chi^2=1331.3$, $df=30$, $p<0.001$), despite being considerably more complex model (34 degrees of freedom vs 4 when modelling as function of citation rate only: $\Delta AIC=1271.3$). In other words, some difference between academic departments explains significant part of the variability in response of CPR to citation rate. Different global publication volumes and citation behaviours in different disciplines are the obvious likely explanation.

Pairwise post-hoc Tukey tests (Supplementary Table S1) show that majority of the department contrasts in coefficient for citation rate on CPR (69 out of 120) are statistically significant (Bonferroni-adjusted p-value < 0.05). Differences are most pronounced between Mathematics & Statistics vs other disciplines: the slope is consistently steeper in Mathematics & Statistics, i.e. CPR responds more rapidly to increase in citation rate. Going from rate of two citations to five citations per 365 days is predicted to move an output in Mathematics & Statistics from bottom third to top third in citation rate in its co-citation set. In contrast at the other extreme, in Physics same change in citation rate means an output moves from bottom 25% only to a rank still below median in its co-citation set.

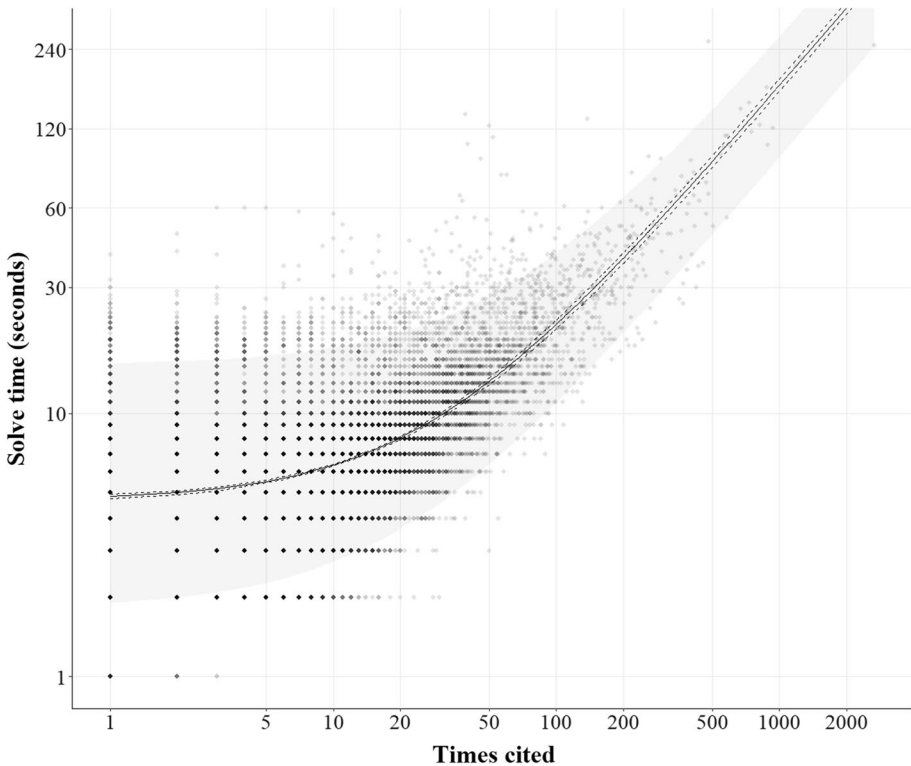


Fig. 2 Log–log plot of the CPR solve time as a function of times the target output is cited. Dot opacity is set at 0.1 in R plotting function so that overlap darkness illustrates density of observations. Quantile regression for median (line) and 90% prediction interval (shaded area) and 99% confidence interval (dashed lines). Solve time $\sim 0.17 * \text{times cited} + 4.66$ ($N=13\ 337$, model fit: $R^1(0.5)=0.29$). The variation in solve time results from variable co-citation set size given times cited (see Fig. 3), and occasional automatic delays in the algorithm as it adheres to rate limit of max 30 requests per minute as mandated in Dimensions’s terms of use

Similarly, post-hoc tests for contrasts in intercept term (Supplementary Table S2) show that majority of the department contrasts (65 out of 120) are statistically significant. The Agora Center (a former multidisciplinary research unit bringing together humanities, social sciences and information technology) differs from all other department by having significantly lower intercept, i.e. CPR of Agora Center outputs was generally lower, given a citation rate. This could partially be a result of the multidisciplinary nature the unit had (see “CPR reflects expected differences between disciplines” below). History & Ethnology on the other hand consistently has significantly higher intercept term than other departments, i.e. CPR for a History & Ethnology output tends to be higher, given a certain citation rate.

