Concerns with bibliometrics in resource allocation decisions

Dan Hicks

2025-07-08

> **Note**
>
> This is an HTML version of a document that I wrote for Leonardo Arriola, Dean of Social Sciences, Humanities, and Arts at UC Merced. There's a PDF version linked in the sidebar.

Executive summary

The UC Merced administration is working to develop a set of "principles and metrics" to be used to make or inform the allocation of faculty lines to departments. This document provides a review of the limitations and problems associated with using publication-based metrics (bibliometrics) for faculty resource allocation decisions. Drawing on my work experience in research evaluation and the established consensus among professionals in that field, I argue that bibliometrics are not fit for purpose for making hiring, tenure, promotion, and faculty line allocation decisions.

Professional consensus

This position is consistent with major professional statements in the field of research evaluation, including the San Francisco Declaration on Research Assessment (DORA) and the Leiden Manifesto, both of which explicitly warn against using journal-based metrics and similar bibliometric measures for personnel decisions. DORA specifically states: "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."

Key technical problems

I review several technical problems with bibliometrics as measures of research quality:

Proxy Limitations: These metrics are at best crude proxies for research quality and impact, measuring outputs and outcomes rather than actual scholarly or social impacts.

Incommensurability: Different types of research outputs are fundamentally incommensurable: there is no meaningful way to compare a historical monograph to a biomedical journal article on a single scale.

Reference Class Problem: Academic disciplines have porous boundaries and individual researchers typically work across multiple communities, making it impossible to accurately normalize metrics across fields.

Data Quality Issues: A fundamental tradeoff exists between data coverage and quality—comprehensive databases suffer from poor data quality and manipulation vulnerabilities, while selective databases systematically underrepresent social sciences and humanities.

Systemic bias concerns

Extensive research demonstrates that bibliometric systems systematically disadvantage women and racial/ethnic minorities through lower citation rates, publication barriers, and exclusionary networks. Rather than providing objective assessment, these systems risk exacerbating existing inequities.

Failure to avoid controversy

Bibliometrics are attractive because they appear to depoliticize controversial resource allocation decisions. But this appearance is incorrect. Instead, they shift political controversies to the kinds of technical implementation issues identified above, creating new sources of institutional conflict while failing to address underlying disagreements about priorities and values.

Recommendations

As an alternative to bibliometrics, I recommend taking an approach to faculty hiring decisions that starts by articulating campus-levels aims for hiring; approaches hiring requests as program impact plans — also known as theories of change and logic models — to explicitly connect proposed hires to those campus-level aims; and incorporates faculty voice through a peer review process, similar to the approaches already used for curricular changes on campus and by research funding agencies.

Introduction

Since last semester, the UC Merced administration has been discussing the development of a set of "principles and metrics" to be used to make or inform the allocation of faculty lines to departments. In SSHA, faculty have participated in these discussions via meetings of the department chairs and Executive Committee, with multiple rounds of feedback on lists of proposed principles and opportunities.

Publication-based metrics have come up repeatedly in these discussions, with both support and criticism from various faculty members. As I explain below, I have some work experience and scholarly expertise on the uses and limitations of bibliometrics. Following the lead of professionals in the relevant fields, I do not believe these metrics are fit for purpose in making hiring, tenure, and promotion decisions, including decisions about how to allocate faculty lines.

In discussions and in writing, I have previously shared brief statements of common technical concerns about bibliometrics with Dean Arriola. The purpose of this document is to provide a more detailed version of these concerns, showing that they are grounded in relevant scholarship and expertise.

In the remainder of this introduction, I first clarify the terms "bibliometrics" and "scientometrics," summarize my background in this area, and review two prominent statements on bibliometrics by research assessment professionals. In the following sections, I argue that bibliometrics are at best crude proxies for the things we academics are trying to do with our scholarship; that research outputs and outcomes are incommensurable in ways that violate the assumptions used to calculate bibliometrics; that bibliometrics are confounded by the "reference class problem"; that there is a sharp tradeoff between bibliometric data coverage and data quality; that bibliometrics can reflect systematic gender and racial/ethnic bias; and finally that they do not serve their broader function of quieting fraught controversies. I close with some procedural recommendations that, I think, would keep controversies from becoming too heated and avoid the favoritism and patronage of the current system.

Bibliometrics and scientometrics

Among specialists, "bibliometrics" refers to "bibliography-derived" measures of scientific productivity and impact, including publication and citation counts as well as impact factors[1] and the h-index[2]. "Scientometrics" is a more general term, covering measures based on patent citations (e.g., Funk and Owen-Smith 2016), coauthor network statistics (Zeng et al. 2016),

[1] Impact factors are usually calculated for journals, not individual researchers, and are defined as the total number of citations to items in the journal divided by the number of items published, over some period of time.

[2] The h-index is used more widely than impact factor, and is defined for some portfolio of publications as the largest number h such that h publications have at least h citations.

semantic novelty (He and Chen 2018; Petersen 2022), or indeed any other quantitative attempt to understand scientific productivity and social impact.

