

OPINION



When Research Metrics Become Targets: A Leadership Perspective

Graham Kendall

GlobalNxt University, Malaysia and University of Nottingham, UK

Corresponding author: Graham Kendall (e-mail: graham.kendall@nottingham.edu.my)

Keywords: Metrics; h-index; Rankings; Goodhart's law; Stanford/Ioannidis database

ABSTRACTS

Research metrics are deeply embedded in academic life. Publication and citation counts, journal quartiles, h-indexes, university rankings and inclusion in the Stanford/Ioannidis database are used to assess individual researchers, allocate resources, inform recruitment and promotion, compare institutions and communicate academic success. Although these measures provide useful information, they are often presented as single values that obscure the underlying components and methodological choices from which they are derived. Moreover, when metrics become targets they shape both institutional and individual behavior. This opinion article examines several widely used research metrics and considers how they are interpreted, aggregated and applied. It argues that the central problem is not the existence of metrics, but their use as substitutes for academic judgement and as objectives in themselves. Six principles are proposed for the responsible leadership of research metrics: transparency, retaining academic judgement, using a basket of evidence, applying disciplinary and career-stage context, anticipating behavioral consequences and reviewing metrics periodically.

1 | Introduction

Research metrics are deeply embedded and highly influential in scholarly publishing, both at the personal and institutional level. Individual researchers cite how many papers they have published, the quartile (if referring to Web of Science or Scopus indexed journals), the ADBC (Australian Business Deans Council) (<https://abdc.edu.au/abdc-journal-quality-list/>, accessed 10 August 2026) rating, their h-index and whether they appear on the "Stanford List". Institutions are concerned about where they appear in the rankings such as QS World University Rankings and the Times Higher Education World University Rankings. From a research perspective this means that their scholars need to publish in Scopus indexed journals (as these are the data source for both QS and Times Higher Education) and that their work is being cited (as this is one of the indicators that the ranking organizations use) (<https://www.topuniversities.com/world-university-rankings/methodology>, <https://www.timeshighereducation.com/world-university-rankings/methodology>, accessed 10 August 2026)

These metrics are used in hiring/firing, promotion, allocating internal resources and benchmarking against competitors. Funding agencies rely on them when assessing research performance and shaping policy.

They are also used as a marketing asset for the university in areas such as student recruitment and to promote itself on social media channels. Individual scholars use these metrics on their CV, their home pages and on social media platforms, in order to demonstrate their academic prowess.

In this opinion article I give my view of various research metrics and critique them, arguing that they are either not being used for what they were initially intended, or they have become targets, rather than measures. That is, they are being used in a way that they were never designed for, but the research sector is now so invested in them that it is difficult to change them, or even disengage from them should it wish to do so.

2 | Rankings

University rankings reduce multiple indicators to a single score and position, for example the QS World University Rankings use ten indicators in 2026 and nine indicators in 2027. Although this simplifies comparison, it can obscure important differences in institutional performance.

Using the QS 2026 World Rankings, the University of Zurich is ranked 100, with an overall score of 68.5 and the University of

Copenhagen is ranked 101, with an overall score of 68.4. One university can state that it is a “*top 100 university*”, while the other cannot. The ranking establishes that Zurich has a marginally higher composite score under the QS methodology, but it does not necessarily establish that it is substantively better than Copenhagen. Figure 1 shows more of the detail. It shows the ten ranking indicators and their score for the two universities. The arrows show where the University of Zurich (ranked 100) is better or worse than the University of Copenhagen (ranked 101). On five of the indicators Zurich is better and on the other five it is worse. If you were interested in gauging the two institutions, you might look at specific indicators. As an example, if you were to look at research, you might note that

the University of Copenhagen is better on academic reputation and international research networks indicators but worse on citations per faculty. How you use this information is a matter of interpretation, but it demonstrates that the headline figure hides a lot of detailed information.

To provide some context, Figure 2 shows the same two institutions for the 2027 World Ranking. Both institutions have risen in the rankings. Zurich rose two places to 98 and Copenhagen rose eleven places to 90. Figure 2 shows the nine indicators (International Students Diversity is no longer present), with the arrows representing which institution performs the best.

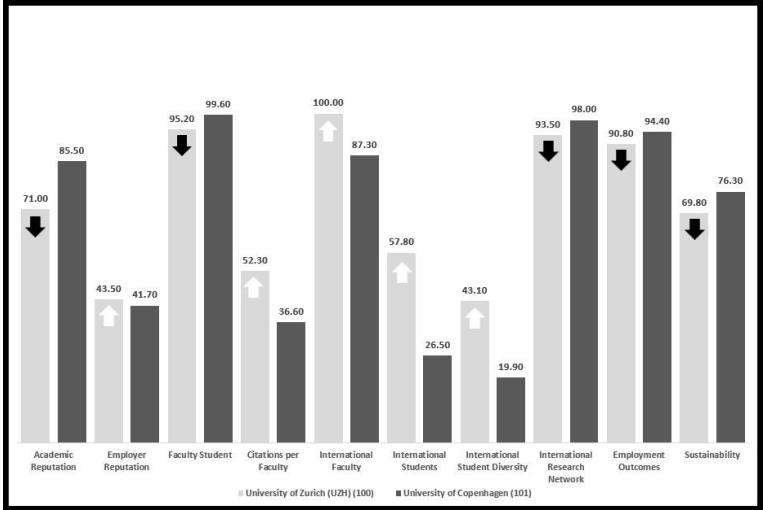


FIGURE 1 | Scores for the ten QS indicators for the University of Zurich and the university of Copenhagen (2026)

