Ranking the Ranker: How to Evaluate Institutions, Researchers, Journals, and Conferences?

Wil M. P. van der Aalst · Oliver Hinz · Christof Weinhardt

1 Introduction

How to evaluate institutions, researchers, journals, and conferences? The ranking of scientific research in all its dimensions is food for discussion and the source of major controversies. As editors of the Business & Information Systems Engineering (*BISE*) journal, we want our journal to score excellently in rankings. As individual *BISE* researchers, we want our research to have a significant impact and see this reflected in rankings. As university employees, we want our university to score well in the global university rankings. Rankings are considered important if one scores well. If one does not score well, then one often finds reasons to downplay the ranking's importance. Due to the availability of data, it has become easier to generate rankings. Also, scholarly interest in rankings has increased, and "ranking the ranker" has become a vibrant area of study (Hazelkorn 2018; Ringel et al. 2021; Moed et al. 1985; Stolz et al. 2010). Rankings also impact individual careers, influence where students

W. M. P. van der Aalst (✉)
Lehrstuhl für Informatik 9, RWTH Aachen, Ahornstr. 55, 52056 Aachen, Germany
e-mail: wvdaalst@pads.rwth-aachen.de

O. Hinz
Faculty of Economics and Business Administration, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany
e-mail: ohinz@wiwi.uni-frankfurt.de

C. Weinhardt
Institute of Information Systems and Marketing (IISM), Karlsruhe Institute of Technology (KIT), Kaiserstr. 89-93, 76133 Karlsruhe, Germany
e-mail: weinhardt@kit.edu

want to study, and play a major role in the distribution of research funding.

Although the different types of rankings are widely used, there are also many concerns. The San Francisco Declaration on Research Assessment (DORA) raised concerns related to the "number-based evaluations" of academics (DORA 2012). The declaration starts with the statement, "There is a pressing need to improve the ways in which the output of scientific research is evaluated by funding agencies, academic institutions, and other parties". The DORA declaration also provides 18 recommendations, grouped according to their intended audience: funding agencies, institutions, publishers, organizations that supply metrics, and researchers. The general recommendation is "Do not use journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions." (DORA 2012). It is hard to disagree with these recommendations, but a decade after the DORA declaration, better mechanisms still seem to be missing.

In some countries and institutions, it is now even *forbidden* to mention numerical data (like H-index and number of citations) in grant applications. However, reviewers immediately search for the Google Scholar pages of the applicants to get a first impression. Due to the broadness of the different scientific disciplines, it is hard to judge work in a purely qualitative manner. Similarly, it is also close to impossible to make objective tenure decisions that are not based on objective data, such as the number of published papers in different categories, citations, and grants. Completely abandoning numerical data ("bibliometric denialism") creates uncertainty and may lead to highly subjective and only allegedly "fairer" decisions (e.g., years of hard work are judged based on someone's presentation skills).

Moreover, we witness fierce international competition to attract both scientific staff and top students. Here, university rankings do play a major role. Therefore, we cannot simply ignore rankings, whether we like them or not. In this editorial, we give an overview of the different types of rankings and discuss their applicability. Figure 1 provides a high-level overview of the three types of rankings considered.

Note that also in science, we can observe the Matthew effect of accumulated advantage. The Matthew principle is also known as "the rich get richer and the poor get poorer" and can be explained by preferential attachment, whereby wealth or credit is distributed among individuals according to how much they already have. This also applies to science. For a highly-ranked university, it is easier to attract excellent researchers, making the university even stronger. For a highly-cited researcher, it is easier to receive research funding, resulting in more PhDs and scientific output. Although the Matthew effect seems unfair, it is also partly inevitable.

There is also a competition between different fields of science. *BISE* researchers compete with researchers in physics, medicine, energy, engineering, and production. Therefore, it is helpful to understand the different rankings

and reflect on them. For example, the databases and rankings by Clarivate have a strong bias toward specific disciplines (e.g., physics) and tend to downplay the impact and volume of *BISE* research (Ioannidis et al. 2019).