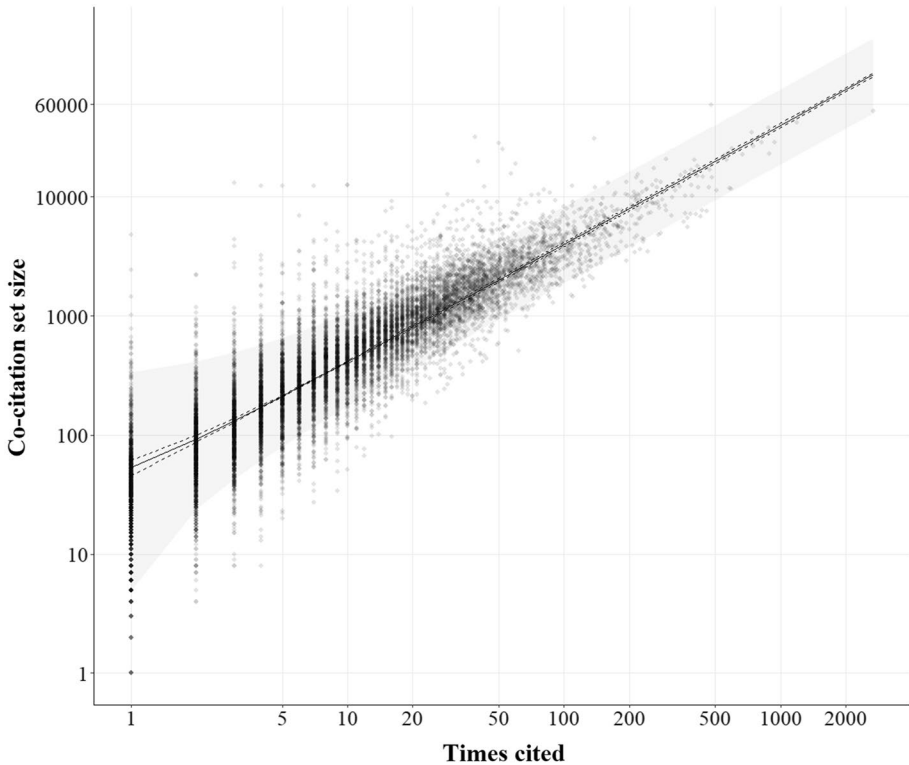


Fig. 3 Log–log plot of the co-citation set size as function of times a peer-reviewed output has been cited. Dot opacity is set at 0.1 in R plotting function so that overlap darkness illustrates density of observations. Quantile regression for median (line) and 90% prediction interval (shaded area) and 99% confidence interval (dashed lines). Co-citation set size $\sim 39.73 * \text{times cited} + 13.82$ ($N = 13,337$, model fit: $R^1(0.5) = 0.54$). The extreme high outliers result from an output being cited by reviews and books, or large reference works, which alone can bring hundreds or thousands of entries, respectively, to output’s co-citation set

Discussion

CPR reflects expected differences between disciplines

The contrast between academic disciplines found are consistent with conventional wisdom and highlight the utility of CPR as a field-normalized citation rate metric. In Mathematics & Statistics, publication volumes are relatively low both individually and collectively, and works often cite just a few foundational references. In Physics, the three particle accelerators and nanoscale materials research at JYU result in outputs that are in very large, fast-moving fields where massively collaborative works get cited quickly. In History & Ethnology, the discipline has smaller and slower publication volumes as many outputs are monographs, and perhaps citations behaviours are also more siloed to fine-grained sub-fields and by language. It should be noted though, that the sample size for History and Ethnology is small here, and furthermore that sparse sample shows more variation than other disciplines (Fig. 5), so inference must be cautious. Future analyses with representative random samples from different disciplines (not