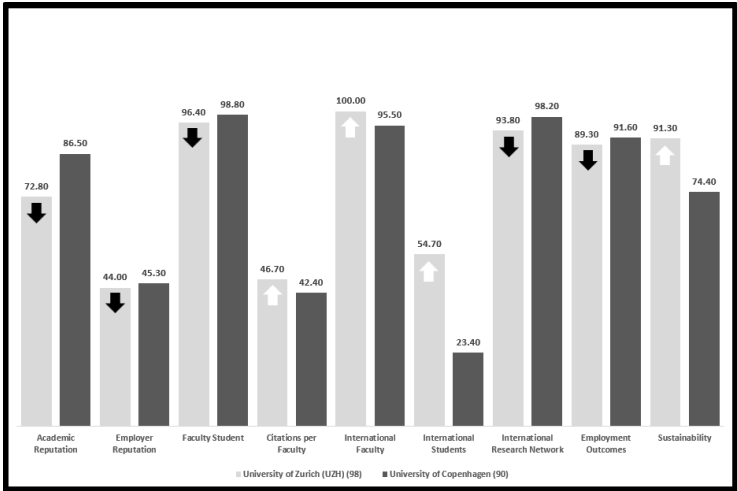


FIGURE 2 | Scores for the nine QS indicators for the university of Zurich and the university of Copenhagen (2027)

2.1 | Aggregation and false precision

University rankings combine a range of indicators into a single composite score. In doing so, they bring together dimensions that may have little in common, such as academic reputation, citations, staff/student ratios, internationalization, employment outcomes and sustainability. Each indicator is normalized, assigned a weight and then aggregated to produce an overall score. These weights are not

neutral. They reflect judgements about what constitutes a successful university and how much importance should be attached to each aspect of institutional performance.

The resulting score can create an impression of precision that the underlying data may not justify. In the example above, a difference of 0.1 points separates the University of Zurich from the University of Copenhagen. Assigning the two institutions ranks of 100 and 101 is

the mechanical consequence of sorting their unequal composite scores and does not, by itself, imply a difference in institutional performance. The interpretive problem arises when stakeholders attach significance to this ordinal distinction. Zurich can describe itself as being in the world's top 100, while Copenhagen cannot, and this categorical label may be understood as evidence that Zurich is materially better. Yet the indicator-level comparison shows that neither university consistently outperforms the other. The overall ordering depends on the way the indicators have been weighted and combined, while the significance attached to the resulting positions depends on how stakeholders interpret and use them.

Several methodological problems have been identified with this process. The nominal weighting assigned to an indicator may not correspond to its actual influence on the final ranking, particularly when indicators have different distributions, while correlations between indicators can result in some dimensions being counted more than once and others contributing little to the composite score^[1-2]. As a result, apparently transparent weighting systems can produce outcomes that differ from their stated methodological intentions. More generally, composite rankings conceal the normative assumptions involved in indicator selection, normalization and aggregation while presenting the result as an unambiguous hierarchy^[3].

One example of this effect is the use of Clarivate's Highly Cited Researchers (HCR) (<https://clarivate.com/highly-cited-researchers/evaluation-and-selection/>, accessed 10 August 2026) as an input to the Academic Ranking of World Universities (ARWU, <https://www.shanghairanking.com/methodology/arwu/2025>, accessed 10 August 2026). The HCR indicator accounts for 20% of ARWU's published weighting. Methodological decisions made by Clarivate concerning database coverage, researcher identification, field assignment, citation patterns and institutional affiliation therefore feed into a separate institutional ranking. HCR is not itself an ordinal ranking of researchers, but its inclusion in ARWU illustrates how the output of one evaluative system can become an input to another. Limitations in the initial identification process may consequently influence the subsequent institutional ranking.

Annual changes in position should be interpreted carefully. A university may rise or fall because its own performance has changed, because other institutions have changed, because the participating population is different, or because the ranking organization has revised its indicators, definitions or weightings. In the example presented here, both Zurich and Copenhagen improved their positions between the 2026 and 2027 QS rankings, but the methodology also changed, with the international student diversity indicator no longer included. A movement of several places cannot automatically be interpreted as evidence that the university itself has improved or declined, although many universities advertise an increase in the rankings without providing any other context.

Rankings are most defensible when treated as broad comparative signals rather than precise measures of institutional quality. Institutions with similar composite scores may have substantially different strengths, while apparently significant boundaries, such as entry into the top 100, may rest on very small numerical differences. Indicator-level profiles may therefore be more informative than the headline position. This problem recurs whenever diverse measures are

aggregated into a single figure that conceals the assumptions and judgements underlying it.

2.2 | Ranking participation and its costs

A university can decide not to participate in rankings but, even if they make this position known, there are still likely to be questions about why this is being done. The justifications might include that it does not want to be judged by an external agency, it does not want its strategy to be informed by the need to chase a ranking position, it does not want to commit financial resources required to participate, or it wants a longer time frame than an annual review cycle to deliver its strategic ambitions. The counter viewpoint is that the institution would not perform well, so does not want to subject itself to that possibility.