2 Ranking Institutions

First, we consider the rankings at the institutional level, i.e., mostly universities. These rankings often also provide a ranking per subject. ShanghaiRanking Consultancy annually publishes the *Academic Ranking of World Universities* (ARWU) and *the Global Ranking of Academic Subjects* (GRAS) (www.shanghairanking.com). The ARWU ranking is based on the number of alumni and staff winning Nobel prizes and Fields medals, the number of highly cited researchers selected by Clarivate, the number of articles published in journals of *Nature* and *Science*, and the number of articles indexed in Science Citation Index Expanded and Social Sciences Citation Index (Web of Science). Note that ARWU heavily relies on Clarivate data, as well as specific awards (e.g., Nobel prizes) and journals (e.g., *Nature*). This means that areas such as Computer Science (where conferences are important and there are
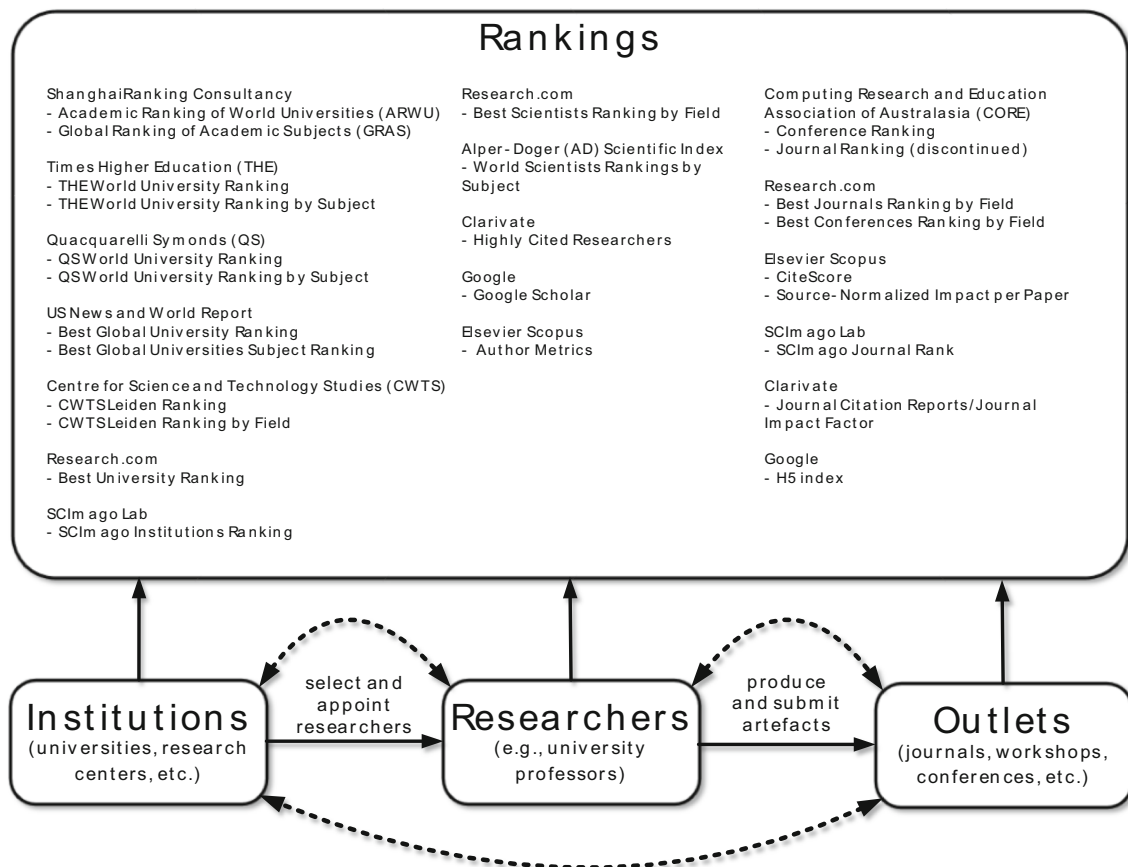

Fig. 1 The interplay between rankings of institutions, researchers, and outlets (e.g., journals and conferences)

"only" Turing award winners instead of Nobel prize winners) are undervalued. The GRAS ranking uses 54 subjects, including Computer Science and Engineering, Economics, Business Administration, and Management.