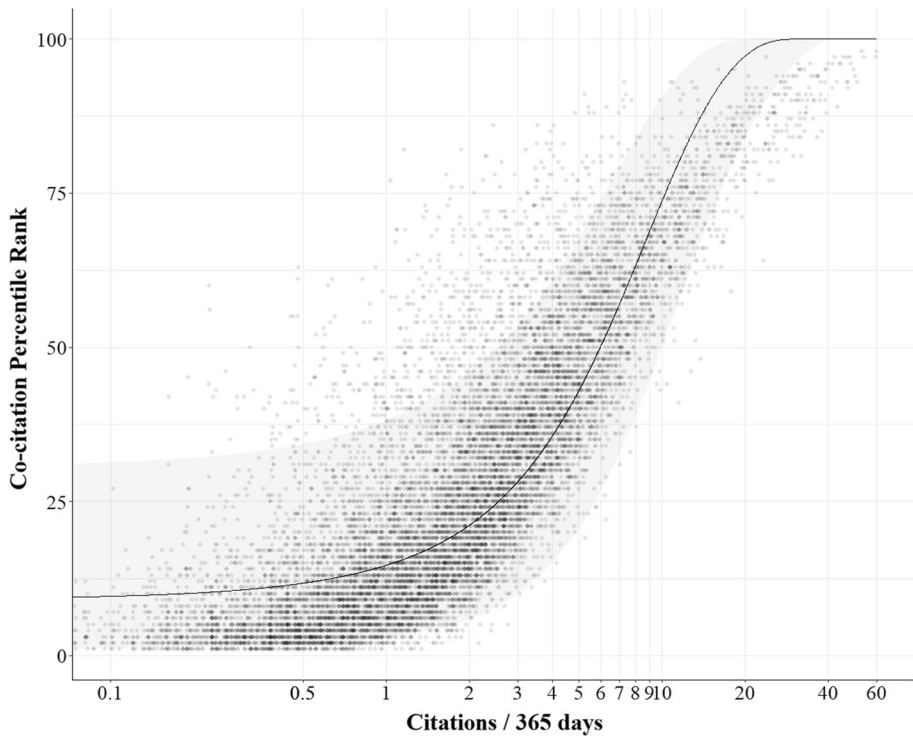


Fig. 4 Co-citation Percentile Rank of outputs as a function of citation rate (citations/365 days since becoming citable). Dot opacity is set at 0.1 in R plotting function so that overlap darkness illustrates density of observations. Beta regression predicted median (line) and 90% prediction interval (shaded area)

from just single university, or delineated based on university departments) are needed to determine how well CPR distributions are comparable between them—if the metric performs well, the distributions should be statistically similar between disciplines (pers.comm. Ludo Waltmann).

Notably, considerable variation of CPR given a certain citation rate remains even after segregating the data by academic department. This suggests CPR further normalizes sub-field or even research question -specific differences in expected citation rates. Future analyses of CPR with more fine-grained and accurate delineation of “discipline” (e.g. expert-curated lists of outputs judged to be narrowly on the same topic) should seek to confirm this.

Implementation using Dimensions API

The efficiency of the Dimensions API is remarkable for this task: for an oeuvre of M target outputs (say, the publication list of a person), only $1 + M$ queries are needed to obtain the IDs of entire co-citation sets of all target outputs. The N IDs in a single co-citation set can often number thousands. But metadata for these can be retrieved in $N/400$ batch queries. Because the algorithm does not have to make separate Dimensions API queries to retrieve either the co-citation set IDs or the metadata for all of the entries, the CPR solve

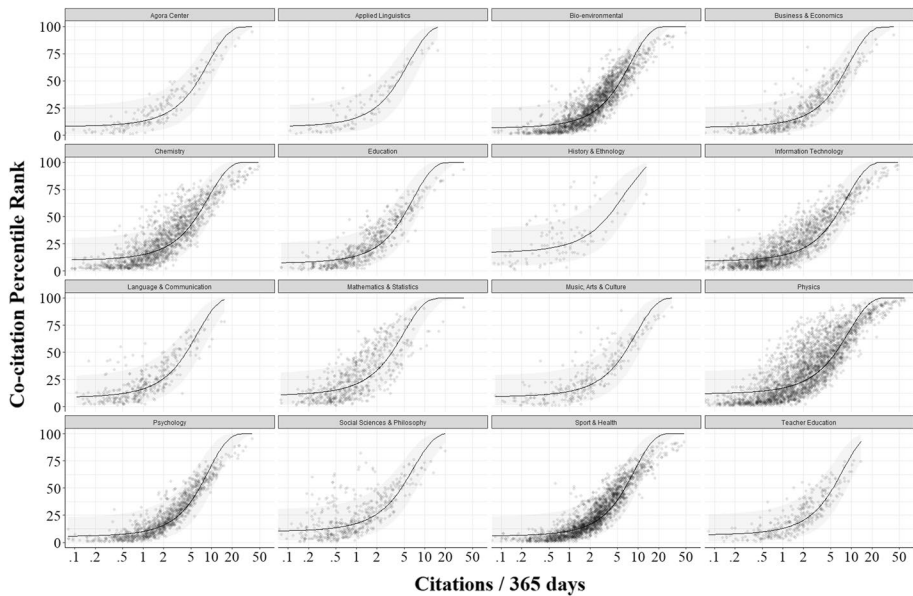


Fig. 5 Co-citation Percentile Rank of outputs as a function of citation rate (citations/365 days since becoming citable) for 16 different academic departments at JYU. Dot opacity is set at 0.1 in R plotting function so that overlap darkness illustrates density of observations. Beta regression predicted mean (line) and 90% prediction interval (shaded area). Note that x-axis is logarithmic

time is reasonably linearly scaled to the number of citers citing the target output (using our computer and network resources, ca. 0.17 s/citer, Fig. 2). Performance could be in principle multiplied by parallelization, but only with consent of the database provider. Here, we operated under the standard consumer terms of service by Dimensions, which set a maximum request rate of 30 / minute.