Some universities have decided to withdraw. For example, Sorbonne University has withdrawn from the rankings (<https://sciencebusiness.net/news/universities/why-sorbonne-pulled-out-university-ranking>, <https://www.sorbonne-universite.fr/en/news/sorbonne-university-decides-withdraw-times-higher-education-world-university-rankings>, accessed 10 August 2026). Among the reasons given are that rankings cannot be used as management tools, the data used to assess a university is not open or transparent and the reproducibility of the results is not guaranteed. It could be argued that Sorbonne does not need the reputation that ranking provides, which is not the case for many, probably the majority, of other universities. This is particularly true in Asia at the current time.

It is a strategic decision for university leaders whether they wish to participate in the rankings. Many, I feel, believe that they do not really have a choice, and the choice is whether they select QS or Times, or both. Another factor, which is often ignored, or probably more correctly not spoken about as much as it should be, are the resources required to participate in the rankings. These resources come in a number of forms, all with a financial cost. For example:

(1) Data collection and maintenance: Data collection may already be part of an institution's day-to-day activities but participation in ranking(s) may require additional data to be collected and/or more refined dashboards to be developed so that progress can be monitored and reported.

(2) Human resources and organizational resources: Participation in rankings requires staff capacity to collect, interpret, verify and submit institutional data. Some universities have responded by investing in their organization's facilities and/or establishing a department to collect data and monitor performance^[4-5].

(3) Additional activities: As well as requiring more staff time to support the ranking activity, there is likely to be more activity which may not be carried out if ranking(s) were not a priority. For example, I have witnessed many more MOUs (Memorandum of Understanding) being signed to support rankings. Primarily, these facilitate student/staff exchange and joint research initiatives, which feed into the ranking indicators. Other initiatives include hosting webinars, seminars and conferences with the aim of extending an institution's network, which directly feeds into the QS academic reputation survey, which accounts for 30% of the ranking score.

(4) Software infrastructure: The institution may be required to

invest in additional software such as Elsevier's Scopus and the products which are developed by the ranking agencies themselves.

(5) Consultancy: A commercial consultancy industry has developed around university rankings. Institutions may purchase external advice, analytics and benchmarking services to better understand ranking methodologies, assess their relative performance and inform institutional strategy. Evidence shows that universities with established ranking strategies and dedicated internal units continue to purchase consultancy services from ranking providers ^[6].

(6) Conferences: A wider ecosystem of conferences, summits, workshops and other events has developed around university rankings. These events may be used to announce ranking results, explain methodologies, provide masterclasses, facilitate networking and enable dialogue and the sharing of best practice between ranking organizations and university leaders ^[7]. Institutions that regard rankings as strategically important may therefore feel some pressure to participate in these conferences.

2.3 | The Red Queen Effect

Participation in rankings can be viewed through the lens of the Red Queen Effect. The term draws on Lewis Carroll's *Through the Looking-Glass*, in which the Red Queen explains that "*it takes all the running you can do, to keep in the same place*"^[8]. In an organizational setting, improvements by one institution provoke responses from its competitors, meaning that an organization may need to invest continuously merely to maintain its relative position ^[9-10]. Applied to university rankings, institutions may therefore devote increasing resources to ranking-related activities without necessarily improving their position, because competing universities are making similar investments

3 | Research metrics

3.1 | DORA and responsible research assessment

There are calls to stop using research metrics as a measure of success, including by the San Francisco Declaration on Research Assessment (DORA) which says:

"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", with one of its recommendations being

"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."

At the time of writing, 23,641 individuals and 3,681 institutions are signatories of DORA.

Signing DORA (<https://sfdora.org/>, accessed 10 August 2026), however, does not necessarily mean that its principles have been fully embedded in institutional practice. Morgan-Thomas et al. ^[11],

examining the 2021 UK Research Excellence Framework, found that the strong association between journal rankings and expert assessment scores had not changed despite institutional endorsements of DORA. This suggests that there may be a significant gap between formally supporting responsible research assessment and changing the practices through which research is evaluated.

3.2 | Journal metrics and quartiles

Despite calls by organizations such as DORA, institutions and ranking agencies use research metrics as the basis for establishing quality. QS, for example, uses citations, saying that "*... a higher volume of citations suggests that academics at those institutions are publishing in respected journals, engaging in strong collaboration and working on topics that merit a wide readership.*"

Times Higher Education also uses Scopus as its primary data source. The fact that Scopus is a primary data source for (arguably) the two leading university indexing services, results in the scholarly community focusing on publishing in journals that are indexed by Scopus, which reinforces its status, increases citations to articles in those journals and leads to non-indexed journals wishing to become indexed.

This results in scholars saying that they want to publish in Q1 and Q2 (and even Q3 and Q4) journals, noting that the quartile of a journal is directly related to its Impact Factor (Web of Science) or CiteScore (Scopus). In my experience, many researchers do not fully understand what the quartile of a journal represents, especially when you start to consider categories and how impact factors change each year ^[12], yet they want, or are instructed, to only publish in those journals.

3.3 | Publication and citation counts

The simplest research metrics are the number of publications published and the number of citations received. Publication counts are commonly treated as an indicator of productivity, while citation counts are used as an indicator of scholarly influence. Both measures can provide useful information, but neither should be interpreted without context.