Times Higher Education (THE) annually publishes *THE World University Ranking* and *THE World University Ranking by Subject* (www.timeshighereducation.com). These rankings use Elsevier's Scopus database. Citations account for 30% of the score. Other elements include student-to-staff ratios, reputation, research income, and proportion of international students. There are 11 subject rankings. Most relevant for *BISE* are Business and Economics, and Computer Science.

Quacquarelli Symonds (QS) publishes the *QS World University Ranking* and the *QS World University Ranking by Subject* (www.topuniversities.com). Like THE, QS also uses Elsevier's Scopus database. Citations account for only 20% of the score. Academic reputation accounts for 40%. Other criteria are international student ratio, international faculty ratio, faculty-to-student ratio, and employer reputation. The QS World University Ranking by Subject covers a total of 54 disciplines, grouped into five broad subject areas. Most relevant for *BISE* are Computer Science and Information Systems, Data Science, Business and Management, and Economics and Econometrics.

As Fig. 1 shows, there are many other university rankings. For example, US News and World Report produces the *Best Global University Ranking* and the *Best Global Universities Subject Ranking* (www.usnews.com/rankings). The Centre for Science and Technology Studies (CWTS) in Leiden publishes the *CWTS Leiden Ranking* and *CWTS Leiden Ranking by Field* (www.leidenranking.com). SCImago Lab publishes the *SCImago Institutions Ranking* (www.scimagoir.com), and Research.com publishes the *Best University Ranking* (research.com). Note that the latter ranking is only provided per subject category and is solely based on researchers with a high Hirsch index.

All of these rankings use different methodologies. Some focus more on scientific output, others more on reputation. Some are more forward-looking, and others are more backward-looking. Therefore, there are differences, but these tend to be smaller than expected (especially for the top 100). Indicators often seem to be selected due to their availability. Also, some measures are size-dependent, making it impossible for smaller or specialized universities to achieve a high overall ranking.

When it comes to research output, the sum of the research outputs of the institution's researchers matters. When it comes to reputation, both current staff and earlier students and staff matter. This shows that hiring and retaining the best researchers is vital for universities. Due to the Matthew effect, this leads to a further concentration of talent.

3 Ranking Researchers

Next, we consider the rankings at the individual level. These rankings are often seen as controversial (Van der Aalst 2022). Whereas university rankings generate revenue through advertisements (and are therefore managed in a professional manner), individual researcher rankings tend to be informal or a side-product of some other service. There are many rankings in specific subfields, e.g., in economics. There are only a few that cover all disciplines. Research.com publishes the *Best Scientists Ranking by Field*. This ranking is based on a scholar's D-index (Discipline H-index), which takes into account only publications and citation metrics for an examined discipline. The fields Business and Management, as well as Computer Science are most relevant for *BISE*. The Alper-Doger (AD) Scientific Index publishes the *World Scientists Rankings by Subject* (www.adscientificindex.com), which is based on the total and last five years' values of the i10 index, H-index, and citation scores in Google Scholar. Clarivate maintains a list of *Highly Cited Researchers* based on the Web of Science (clarivate.com/highly-cited-researchers). Finally, Elsevier Scopus provides several author metrics that can be used to create rankings easily.

The easiest way to evaluate productivity and impact is to simply count the number of published papers and the number of citations. Clearly, this is very naïve because it is possible to publish many papers that are incremental or of low quality. Counting the total number of citations is also problematic because a researcher may be an "accidental co-author" of a high-cited paper. This does not say much about the contribution of the author, and citations tend to follow a power-law distribution (i.e., just a few papers attract most of the citations). To address the limitations of simply counting papers and citations, the scientific community has created journal and conference rankings, and metrics like the well-known Hirsch index. This H-index was first proposed by Jorge E. Hirsh in 2005 and adapted in many different ways (Harzing and Alakangas 2016).