Interestingly, whereas Hutchins et al., (2016a, 2016b) found that each citer added on average ca. 18 new outputs to the co-citation set, we find that each new citer adds ca. 40 (Fig. 3). This may result from the narrower focus as Hutchins et al. limited their analysis to articles in those journals in which NIH R01-funded researchers published between 2002 and 2012, i.e. largely biomedical research. In contrast, the dataset we use here to illustrate CPR covers all research outputs 2007–2019 from one university, spanning 18 different departments. Second, and likely a more consequential difference is that Hutchins et al. used Web of Science, whereas we use Dimensions citation database. Dimensions indexes citations much more inclusively, particularly in dissertations, textbooks, monographs and large reference works such as encyclopaedias. Such book-length sources often include much longer lists of references—and hence large contributions to co-citation sets—than typical articles.

Faster accrual of the co-citation set using Dimensions also means our implementation of CPR is not very vulnerable to finite number effects. For median, typical target output, mere three citers bring its co-citation set size above 100, where percentile can be confidently stated as integer value without having to interpolate between observed values.

CPR and RCR

Definition of the reference set, and inspiration for developing CPR, is the Relative Citation Ratio (RCR, Hutchins et al., 2016a, 2016b). The foundational idea of RCR—comparing the citation rate of target output to those of ‘peer’ outputs which are cited alongside the target, the co-citation set—makes intuitive sense and is a truly journal-independent research field -normalization for quantitative comparison of academic impact.

All three sets of outputs in an output’s citation network (cited, citing and co-citations, see Fig. 1) can be intuitively seen to “define empirical field boundaries that are simultaneously more flexible and more precise” (Hutchins et al., 2016a, 2016b). But the co-citation set is clearly the superior choice. While the cited set is defined just by the authors, the other two sets grow dynamically over time and can be seen to reflect current aggregate expert opinion of what constitutes the field of the target output. The citing set might be considered to be the most direct expression of that aggregate expert opinion, but as a reference set for comparing citation performance, the set suffers from two significant problems: by definition it only includes newer outputs than the target, and it usually grows slowly. Hence, typically for years there would not be a meaningful comparison set. Then a set would begin to exist but comparison would be meaningless as the citing outputs would not yet have had time to get cited themselves. And if and once, eventually, those issues begin to fade then the target output gets compared to a set that contains more currently relevant and derived research, only, several years after the target itself was at its most influential. In slow-moving fields, this could easily take a decade or longer.

In contrast, the co-citation set grows by multiple outputs every time the target is cited. Even with partial overlap, Hutchins et al., (2016a, 2016b) found that each citer added on average ca. 18 new outputs to the co-citation set in their example in biomedical sciences. Just as importantly, the temporal coverage of the co-citation set encompasses all research from oldest seminal outputs to the newest research that can reasonably appear in reference lists today.

In addition, the text similarity (Lewis et al., 2006) was greater among co-cited outputs than among articles appearing on same platform.

However, RCR is problematic in several ways, and CPR has been developed to remedy those failures.

Most glaring discrepancy to the promise of the idea of truly output-level metric, was that Hutchins et al. resorted to using Journal Impact Factors (JIF) in the normalization. They define Field Citation Rate (FCR) as the average JIF of journals where the co-cited articles appeared. The only justification for this decision was a vague claim that average of citation rates of articles would be “vulnerable to finite number effects”. An unmentioned but possible explanation was to make the introduction presented in their scholarly article (based on Web of Science citation data) comparable with the online implementation (based on PubMed data). PubMed indexes an extremely narrow scope of the published literature and even narrower slice of the citation network, so most articles in the co-citation set would show zero or very few citations. Such zero-inflated and relatively small data indeed would be susceptible to instability of average, but that is more a property of the data source, rather than the co-citation set itself.

Regardless of justification, JIF itself is (sort of) an average, from an extremely skewed distribution. Furthermore, calculation of JIF is a negotiated and extensively gamed process in the publishing industry, and most importantly, it is trivially obvious that neither the true

importance nor the citation rates of individual outputs are meaningfully correlated with the JIF of the journal where the output appears.

The target Article Citation Rate (ACR) was defined as the accrued citations divided by the number of calendar years since publishing, excluding the publication year. This introduces another serious yet unnecessary problem: some articles had had 364 days more to accrue citations than others divided by the same number of calendar years. Especially for recent articles this is a significant source of error. Modern citation databases are updated continuously and outputs and their citations appear within days, sometimes before, of the date of publication, and the date is listed in the metadata. There is no reason to obscure that temporal resolution of citation data by aggregating counts by calendar years, when citation rates can be easily expressed as citations / days since citable.