Publication counts record outputs rather than their quality, originality or significance. One researcher may publish a small number of substantial studies, while another produces a larger number of incremental papers. The total may also depend on disciplinary conventions, career length, access to research funding and infrastructure, and the extent to which authorship is shared across large research teams. Publication counts reward volume but overlook the nature of the contribution. When institutions set numerical publication expectations, researchers may be encouraged to divide a coherent body of work into several smaller papers (so called salami slicing), adding to the literature but each article not necessarily significantly adding to knowledge.

Citation counts may appear to be a more direct indication that research has attracted attention, but it does not necessarily demonstrate that the work is correct, important or positively regarded. Papers may be cited because they provide a useful method or dataset, because they are

being criticized, or simply because they have become the conventional reference for a particular claim. Citation counts are better understood as indicators of scholarly attention rather than as a measure of research quality.

Comparisons are further complicated by substantial differences in citation practices. Citation rates vary across disciplines, document types and publication years. Researchers in large and rapidly developing fields generally have more opportunities to be cited than those working in smaller fields, while review articles frequently attract more citations than original research papers. Older publications are able to accumulate citations, and bibliographic database coverage also affects citations, as does whether the paper is open access or not. Citation distributions are highly skewed, with many papers receiving few, or zero, citations and a relatively small number attracting very large totals [13-14].

The source of the data is also important. As discussed later in relation to the h-index, Google Scholar, Scopus and Web of Science differ in the publications and citing documents they cover. A citation count should therefore be accompanied by the name of the database and the date on which it was retrieved [12]. Comparisons should ideally be made within the same discipline, career stage, publication period and data source, with appropriate normalization.

Publication and citation counts can be informative when used as part of a broader assessment. They may reveal sustained productivity, identify work that has influenced a field or help institutions understand patterns in their research portfolios. The problem arises when quantity is equated with quality, or citation volume with excellence. Neither measure captures methodological rigor, originality, research integrity, local relevance, societal benefit or the intellectual contribution made by an individual author. These aspects require informed qualitative judgement.

The Leiden Manifesto advises that quantitative evaluation should support, rather than replace, expert assessment and that differences between fields in publication and citation practices must be considered [14]. The leadership challenge is therefore not to reject counts, but to ensure that they remain descriptive evidence rather than definitive judgements about the value of a researcher or their work.

3.4 | Hirsch' s h-index

Hirsch [15] introduced the h-index; one of the measures that scholars and journals are often incentivized to maximize. Hirsch, in his original article, advised caution. For example, “... a single number can never give more than a rough approximation to an individual's multifaceted profile, and many other factors should be considered in combination in evaluating an individual.” and “... the fact that there can always be exceptions to rules should be kept in mind, especially in life-changing decisions such as the granting or denying of tenure.”

Hirsch [15] recognized that citation practices differ across disciplines and argued that self-citations should ideally be excluded. Subsequent studies have identified further limitations of the h-index, while numerous alternative and modified indices have been proposed to address weaknesses relating to career length, disciplinary differences, co-authorship and the distribution of citations [16]. Despite these

limitations, the h-index is widely used and is often reported as a single figure without sufficient context. At a minimum, the database from which the value was obtained, and the date of retrieval should be stated [12]. This is important because differences in coverage mean that Google Scholar, Scopus and Web of Science can produce substantially different citation counts and h-index values, with Google Scholar generally returning the highest value because of its broader coverage [17-18]. Google Scholar is also susceptible to manipulation through the indexing of fabricated or strategically created documents that inflate citation-based indicators [19].

3.5 | Goodhart' s Law

Academic publishing provides a particularly clear illustration of Goodhart's Law [20]. Goodhart's original observation concerned the use of statistical indicators in economic policy. It was later expressed more broadly by Hoskin [21] and Strathern [22], leading to the formulation now commonly known as Goodhart's Law:

“When a measure becomes a target, it ceases to be a good measure.”

This principle is highly relevant to academic publishing and research assessment. Measures such as citation counts, h-indices, journal impact factors and journal quartiles were initially intended to serve as indicators of research influence or quality. Once they become targets, however, researchers, institutions and journals may change their behavior so as to improve the measure itself.

Researchers may prioritize work that is more likely to attract citations, select journals primarily because of their quartile, or divide findings across several papers in order to increase publication output [23]. They may also engage in excessive self-citation or participate in citation cartels intended to inflate apparent research impact [24-25]. Institutions may incorporate publication counts, citation targets or journal rankings into staff performance indicators and promotion criteria [26]. Journals may adopt practices intended to improve citation-based measures, including encouraging additional citations or pursuing editorial strategies designed to improve their Journal Impact Factor, CiteScore or quartile position.

These behaviors do not necessarily indicate that the underlying research has improved. Instead, they illustrate the central problem identified by Goodhart's Law: once a metric becomes an objective, it may become less reliable as an independent measure of the quality or performance it was originally intended to represent.

3.6 | Metrics in recruitment, promotion and performance management

Research metrics become particularly consequential when they are incorporated into decisions about recruitment, promotion and individual performance. They no longer describe aspects of a researcher's record. They now influence who is appointed, who progresses, which activities are rewarded and how academic success is defined.

Metrics can appear attractive to decision-makers because they appear to offer consistency and efficiency. Publication and citation counts, h-

indexes, journal impact factors and journal quartiles can be compared more quickly than a diverse portfolio of research, presented in different ways. They may also provide committees with a degree of reassurance that decisions are supported by objective evidence. However, this apparent objectivity conceals differences between disciplines, career stages, publication practices and access to research resources.