The DORA declaration mentioned before advocates *not* using such measures (DORA 2012). In the Netherlands, the "Recognition and Rewards" ("Erkennen en Waarderen") program (NWO 2019) was initiated to improve the evaluation of academics and to give credits to people working in teams or focusing on teaching. Similar initiatives can be seen in other countries and at the European level (COARA 2022). Although the goals of such programs are reasonable, and it is impossible to disagree with statements such as "quality is more important than quantity" and "one should recognize and value team performance and interdisciplinary research", suitable measures are lacking. Such initiatives are often used to dismiss any attempt to quantify and evaluate productivity and impact. In some universities,

it has even become "politically incorrect" to talk about published papers and the number of citations. In Torres-Salinas et al. (2023), this phenomenon is described as "Bibliometric denialism" and an incorrect interpretation of the DORA declaration, which primarily focused on abuse and misuse of the Journal Impact Factor (JIF). When evaluating and selecting academics, committee members typically still secretly look at the data provided by Google Scholar, Scopus, and Web of Science. This is because it is challenging to evaluate and compare academic performance in an objective and qualitative way. In fact, not using quantitative data creates the risk that evaluations and selections become highly subjective, e.g., based on taste, personal preferences, and criteria not known to the individuals evaluated. Moreover, in such processes, quantitative data are often still used, but in an implicit, secretive, and inconsistent manner.

Therefore, despite all the problems, we often still need to resort to data-driven approaches to evaluate productivity and impact. Of course, quantitative measures should only support expert assessment and are not a substitute for informed judgment. When using citation scores, one should definitely consider the "Leiden Manifesto for research metrics" (Hicks et al. 2015), which provides ten principles to guide research evaluations.

As elaborated in Sect. 4, it is also not easy to rank outlets (journals, conferences, workshops, etc.). Therefore, in this section, we confine ourselves to counting output and impact in terms of citations. There are multiple databases that can be used to evaluate productivity and impact, e.g., Elsevier's Scopus and Google Scholar (both released in 2004) and Web of Science (online since 2002). Also, dedicated tools running on top of these platforms, such as InCites (using the Web of Science) and SciVal (using Scopus), have been developed. Web of Science has a strong focus on journals published in the US and favors traditional disciplines such as physics. Conferences are only partially covered. For a *BISE* researcher, the number of citations in Google Scholar may be twice the number of citations in Scopus, and over eight times the number of citations in Web of Science. For a researcher in Physics, the differences between Google Scholar, Scopus, and Web of Science tend to be much smaller. This means that the Web of Science should not be used for underrepresented disciplines like *BISE*. Google Scholar has the most extensive coverage, but also data quality problems. Google Scholar simply crawls academic-related websites and also counts non-peer-reviewed documents. One may also find stray citations where minor variations in referencing lead to duplicate records for the same paper (Harzing and Alakangas 2016). Also, the output of different authors may be merged into one user profile. Also Scopus and Web of Science have such problems, but to a lesser degree. These examples illustrate that the impact of data quality problems and limited coverage are not equally distributed. Considering data quality and coverage, Scopus can be seen as the "middle road" when counting publications and citations (Baas et al. 2020; Harzing and Alakangas 2016; Van der Aalst 2022).

Another complication is that there are different publication traditions that significantly impact the most common measures used today. In many disciplines, the average number of authors is around two. However, in areas like physics, the average is above ten authors, and there are papers with hundreds or even thousands of authors. An article on measuring the Higgs Boson Mass published in Physical Review Letters has 5,154 authors (Aad et al. 2015). This 33-page article has 24 pages to list the authors, and only nine pages are devoted to the actual paper. When counting H-indices in the standard way, this paper will increase the H-index by one for more than 5,000 authors. Also, the order in which authors are listed varies from discipline to discipline. In mathematics, it is common to list authors alphabetically. In other fields, the order is based on contribution. Also, the "last author" position may have a specific meaning (e.g., the project leader or most senior researcher). In Computer Science, conference publications are regarded as important and comparable to journal publications. In other areas, conference publications "do not count", and all work is published in journals. The above shows that counting just journal papers while ignoring the number of authors may have hugely diverging consequences for different disciplines.