Then, instead of straightforward definition of $RCR = ACR/FCR$, Hutchins et al. (2016a, 2016b) present a convoluted regression normalization procedure requiring very large datasets of other “benchmarking” articles. This decision may reflect an unmentioned recognition that their FCR, as it is an average from an extremely skewed distribution of JIF values which themselves are (sort of) averages from extremely skewed distribution of two-year article citation counts, inevitably is a much larger number than the expected citation rate of a typical article. As Janssens et al (2017) notes, the FCR-to-RCR regressions had intercepts ranging from 0.12 to 1.95 and slopes from 0.50 to 0.78, which is entirely expected given the way FCR is calculated. The regression normalization procedure yielded Expected Citation Rate (ECR, detailed in Hutchins et al., 2016b) for the target article, and then $RCR = ACR/ECR$ so that for articles in the dataset with the same FCR, the article with average citation rate had $RCR = 1.0$. Again, given the skewed nature of citation accrual, this inevitably means that most articles will have an RCR (much) below 1.0 despite many of them being cited more frequently than majority of articles in their co-citation sets.

Final shortcoming was the subsequent implementation of RCR in iCite, building the algorithm on top of PubMed citation data. That data is woefully sparse for most fields of science. Most outputs simply are not indexed by PubMed at all. And even when they are indexed, PubMed fails to capture vast majority of citations to them.

JYUcite

We have built a publicly available web resource allowing anyone to retrieve CPR value for any output that has a DOI and is indexed in the Dimensions database. The resource is available at <https://oscsolutions.cc.jyu.fi/jyucite>.

Additionally, JYUcite graphically and numerically displays the observed citation rate, and size N and quartiles of citation rates in the co-citation set (Fig. 6). The citation rates are expressed rescaled as citations/365 days (note: not a calendar year, but daily rate scaled to one year for convenience—see “Methods”).

At the time of publication of this article, JYUcite limits number of DOIs per request to 10 to ensure performance as we develop the resource further. JYUcite also limits the number of newly calculated CPR values returned (i.e. those requiring JYUcite server to make calls to Dimensions API) to 100 per IP-address per day, as contractually agreed with Dimensions database. If a CPR value for a DOI has already been calculated previously and is saved in JYUcite’s own database and is younger than 100 days, it is returned immediately without new calculation and does not count in the daily rate limit.

University of Jyväskylä | Open Science Centre | JYUcite

JYUcite: research metrics by University of Jyväskylä Open Science Centre

Get CPR | What is CPR?

DOIs (separate by comma, max 10):

GET

| # | Title | CPR | Co-Citation Quartiles, citations / 365 days (hover mouse or tap to see details) |
|----|--|-----|--|
| 1. | The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits link | 83 | |
| 2. | Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods link | 73 | |
| 3. | Prestigious Science Journals Struggle to Reach Even Average Reliability link | 66 | |
| 4. | Grant funding: Playing the odds link | 12 | |

JYU/Open Science Centre

Fig. 6 Screenshot from JYUcite web service, showing results view when requesting CPR for some articles listed in References in this article

Remaining issues

Interdisciplinary influence

An output may suffer lowering CPR if it begins to get cited in another field where outputs typically get much more citations (e.g. due to field size and typical lengths of reference lists). It would begin to gain (proportionally large numbers if the other field typically has longer length of reference lists) relatively frequently cited co-citations into its co-citation set, repressing its CPR. Thus interdisciplinary influence, which typically is seen as a merit, may actually erode the CPR value of an output (see also Waltman, 2015). On the other hand, it could also be argued that once an output begins to become relevant in another discipline, it should start to get compared to outputs in that field, otherwise it would appear unduly influential compared to those.

A possible remedy could be to pay attention to relative $\frac{\text{co-citation set size}}{\text{citation count}}$ ratio: if an output in e.g. individual researcher's publication list stands out from others with a noticeably

larger ratio, it might indicate interdisciplinary influence, and that output with its citing set should then be flagged for individual judgement by a human, for potential extra merit.

The reverse (an output from a field with high typical citation rates begins to get cited in a field with typically much lower citation rates, thereby gaining low-cited co-citations, inflating its CPR) is likely to have less significant impact as the gain of low-cited co-citations is slow compared to gain of co-citations in its original field, and the proportion of those co-citations remains marginal in the set.

Co-citations in reference works

Being cited in a large reference work, such as an encyclopaedia, can bring thousands of entries into the co-citation set. On one hand this is an advantage, e.g. when “Encyclopedia of Library and Information Sciences” creates relevant co-citation set of 4608 entries for large number of outputs, which otherwise would wait many years to accrue similar sized comparison set.