Journal-level metrics feature in formal career assessment. McKiernan et al. ^[26] examined review, promotion and tenure documents from 129 universities in the United States and Canada. Journal Impact Factors or closely related terms were mentioned by 40% of research-intensive institutions and 18% of master's institutions. Among institutions that mentioned the Journal Impact Factor, 87% had at least one statement that supported its use, while none explicitly prohibited it. The metric was commonly associated with judgements about research quality, impact, importance or prestige, despite the Journal Impact Factor being a journal-level measure rather than a measure of the quality of an individual article.

A recent analysis examined 532 promotion policies issued by academic institutions and government agencies across 121 countries. Research outputs appeared in 97% of the policies, and quantitative approaches to evaluating those outputs were more common than qualitative approaches. Publication counts were among the most frequently used quantitative criteria, appearing in 61% of the policies examined ^[27].

Such measures may be used as thresholds, targets or supporting evidence. An institution might require a specified number of publications, place greater value on articles in particular journal quartiles, establish citation expectations or use the h-index as evidence of sustained research influence. Metrics may also enter annual appraisal systems, workload models, probation reviews and performance improvement processes. The choice of metric communicates what the institution values, influencing how academics allocate their time.

This can have unintended consequences. Researchers may favor projects that are likely to generate publications quickly, select topics with greater citation potential, divide findings into multiple papers or avoid locally important research that is unlikely to appear in highly ranked journals. Early-career researchers, scholars in less citation-intensive disciplines and those who have taken career breaks may be disadvantaged by measures that reward accumulated output. Metrics may also undervalue collaborative, interdisciplinary or applied contributions where individual credit is difficult to separate.

The problem is not that quantitative evidence has no place in career assessment. Publication and citation data can provide useful information, particularly when interpreted alongside disciplinary norms and the candidate's career stage. The difficulty arises when the metric becomes a substitute for reading the work, assessing its contribution or exercising academic judgement. A high journal impact factor does not establish that a particular paper is rigorous or important, just as a modest citation count does not demonstrate that the research lacks value.

University leaders need to consider not only which metrics are used,

but what behaviors those metrics are likely to encourage. Recruitment and promotion criteria are powerful statements of institutional priorities. When narrow bibliometric targets are embedded in them, the institution may unintentionally reward success in producing the numbers rather than success in producing valuable research. Responsible assessment requires metrics to inform judgement, not predetermine it.

3.7 | The Stanford/Ioannidis database

The database commonly referred to as the "Stanford List" was developed by Ioannidis et al. to provide standardized information about the citation impact of researchers across different scientific fields. The original study combined six indicators: total citations, h-index, a co-authorship-adjusted hm-index, and citations to papers in which the researcher was the sole author, sole or first author, or sole, first or last author. These measures were combined into a composite citation score, with researchers classified across scientific fields and subfields to allow more meaningful comparisons between disciplines ^[28].

The database contains separate measures of career-long and single-year citation impact. Inclusion is not based solely on being among the top 2% of all researchers worldwide. Researchers are included if they are among the 100,000 highest-ranked authors according to the composite indicator or fall within the top 2% of their principal subfield among authors who have published at least five papers ^[29]. The commonly used description of the database as a list of the "world's top 2% of scientists" is therefore an oversimplification. It is more accurately understood as a database of researchers with high citation impact according to a particular composite bibliometric methodology.

The career-long and single-year datasets provide different temporal views. The career-long dataset reflects accumulated citation impact, whereas the single-year dataset captures citations received during a recent calendar year. This distinction provides useful information about sustained and more recent citations. However, the availability of two variants also creates scope for ambiguity when researchers or institutions report inclusion among the "world's top 2%" without identifying the relevant dataset, citation year, field or selection criterion. These details should therefore be stated whenever inclusion is communicated.

The Stanford/Ioannidis database is examined here because its structure and accessibility make it particularly suitable for considering how researcher-level citation metrics are constructed, interpreted and utilized. The database provides a large, publicly downloadable dataset containing individual citation indicators, a composite citation score, field and subfield classifications, and values with and without self-citations. This level of transparency permits scrutiny of both the methodology and individual outcomes. Its selection should not be interpreted as a claim that it is superior to other systems, including Clarivate HCR. HCR uses Web of Science data and combines quantitative citation criteria with additional screening, but it does not publish an equivalent comparison population or the underlying researcher-level information required for the examination undertaken here.

The database provides useful information. It adjusts, at least partly, for disciplinary differences, co-authorship and author position, and it

reports results both including and excluding self-citations. It is therefore more sophisticated than relying on total citations or an unqualified h-index. However, the authors themselves acknowledge that citation indicators are widely misused and that no bibliometric method can identify all of the “best” scientists^[13, 28]. Citation impact is not equivalent to research quality, originality, integrity or societal contribution, and inclusion in the database should not be interpreted as a comprehensive assessment of scientific excellence.

The more significant concern is how the database has come to be used. Universities and individual researchers increasingly present inclusion as an award, distinction or external validation of quality. Institutions announce how many of their staff appear in the database, researchers include the designation on their CV and social-media profiles, and the results may be used as evidence of institutional research strength.