An interesting approach to address some of these concerns was proposed by John Ioannidis and his colleagues (Ioannidis 2022; Ioannidis et al. 2016, 2019, 2020). They propose to use a composite indicator (called *C-score*), which is the sum of the standardized six log-transformed citation indicators (NC, H, Hm, NS, NSF, NSFL):

- the total number of citations received (NC),
- the Hirsch index for the citations received (H),
- the Schreiber co-authorship adjusted Hm index for the citations received (Hm).
- the total number of citations received to papers for which the scientist is single author (NCS),
- the total number of citations received to papers for which the scientist is single or first author (NCSF), and
- the total number of citations received to papers for which the scientist is single, first, or last author (NCSFL).

For a detailed explanation of these indicators, we refer to Ioannidis et al. (2016) and Ioannidis et al. (2019). The resulting C-score focuses on impact (citations) rather than productivity (number of publications) and incorporates information on co-authorship and author positions (single,

first, last author). Each NC, H, Hm, NS, NSF, NSFL score is normalized to a value between 0 and 1, and these are summed up. Hence, the C-score has a range between 0 and 6. In the dataset (Ioannidis 2022), data for 194,983 scientists are reported. The selection is based on the top 100,000 scientists by C-score (with and without self-citations) or a percentile rank of 2% or above in the subfield. The researchers are classified into 22 scientific fields and 174 sub-fields. The dataset is based on all Scopus author profiles as of September 1, 2022, because Scopus can be seen as the middle ground between Google Scholar and Web of Science.

Currently, the C-score seems to be the best way to measure the impact of an author based on her publications. Although the C-score definitely has its limitations and only paints a one-dimensional picture, it removes some of the biases and creates a level playing field when quantifying scientific impact.

4 Ranking Outlets (Journals, Conferences, Etc.)

Researchers produce artifacts such as papers, datasets, prototypes, and software. For software and datasets, one can measure the number of downloads. This can also be done for papers. Downloads and citations are definitely indicators of impact. However, the impact of an artifact can only be measured after some time. This delay complicates decision-making. When a paper is published, it is unclear what impact it will have in five or ten years. Similarly, it is hard to judge the future impact of a PhD thesis for people not directly involved. The PhD student may have left academia before there is "bibliometric evidence" that the thesis realized major breakthroughs. Due to this delay, it is tempting to assign value to the "outlet" of a paper (e.g., journal, conference, or workshop). A paper published in *Science* or *Nature* is expected to have more impact than a paper published in some informal workshop proceedings. A paper accepted for a conference with an acceptance rate of 10% is expected to have more impact and higher quality than a paper accepted for a conference with an acceptance rate of 90%. Therefore, there is a desire to "rank outlets". This has the advantage that one can assign "value" to a paper the moment it is accepted and remove the delay mentioned before. This results in ranked lists of journals and conferences.

However, focused lists of journals and conferences tend to have a topical or geographical bias. For example, in the field of Information Systems (IS), the "College of Senior Scholars" selected a "basket" of journals as the top journals in their field. The goal was to address the problem that few "Information Systems" (IS) journals were widely considered elite-level journals in tenure and promotion cases. However, looking at the selected journals, the field of IS was interpreted in a particular manner. In Europe, IS also includes more technical subjects (e.g., building prototype systems, developing algorithms, and using formal reasoning). This side of IS is not well-represented in the current basket. Some universities create their own local journal lists for specific areas and use these for tenure decisions. This heavily influences the research conducted by young researchers. The CORE ranking of conferences (CORE 2023) is much broader, but has similar problems (e.g., the ranking was established by a few computer departments in Australia and New Zealand and is now used all over the globe to decide on research funding and travel budgets). The intentions behind these lists are good. However, it is unavoidable that there are topical biases and scoping issues. Moreover, such rankings are like a self-fulfilling prophecy. This again leads to a variant of the Matthew effect, i.e., the higher the ranking of a conference or journal, the more people want to submit to it, which automatically improves its status. This, combined with a narrow focus, leads to a degenerate view of research quality and discourages innovations in new directions. Although research is changing rapidly, these journal lists tend to be relatively stable. Also, the editorial boards of these journals aim for a particular type of papers. Excellent, highly innovative papers may be rejected due to scope issues and end up in lower-ranked journals. As a result, young researchers are encouraged to write "what is expected" rather than exploring new research directions.