However, “Encyclopedia of Library and Information Sciences” also cites articles from completely unrelated fields, e.g. a conservation ecology article from 1968 titled “Chlorinated Hydrocarbons and Eggshell Changes in Raptorial and Fish-Eating Birds”. While this is likely to be an insignificant source of bias for the CPR of outputs in Library and Information Sciences (relatively few of the thousands of co-citations accrued are from outside fields), it could be significant for the CPR of the out-field article cited. It seems likely though, that outputs that end up cited in reference works outside their own field are cited extensively in their old fields before that, and tend to be aged. The example article above has, at time of writing this, been cited 301 times and has a co-citation set of 21.616 outputs, so contribution of the encyclopaedia is at most 20% of the co-citation set.

A possible remedy could be to exclude co-citations coming from citers that have extreme outlier count of references. A typical scholarly article has ca 30 references, so a threshold of 300 would likely retain all original research and most review articles, while excluding large reference works. However, the trade-off of such filter would be to lose relevant co-citations for many little-cited articles, which could be an issue in disciplines and fields where co-citation sets accrue slowly.

Hyper-cited co-citations

Some scholarly outputs, particularly in natural sciences, medicine and engineering, may receive many thousands of citations every year, and consequently are likely to appear in the co-citation set of most outputs. These hyper-cited outputs tend to be statistical analysis textbooks (e.g. “Statistical Power Analysis for the Behavioral Sciences (2015)”, cited more than 37.000 times in total, or over 5000 times per every 365 days) or companion articles to statistical packages in the R -software (e.g. “Fitting Linear Mixed-Effects Models Using lme4 (2015)”, cited more than 21,000 times in total, or almost 4000 times per every 356 days) or descriptions of standard hardware or software infrastructure (e.g. “A short history of SHELX (2008)”, cited more than 68.000 times in total, or over 5000 times per every 365 days).

Is inclusion of such “boilerplate” general methodology outputs in the comparison set a source of bias? They are always far more frequently cited than a typical target output, and thus outrank them, lowering the target output’s CPR. On the other hand, they often

objectively **do** have more significant influence on research than any field-specific research output. Also, they cannot be automatically excluded using either blacklists or citation rate thresholds, as the target output may well be directly in the same field as, or a co-citation have citation rate rivalling, e.g. a statistical methodology “boilerplate” reference.

But even where hyper-cited co-citations are seen as a source of bias for CPR, the magnitude of such bias decreases rapidly as target output gains more citers and thereby median of 40 new co-citations per new citer (Fig. 3). Single or few hyper-cited co-citations in a large set do not have a large effect on the percentile rank, while they would easily distort a metric relating the target to averages in the set.

Delayed recognition and instant fading

The value of a scholarly output is not necessarily present or recognized immediately upon publication. Some other, later discovery may suddenly reveal an important aspect of an earlier output, or its relevance in another field of research may be get re-discovered serendipitously, awakening the output to suddenly increased influence, many years after first becoming citable. Ke et al. (2015) present analysis of temporal patterns for such “sleeping beauties”, and find that while there are cases of very pronounced delayed recognition and they are often associated with delayed accrual of interdisciplinary citations, such articles are not exceptional rare outliers in their own category, but simply extreme cases in heterogenous but continuous distribution.

Conversely, the influence of a scholarly output may be immediate and intense but then quickly fades (Ye & Bornmann, 2018). This can happen in very rapidly advancing fields where an important discovery can spur quick succession of derivative further discoveries or improvements which overshadow their ancestors, or for big or controversial claims that generate large initial interest which however proves the claim false and the output is thereafter ignored and forgotten.

Consider three outputs with similar citation rates, where one is an old instant fade, another equally old but recently awakened sleeping beauty, and third a new output receiving few but immediate citations. Their CPRs would be similar. Whether this would be an unfair similarity is not necessarily clear: should an awakened sleeping beauty be considered more valuable than a new discovery, or should we acknowledge that CPR reflects whatever it was that kept the sleeping beauty from gaining influence sooner? Notably, the direction of change in CPR would be different for these three examples.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11192-022-04393-8>.

Acknowledgements We thank colleagues at Open Science Centre at JYU for enabling assembly of the dataset used in this article, and Jyrki Laitinen and colleagues at Digital Services at JYU for expertise and tireless help in setting up servers and keeping them running. Stacy Konkiel and other experts and customer support staff at Digital Science have provided valuable and timely guidance. Lutz Bornmann and Ludo Waltman gave valuable comments on the preprint version of this article (Seppänen et al., 2020), available at <https://www.biorxiv.org/content/10.1101/2020.09.23.310052v1.full>

Author contributions Conceptualization: J-TS, HV, IY; Methodology: J-TS; Data curation: J-TS; Software: J-TS; Formal analysis: J-TS; Investigation: J-TS; Writing—original draft preparation: J-TS; Writing—review and editing: J-TS, HV, IY; Visualization: J-TS; Funding acquisition: not applicable; Supervision: J-TS, IY; Project administration: IY; Resources: IY; Validation: HV, IY.