A bibliometric dataset designed to provide standardized citation information risks becoming a target rather than a measure. Once inclusion is treated as evidence of individual or institutional success, researchers and universities have incentives to maximize the indicators on which it depends. Some of these improvements may reflect influential research, but others may result from strategic publication, authorship or citation practices that increase bibliometric standing without necessarily improving research quality.

There is evidence that self-citation can materially affect positions within the database. An analysis of researchers included in the dataset found that excluding self-citations produced substantial changes in the rankings of countries, institutions, fields and subfields. The effect varied considerably across fields and national systems, demonstrating that apparently comparable positions may be influenced by different citation practices^[30]. The database itself reports indicators with and without self-citations, but public claims of inclusion rarely provide this context.

Other concerns relate to data quality and interpretation. Del Fabbro et al.^[31] identify persistent anomalies in the Stanford/Ioannidis database arising from the inclusion of non-research publication types, incomplete historical Scopus coverage and researchers whose citation profiles are substantially associated with retracted work. They also note that the presence of retractions does not automatically exclude a researcher from the database. Importantly, however, their analysis shows that the overall ranking is highly stable, with the great majority of the top-ranked researchers remaining in place after the proposed corrections were applied. Their critique is not that the database is fundamentally unreliable, but that a relatively small number of visible anomalies can weaken confidence in it and should be addressed through more robust filtering and reporting.

Some of these concerns arise from the design and data sources of the Stanford/Ioannidis database. These include its reliance on Scopus author profiles, the construction of its composite citation score, its career-long and single-year variants, and its treatment of fields, self-citations and retracted publications. Their prevalence should not be assumed to be the same under HCR, which uses a different bibliographic database, citation window and selection process. Other concerns are more structural and extend beyond any single system. These include disciplinary and career-stage differences in citation practices, threshold effects created by inclusion and exclusion,

institutional incentives associated with recognized researcher lists, and the interpretation of citation-based recognition as evidence of broader research quality. It is therefore important to distinguish limitations arising from the design of a particular system from those associated more generally with the evaluative use of researcher citation benchmarks.

It would be too strong to conclude that everyone who appears in the database has gamed the system. Most researchers are likely to be included because their work has received substantial and sustained attention. Nor does the available evidence establish how many researchers have deliberately altered their behavior specifically to gain inclusion. The more defensible concern is that, once the database is treated as a prestigious ranking, it creates incentives for behavior that can increase the underlying indicators without necessarily improving the quality, integrity or value of the research.

Inclusion should therefore be reported with care. At a minimum, claims should specify whether they relate to career-long or single-year impact, identify the year of the dataset, state the relevant field or subfield, and avoid presenting the database as a definitive list of the world’s best scientists. It is a useful bibliometric resource, but it should remain an indicator rather than becoming an accolade that substitutes for academic judgement.

A like-for-like replication using Clarivate HCR would not be possible because HCR does not publish an ordinal ranking of all included researchers, an equivalent comparison population or the complete underlying information used in its final selection process. The two systems also represent somewhat different approaches: the Stanford/Ioannidis database orders researchers using a published composite citation indicator, whereas HCR identifies a much smaller group through citation thresholds followed by additional quantitative and qualitative screening. Future research could examine whether the institutional and behavioral consequences discussed here are evident when HCR status is used in recruitment, promotion, institutional comparison or research assessment.

4 | Principles for responsible leadership of research metrics

The preceding sections have reviewed how research metrics are constructed, interpreted and used, together with some of their limitations and unintended consequences. This section gives my views on how university leaders can use metrics responsibly without allowing them to replace academic judgement or become objectives in themselves.

4.1 | Be transparent about the metric

A metric should never be presented as a number without context. Leaders should require clarity about what has been measured, how it has been calculated and where the underlying data came from. For citation-based measures, this should include the database, date of retrieval, period covered and whether self-citations have been included. Journal metrics should identify the year, database and subject category, while claims based on researcher citation benchmarks should identify the system and edition used. In the

Stanford/Ioannidis case, this should also include whether the claim relates to career-long or single-year impact and the relevant field or subfield.

The same principle applies to institutional rankings and composite measures. The indicators, weightings and methodological assumptions that produce the final figure should be understood before the result is used in strategy, marketing or performance management. Where the calculation cannot be independently reproduced, or where the provenance of the data is unclear, the result should be treated with caution.

Transparency does not remove the limitations of a metric, but it makes those limitations visible. It allows decision-makers to judge whether comparisons are meaningful and helps prevent an apparently precise number from being accepted without sufficient scrutiny.

4.2 | Use metrics to inform, not replace, judgement

Metrics can provide useful evidence, but they should not be the primary factor for any decision-making process. A citation count, h-index, journal quartile or ranking position may highlight aspects of performance, but none can capture the full quality, significance or context of the work being assessed. Leaders should therefore treat metrics as prompts for further enquiry rather than as conclusions.

This is especially important in recruitment, promotion, resource allocation and institutional strategy. A strong metric may justify closer attention, but it should not remove the need to read the work, understand the contribution and consider disciplinary norms, career stage and institutional circumstances. Equally, a modest metric should not automatically be interpreted as evidence of weak performance.

The purpose of judgement is not to override evidence, but to interpret it responsibly. Metrics can help structure discussion and identify patterns, but the final decision should remain an informed academic and leadership judgement. Once the number becomes the decision, the institution has replaced evaluation with calculation.