To avoid subjectivity in ranking journals and conferences, one can use quantitative measures based on citations. Instead of evaluating a researcher, one now evaluates the work published by a journal or conference in a given time period. Figure 1 shows some of the journal and conference rankings.

Well-known metrics based on Elsevier's Scopus are CiteScore, SNIP (Source Normalized Impact per Paper), and SJR (SCImago Journal Rank) (Roldan-Valadez et al. 2019). Well-known metrics based on Clarivate 's Web of Science are JIF (Journal Impact Factor) and 5yIF (Five-year Impact Factor). Google is used to compute the H5 Index. To understand how such metrics are computed, let us consider the way CiteScore, JIF, and H5 are computed for *BISE* for 2023. The *CiteScore* for 2023 is the number of citations in Scopus to *BISE* papers published in 2000, 2021, 2022, and 2023 (four years) divided by the number of papers published by *BISE* in the same period. The *JIF* (Journal Impact Factor) for 2023 is the number of citations to *BISE* papers published in 2021 and 2022 by Web of Science papers in 2023 divided by the number of *BISE* papers published in 2021 and 2022. The *H5* score is the Hirsch index for articles published in the last five years. For 2023, the H5-index for *BISE* is the largest number *X,* such

that *X* articles published in *BISE* in 2018–2022 have received at least *X* citations each (using Google Scholar). As can be noted, the intent of these measures is similar: Measuring impact based on citations. However, the underlying data sources and time scales are different.

The San Francisco Declaration on Research Assessment (DORA 2012) movement was triggered by the obsession of the scientific community with the JIF. Even for journals with astronomical impact factors, the citations of individual papers vary widely. As shown in (Schmid 2018), the average number of citations of the top 10% and bottom 10% of papers published in Nature shows a 20-fold difference. Hence, it is odd to judge a paper based on the JIF of the journal that happened to publish it. Just looking at the outlet itself is not enough to evaluate the quality, novelty, and impact of the work. This was the main trigger for the DORA movement. Unfortunately, this also resulted in widespread "bibliometric denialism", as described in (Torres-Salinas et al. 2023). Peer review and qualitative judgment are difficult to implement and tend to be subjective. Therefore, completely denying quantitative indicators based on bibliometric data seems counterproductive.

5 Implications

As expected, we were not able to answer the question "How to evaluate institutions, researchers, journals, and conferences?" in a satisfactory manner. However, by posing the question and providing an overview of the different types of rankings, we hope to trigger a discussion about what these rankings mean for the *BISE* community. Although these rankings have many limitations and measure what *can* be measured rather than what *should* be measured, they remain highly relevant for *BISE* researchers. We often use the phrase "you get what you measure" to indicate that rankings influence the behavior of students and researchers. It may also explain why particular types of research are conducted in particular countries. In countries with a focus on publishing in a few top-journals that enforce specific research methods, certain types of research cannot flourish. For example, in Computer Science and Europe, there is a stronger focus on conference publications. In Management Science and in the US, there is a stronger focus on journal publications. Academics working on "Information Systems" (IS) in the US tend to work on rather different things than academics working on IS in Europe, e.g., US-based IS researchers tend to have a more social-sciences focus, and European IS researchers tend to work on more technical and conceptual topics. This may explain why Business Process Management (BPM) research thrives in Europe and parts of Asia (e.g., Australia), but is almost non-existent in the US. Of course, this

is not just due to rankings; also, cultural aspects play a significant role. However, for *BISE* researchers, it is good and important to reflect on all these phenomena.