Funding Open Access funding provided by University of Jyväskylä (JYU). No specific funding was received for conducting this study. Subscription to Dimensions was paid from general institutional budget of University of Jyväskylä.

Data and code availability The source code of the algorithm, the example data, and the R code to reproduce the statistical analysis and figures here is publicly available at https://gitlab.jyu.fi/jyucite/publicshed_cpr (<https://doi.org/10.17011/jyx/dataset/71858>) and published in Seppänen (2020).

Declarations

Conflict of interests Access to Dimensions API was obtained by otherwise standard paid subscription contract between its owner Digital Science Ltd and University of Jyväskylä, but special provisions were negotiated and included in that contract to allow public display of data and derived metrics at JYUcite, and to prominently display Dimensions logo and name in that web service. No provisions were made regarding academic outputs and none of the authors are affiliated or compensated by Digital Science Ltd, and their employees were not involved in carrying out this research. JYUcite is a publicly available free service running on resources provided by University of Jyväskylä, providing no financial benefits to authors or their institutions.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics*, 6(1), 36–43. <https://doi.org/10.1016/j.joi.2011.08.002>
- Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *C&RL News* 68, 314–316. Retrieved from <https://crln.acrl.org/index.php/crlnews/article/viewFile/7804/7804>
- Bornmann, L., & Leydesdorff, L. (2013). The validation of advanced bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics*, 7, 286–291. <https://doi.org/10.1016/j.joi.2012.12.003>
- Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, 7(1), 158–165. <https://doi.org/10.1016/j.joi.2012.10.001>
- Bornmann, L., & Marx, W. (2013). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, 98, 487–509. <https://doi.org/10.1007/s11192-013-1161-y>
- Brembs, B. (2018). Prestigious science. *Journals Struggle to Reach Even Average Reliability*. <https://doi.org/10.3389/fnhum.2018.00037>
- Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, 34(2), 1–24. <https://doi.org/10.18637/jss.v034.i02>
- Fang, F. C., & Casadevall, A. (2016). Grant funding: Playing the odds. *Science*, 352(6282), 158. <https://doi.org/10.1126/science.352.6282.158-a>
- Garfield, E. (1955). Citation indexes to science: A new dimension in documentation through association of ideas. *Science*, 122, 108–111. Retrieved from <https://garfield.library.upenn.edu/essays/v6p468y1983.pdf>
- Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association*, 295, 90–93. <https://doi.org/10.1001/jama.295.1.90>
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569–16572. <https://doi.org/10.1073/pnas.0507655102>
- Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. (2016a). Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLoS Biology*, 14(9), e1002541. <https://doi.org/10.1371/journal.pbio.1002541>

- Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. (2016b). S1 Text. *Supporting Text and Equations*. <https://doi.org/10.1371/journal.pbio.1002541.s018>
- Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Sleeping beauties in science. *Proceedings of the National Academy of Sciences United States of America*, 112(24), 7426–7431. <https://doi.org/10.1073/pnas.1424329112>
- Janssens, A. C. J. W., Goodman, M., Powell, K. R., & Gwinn, M. (2017). A critical evaluation of the algorithm behind the Relative Citation Ratio (RCR). *PLoS Biology*, 15(10), e2002536. <https://doi.org/10.1371/journal.pbio.2002536>
- Johnston, M. (2013). We have met the enemy, and it is us. *Genetics*, 194, 791–792. <https://doi.org/10.1534/genetics.113.153486>
- Koener, R. (2020). *quantreg: Quantile Regression. R package version 5.67*. <https://CRAN.R-project.org/package=quantreg>
- Koener, R., & Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448), 1296–1310. <https://doi.org/10.1080/01621459.1999.10473882>
- Lewis, J., Ossowski, S., Hicks, J., Errami, M., & Garner, H. R. (2006). Text similarity: An alternative way to search MEDLINE. *Bioinformatics*, 22(18), 2298–2304. <https://doi.org/10.1093/bioinformatics/bt1388>
- Leydesdorff, L., & Ophof, T. (2010). Normalization at the field level: Fractional counting of citations. *Journal of Informetrics*, 4, 644–646. <https://doi.org/10.1016/j.joi.2010.05.003>
- Leydesdorff, L., Bornmann, L., Mutz, R., & Ophof, T. (2011). Turning the tables in citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, 62(7), 1370–1381. <https://doi.org/10.1002/asi.21534>
- Liu, M., Choy, V., Clarke, P., et al. (2020). The acceptability of using a lottery to allocate research funding: A survey of applicants. *Research Integrity and Peer Review*, 5, 3. <https://doi.org/10.1186/s41073-019-0089-z>
- Long, J. A. (2020). *jtools: Analysis and Presentation of Social Scientific Data. R package version 2.1.0*. <https://cran.r-project.org/package=jtools>
- Ministry of Education and Culture (Finland). (2019). *Publication data collection instructions for researchers 2019*. Retrieved from <https://wiki.eduuni.fi/download/attachments/39984924/Publication%20data%20collection%20instructions%20for%20researchers%202019.pdf>
- Moed, H. F., Burger, W. J. M., Frankfort, J. G., & Van Raan, A. F. J. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy*, 14, 131–149. [https://doi.org/10.1016/0048-7333\(85\)90012-5](https://doi.org/10.1016/0048-7333(85)90012-5)
- Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4, 265–277. <https://doi.org/10.1016/j.joi.2010.01.002>
- Moore, S., Neylon, C., Paul Eve, M., et al. (2017). “Excellence R Us”: University research and the fetishisation of excellence. *Palgrave Communications*, 3, 16105. <https://doi.org/10.1057/palcomms.2016.105>
- Norris, M., & Oppenheim, C. (2010). The h-index: A broad review of a new bibliometric indicator. *Journal of Documentation*, 66(5), 681–705. <https://doi.org/10.1108/00220411011066790>
- Ophof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS (“Leiden”) evaluations of research performance. *Journal of Informetrics*, 4, 423–430. <https://doi.org/10.1016/j.joi.2010.02.003>
- Perianes-Rodríguez, A., & Ruiz-Castillo, J. (2015). Multiplicative versus fractional counting methods for co-authored publications. The case of the 500 universities in the Leiden Ranking. *Journal of Informetrics*, 9(4), 974–989. <https://doi.org/10.1016/j.joi.2015.10.002>
- Pulverer, B. (2013). Impact fact-or fiction? *EMBO J*, 32, 1651–1652. <https://doi.org/10.1038/emboj.2013.126>
- Radicchi, F., & Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, 6, 121–130. <https://doi.org/10.1016/j.joi.2011.09.002>
- Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E*, 83, 046116. <https://doi.org/10.1103/PhysRevE.83.046116>
- Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 17268–17272.
- Schreiber, M. (2009). A case study of the modified Hirsch index hm accounting for multiple co-authors. *Journal of the American Society for Information Science and Technology*, 60(6), 1274–1282. <https://doi.org/10.1002/asi.21057>
- Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314, 498–502. <https://doi.org/10.1136/bmj.314.7079.497>

- Seppänen, J. T., Värri, H., & Ylönen, I. (2020). Co-Citation Percentile Rank and JYUcite: A new network-standardized output-level citation influence metric and its implementation using Dimensions API. *bioRxiv*. <https://doi.org/10.1101/2020.09.23.310052>
- Seppänen, J. T. (2020). *Source Code and Example Data for Article: Co-Citation Percentile Rank and JYUcite: A New Network-Standardized Output-Level Citation Influence Metric*. <https://doi.org/10.17011/jyx/datas-et/71858>
- Stringer, M. J., Sales-Pardo, M., & Nunes Amaral, L. A. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, 3, e1683. <https://doi.org/10.1371/journal.pone.0001683pmid:18301760>
- Ye, F. Y., & Bornmann, L. (2018). “Smart girls” versus “sleeping beauties” in the sciences: The identification of instant and delayed recognition by using the citation angle. *Journal of the Association for Information Science and Technology*, 69(3), 359–367. <https://doi.org/10.1002/asi.23846>
- Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011a). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 2011(5), 37–47. <https://doi.org/10.1016/j.joi.2010.08.001>
- Waltman, L., Yan, E., & van Eck, N. J. (2011b). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics*, 89, 301–314. <https://doi.org/10.1007/s11192-011-0449-z>
- Waltman, L. (2015). *NIH’s new citation metric: A step forward in quantifying scientific impact?* Retrieved from <https://www.cwts.nl/blog?article=n-q2u294&title=nihs-new-citation-metric-a-step-forward-in-quantifying-scientific-impact#sthash.w1KC3A1O.dpuf>
- Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2), 365–391. <https://doi.org/10.1016/j.joi.2016.02.007>
- Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, 342, 127–132. <https://doi.org/10.1126/science.1237825>
- Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, 101(1), 125–158. <https://doi.org/10.1007/s11192-014-1423-3>
- Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology*, 59, 1856–1860. <https://doi.org/10.1002/asi.20880>