4.3 | Use a basket of evidence

No single metric can capture the full value of a researcher, a research output or an institution. Publication counts may indicate productivity, citation counts may suggest scholarly attention and journal metrics may provide information about the standing of a publication venue, but each offers only a partial view. Responsible assessment therefore requires a basket of evidence rather than reliance on one dominant measure.

That basket may include quantitative indicators, peer judgement, research quality and originality, integrity, societal or local contribution, collaboration, mentoring and leadership. The balance should vary according to the purpose of the assessment, since recruitment, promotion and institutional benchmarking do not require identical evidence.

Using several forms of evidence does not eliminate subjectivity, but it reduces the risk that one imperfect metric will distort the decision. The aim is not to create a larger composite score, but to consider different

dimensions alongside one another and preserve the distinctions between them.

4.4 | Apply disciplinary and career-stage context

Research metrics are not directly comparable across all disciplines or career stages. Publication rates, authorship conventions, citation practices and the time needed for work to attract attention vary across disciplines. A publication and citation profile that is strong in one discipline may appear modest in another, even though the underlying contribution is equally significant.

Career stage also matters. Established researchers have had more time to publish, accumulate citations, build collaborations and supervise research teams. Early career researchers, those who have taken career breaks and those working in less well-resourced environments should not be assessed against the same accumulated totals without appropriate context.

Leaders should therefore avoid universal thresholds and comparisons that ignore disciplinary norms and individual circumstances. Metrics are most useful when interpreted relative to the field, opportunity and career stage. Context should not be treated as an exception to the assessment process. It is part of what makes the assessment fair and meaningful.

4.5 | Anticipate behavioral consequences

Metrics do not simply measure behavior, they also shape it. Leaders should ask not only whether a metric is easy to collect or appears objective, but what people are likely to do in response to it.

A target for publication volume may encourage salami slicing. A requirement to publish in particular journal quartiles may narrow research choices. Citation targets may increase self-citation or reciprocal citation practices, while ranking targets may redirect institutional effort towards activities that improve the indicator rather than supporting the institution's strategy.

Leaders should ask what behavior this measure will reward? Before adopting a metric as a key performance indicator (KPI), institutions should consider unintended consequences and whether the incentives align with research quality and integrity.

4.6 | Review metrics periodically

Metrics should not become permanent because they are familiar or convenient. Databases change, ranking methodologies are revised, disciplinary practices evolve and institutional priorities shift. A measure that was once useful may therefore become less relevant and less reliable.

Leaders should review not only the technical validity of a metric, but also how it is being used and what effects it is producing. This includes asking whether the measure still aligns with institutional strategy, whether it disadvantages particular disciplines or groups, and whether it is leading to unintended consequences.

Periodic review should also allow institutions to retire metrics that no

longer add value, revise thresholds that have become inappropriate and introduce better forms of evidence where necessary. Responsible leadership requires the willingness to change the measurement system, rather than expecting people to continue adapting to a metric that no longer serves its purpose.

5 | Discussion and conclusions

Research metrics are now embedded in the systems through which universities assess researchers, allocate resources, compare institutional performance and communicate success. They reduce complex activity to (typically) single figures that can be reported, compared and monitored. However, that simplicity obscures the underlying complexity in how those values are derived.

Publication counts, citations, h-indexes, journal metrics and rankings can all provide useful information, when interpreted appropriately. The problem arises when they move from being indicators to becoming targets. Researchers respond to these targets which become embedded in recruitment, promotion and performance systems, while institutions redirect effort and resources towards activities that improve their ranking position.

University leaders cannot easily disengage from this environment. Rankings and bibliometric indicators influence reputation, recruitment, partnerships, funding and stakeholder perceptions. An institution that steps away from them may find that its decision is interpreted not as a principled rejection of inappropriate measurement, but as evidence of weak performance. This creates a genuine leadership dilemma: metrics cannot be ignored, but neither should they be allowed to determine institutional strategy.

Responsible leadership therefore requires balance. Metrics should be transparent, interpreted in context, considered alongside other evidence and reviewed for the behaviors that they encourage. They should inform academic judgement rather than replace it. The greatest risk is not that universities measure research performance, but that the measures become targets and eventually substitutes for research quality, integrity and institutional mission. Metrics should help leaders ask better questions. They should not become the answer.

ACKNOWLEDGEMENT

The author received no financial support for this article.

ETHICAL STATEMENT

This study does not contain any studies with human or animal subjects performed by the author.

CONFLICTS OF INTEREST

The author declares no conflicts of interest.

AUTHOR CONTRIBUTION STATEMENT

The sole author of this article is responsible for all aspects of the article.

DATA AVAILABILITY STATEMENT

There is no data associated with this article.