References

Aad G et al (2015) Combined measurement of the Higgs Boson Mass in pp collisions at $\sqrt{s}$= 7 and 8 TeV with the ATLAS and CMS experiments. Phys Rev Lett 114(19):191803. https://doi.org/10.1103/PhysRevLett.114.191803

Baas J, Schotten M, Plume A, Côté G, Karimi R (2020) Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quant Sci Stud 1(1):377–386. https://doi.org/10.1162/qss_a_00019

COARA (2022) Agreement on reforming research assessment. https://coara.eu/. Accessed 29 August 2023

CORE (2023) CORE rankings portal. https://www.core.edu.au/. Accessed 29 August 2023

DORA (2012) San Francisco declaration on research assessment (DORA). https://sfdora.org/. Accessed 29 August 2023

Hazelkorn E (2018) Reshaping the world order of higher education: the role and impact of rankings on national and global systems. Policy Rev High Educ 2(1):4–31. https://doi.org/10.1080/23322969.2018.1424562

Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I (2015) Bibliometrics: the Leiden Manifesto for research metrics. Nature 520:429–431. https://doi.org/10.1038/520429a

Ioannidis J (2022) September 2022 data-update for "Updated science-wide author databases of standardized citation indicators. Elsevier Data Repos. https://doi.org/10.17632/btchxktzyw.5

Ioannidis J, Klavans R, Boyack K (2016) Correction: multiple citation indicators and their composite across scientific disciplines. PLOS Biol 14(8):e1002548. https://doi.org/10.1371/journal.pbio.1002548

Ioannidis J, Baas J, Klavans R, Boyack K (2019) A standardized citation metrics author database annotated for scientific field. PLoS Biol 17(8):e3000384. https://doi.org/10.1371/journal.pbio.3000384

Ioannidis J, Boyack K, Baas J (2020) Updated science-wide author databases of standardized citation indicators. PLoS Biol 18(10):e3000918. https://doi.org/10.1371/journal.pbio.3000918

Moed HF, Burger WJM, Frankfort GJ, Van Raan A (1985) The use of bibliometric data for the measurement of university research

performance. Res Policy 14(3):131–149. https://doi.org/10.1016/0048-7333(85)90012-5

NWO (2019) Recognition and Rewards ("Erkennen en Waarderen") program, an initiative by VSNU, NFU, KNAW, NWO and ZonMw. https://recognitionrewards.nl/. Accessed 29 August 2023

Roldan-Valadez E, Salazar-Ruiz SY, Ibarra-Contreras R, Rios C (2019) Current concepts on bibliometrics: a brief review about impact factor, Eigenfactor score, CiteScore, SCImago Journal Rank, Source-Normalised Impact per Paper, H-index, and alternative metrics. Ir J Med Sci 188:939–951. https://doi.org/10.1007/s11845-018-1936-5

Harzing AW, Alakangas S (2016) Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics 106:787–804. https://doi.org/10.1007/s11192-015-1798-9

Ringel L, Espeland W, Sauder M, Werron T (2021) Worlds of rankings. In: Ringel L et al (eds) Worlds of rankings (research in the sociology of organizations). Bingley, Emerald

Schmid SL (2018) Five years post-DORA: promoting best practices for research assessment. Mol Biol Cell 28(22):2941–2944. https://doi.org/10.1091/mbc.E17-08-0534

Stolz I, Hendel DD, Horn AS (2010) Ranking of rankings: benchmarking twenty-five higher education ranking systems in Europe. High Educ 60:507–528. https://doi.org/10.1007/s10734-010-9312-z

Torres-Salinas D, Arroyo-Machado W, Robinson-Garcia N (2023) Bibliometric denialism. Scientometrics 128:5357–5359. https://doi.org/10.1007/s11192-023-04787-2

Van der Aalst WMP (2022) Yet another view on citation scores. LinkedIn Pulse. https://www.linkedin.com/pulse/yet-another-view-citation-scores-wil-van-der-aalst. Accessed 29 August 2023