REFERENCES

- [1] Soh, K. (2013). Misleading university rankings: Cause and cure for discrepancies between nominal and attained weights. *Journal of Higher Education Policy and Management*, 35(2), 206–214. DOI: 10.1080/1360080X.2013.775929
- [2] Soh, K. (2015). Multicollinearity and indicator redundancy problem in world university rankings: An example using Times Higher Education World University Ranking 2013–2014 data. *Higher Education Quarterly*, 69(2), 158–174. DOI: 10.1111/hequ.12058
- [3] Latifis, K., Sidiropoulos, A., & Evangelidis, G. (2026). International university rankings: Composite indicators and methodological approaches. *Frontiers in Research Metrics and Analytics*, 11, 1828850. DOI: 10.3389/frma.2026.1828850
- [4] Hazelkorn, E. (2008). Learning to live with league tables and ranking: The experience of institutional leaders. *Higher Education Policy*, 21, 193–216. DOI: 10.21427/d7pp7z
- [5] Chun, H., & Sauder, M. (2023). The power in managing numbers: Changing interdependencies and the rise of ranking expertise. *Higher Education*, 86(4), 771–790. DOI: 10.1007/s10734-022-00823
- [6] Chirikov, I. (2023). Does conflict of interest distort global university rankings? *Higher Education*, 86, 791–808. DOI: 10.1007/s10734-022-00942-5
- [7] Lim, M. A. (2018). The building of weak expertise: The work of global university rankers. *Higher Education*, 75(3), 415–430. DOI: 10.1007/s10734-017-0147-8
- [8] Carroll, L. (1871). *Through the Looking-Glass, and What Alice Found There*. London: Macmillan
- [9] Barnett, W. P. (2008). *The Red Queen among Organizations: How Competitiveness Evolves*. Princeton University Press
- [10] Derfus, P. J., Maggitti, P. G., Grimm, C. M., & Smith, K. G. (2008). The Red Queen effect: Competitive actions and firm performance. *Academy of Management Journal*, 51(1), 61–80. DOI: 10.5465/AMJ.2008.30708624
- [11] Morgan-Thomas, A., Tsoukas, S., Dudau, A., & Gaska, P. (2024). Beyond declarations: Metrics, rankings and responsible assessment. *Research Policy*, 53(10), 105093. DOI: 10.1016/j.respol.2024.105093
- [12] Kendall, G. (2024) More Transparency is Needed When Citing h-Indexes, Journal Impact Factors and CiteScores. *Publishing Research Quarterly*, 40(1): 80-99. DOI: 10.1007/s12109-024-09983-3

- [13] Ioannidis, J. P. A., Klavans, R., & Boyack, K. W. (2016). Multiple citation indicators and their composite across scientific disciplines. *PLOS Biology*, 14(7), e1002501. DOI: 10.1371/journal.pbio.1002501
- [14] Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520, 429–431. DOI: 10.1038/520429a
- [15] Hirsch, J.E. (2005) An index to quantify an individual's scientific research output. *PNAS* 102 (46): 16569–16572. DOI: 10.1073/pnas.0507655102
- [16] Bihari, A., Tripathi, S., & Deepak, A. (2023). A review on h-index and its alternative indices. *Journal of Information Science*, 49(3), 624–665. DOI: 10.1177/01655515211014478
- [17] Bar-Ilan, J. (2008). Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257–271. <https://doi.org/10.1007/s11192-008-0216-y>
- [18] Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177. DOI: 10.1016/j.joi.2018.09.002
- [19] Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, 65(3), 446–454. DOI: 10.1002/asi.23056
- [20] Goodhart, C. (1989) *Money, Information and Uncertainty*, Macmillan, London.
- [21] Hoskin, K. (1996) *The 'awful idea of accountability': inscribing people into the measurement of objects. Accountability: Power, ethos and the technologies of managing*. London: International Thomson Business Press. pp. 265–282
- [22] Strathern M. (1997) 'Improving ratings': audit in the British University system. *European Review*, 5(3): 305–321. DOI:10.1017/s1062798700002660
- [23] Pasternak, J. (2023). Ethics in publication: Salami slicing, Einstein (São Paulo), 21, eED0837. DOI: 10.31744/einstein_journal/2023ED0837
- [24] Krystal, J. H., Barch, D. M., Bugno, R. M., & Carter, C. S. (2026). Addressing new challenges to scientific integrity including paper mills, citation cartels, coercive citation practices, fake reviewers, and artificial intelligence-generated papers. *Biological Psychiatry*, 99(4), 262–263. DOI: 10.1016/j.biopsych.2025.11.017
- [25] Zaidi SJA & Taqi M. (2023) Citation Cartels in Medical and Dental Journals. *Journal of the College of Physicians and Surgeons Pakistan*, 33(06):700–701. DOI: 10.29271/jcpsp.2023.06.700
- [26] McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *eLife*, 8, e47338. DOI: 10.7554/eLife.47338
- [27] Lim, B. H., D'Ippoliti, C., Dominik, M., et al. (2025). Regional and institutional trends in assessment for academic promotion. *Nature*, 637, 607–615. DOI: 10.1038/s41586-024-08422-9
- [28] Ioannidis, J. P. A., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLOS Biology*, 17(8), e3000384. DOI: 10.1371/journal.pbio.3000384
- [29] Ioannidis, J. P. A., Boyack, K. W., & Baas, J. (2020). Updated science-wide author databases of standardized citation indicators. *PLOS Biology*, 18(10), e3000918. DOI: 10.1371/journal.pbio.3000918
- [30] Mishra, S., et al. (2025). Self-citation pattern among world's top 2% of the scientists. *Heliyon*, 11(3), e42471. DOI: 10.1016/j.heliyon. 2025.e42471
- [31] Del Fabbro, M., Stabene, S., Khijmatgar, S., et al. (2026). Persistent anomalies in the Stanford/Ioannidis top 2% scientist's database: Empirical impact of publication-type filtering, temporal coverage correction, and retraction-rate thresholds. *Scientometrics*. DOI: 10.1007/s11192-026-05682-2