"Scientometrics" is also used as the name of a research field — the scholarly side of the practice of research assessment or research evaluation — that operates at the intersection of science and technology studies, science policy, library and information sciences, and sociology. As a research field, research evaluation/scientometrics has obvious conceptual and methodological ties to science-of-science and metascience. Unlike these other two fields, research evaluation tends to focus more on supporting decisionmaking by funding bodies and research organizations.

My background

From 2015-2017 I served as an AAAS Science and Technology Policy fellow, hosted by the EPA Office of Research and Development's Chemical Safety for Sustainability program (first year) and then by NSF's National Robotics Initiative (second year). In both positions my responsibilities included developing and conducting evaluation projects for the research done (EPA) or funded (NSF) by my host office. Two of these evaluation projects were published and appear on my CV (Daniel J. Hicks 2016; D. Hicks and Simmons 2019).

From 2017-2019 I was a postdoctoral researcher at the Data Science Initiative (now known as DataLab: Data Science and Informatics) at UC Davis. My position was funded by Elsevier[3] and the general aim of the postdoc was to develop new approaches to scientometrics that avoided common criticisms of bibliometrics (as discussed in the rest of this document). In this role I had opportunities to meet with deans and campus-level administrators to develop projects, and worked closely with both faculty and staff. Again, some of these projects were published and appear on my CV (Daniel J. Hicks, Stahmer, and Smith 2018; Daniel J. Hicks et al. 2019; Daniel J. Hicks 2021).

The DORA Statement and Leiden Manifesto

Bibliometrics have played a significant and highly contested role in faculty hiring, tenure, and promotion decisions in a number of countries for several decades, though less so in the United States. In this context, research evaluation experts have released multiple statements providing guidance and criticizing certain uses of bibliometrics. I review two especially influential statements, the San Francisco Declaration on Research Assessment (DORA, published in multiple venues including Cagan 2013) and the Leiden Manifesto for Research Metrics (Diana Hicks et al. 2015).[4]

[3] I wasn't aware of this when I accepted the offer, and had no contact with Elsevier representatives during the last eight months or so of the postdoc, or since.

[4] Note that Diana Hicks is a prominent scholar in this field. We have never met, and are not related.

DORA first offers a number of criticisms of the impact factor: that it was originally created to help librarians make purchasing decisions, not assess research; that means are misleading measures of central tendency for highly skewed distributions such as citation counts; that it fails to distinguish different kinds of articles (commentaries from primary research from reviews) and neglects differences across research fields; and that the data used to calculate impact factors are generally neither public nor subject to peer review. It then lists 18 recommended practices for research assessment, the first of which 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.

For institutions specifically, recommendation 5 states:

> For the purposes of research assessment, consider the value and impact of all research outputs (including datasets and software) in addition to research publications, and consider a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice.

The Leiden Manifesto is motivated by a concern that research evaluation and metric development is "usually well intentioned, not always well informed, often ill applied …. by organizations without knowledge of, or advice on, good practice and interpretation." The Manifesto offers a "distillation of best practice in metrics-based research assessment," in the form ten principles. Some of these principles include:

1. *Quantitative evaluation should support qualitative, expert assessment ….* assessors must not be tempted to cede decision-making to the numbers.
2. *Measure performance against the research missions of the institution, group or researcher.* Programme goals should be stated at the start, and the indicators used to evaluate performance should relate clearly to those goals.
3. *Allow those evaluated to verify data and analysis.* To ensure data quality, all researchers included in bibliometric studies should be able to check that their outputs have been correctly identified.
4. *Account for variation by field in publication and citation practices.* Best practice is to select a suite of possible indicators and allow fields to choose among them.
5. *Avoid misplaced concreteness and false precision.* Science and technology indicators are prone to conceptual ambiguity and uncertainty and require strong assumptions that are not universally accepted.
6. *Recognize the systemic effects of assessment and indicators ….* a suite of indicators is always preferable — a single one will invite gaming and goal displacement (in which the measurement becomes the goal).

These points imply important limitations and concerns about bibliometrics that will be echoed below: the need to base the choice of metrics on organizational goals rather than what's

convenient to measure; significant variation in data coverage and quality; and variations across fields. Point nine is also known as "Goodheart's Law": "when a measure becomes a target, it ceases to be a good measure." Tying resource allocations to publication counts encourages "salami slicing," the rapid publication of many minimally-informative or low-quality (or even fraudulent) papers; while using citation counts incentivizes the formation of "citation cartels" (Kojaku, Livan, and Masuda 2021) or even manufacturing fake papers to inflate citation counts (Ibrahim et al. 2024).

Outputs, outcomes, and impacts

In program evaluation, there's a useful distinction between outputs, outcomes, and impacts. *Outputs* are the direct products of the program's activities, and the things the program has most control over. For research organizations, outcomes include things like scholarly publications, grey literature reports, social media posts by the organization or its members, grant applications submitted, and public-facing events. *Outcomes* are the short-term consequences that follow from those outputs: citations to publications, policymakers and advocates referencing reports, social and traditional media coverage, grant awards, attendance at public events. *Impacts* are the longer-term goals that the organization hopes to achieve by way of its outputs and outcomes. Daniel J. Hicks, Stahmer, and Smith (2018) further distinguish between inward-facing goals ("the value of research for the relevant scholarly community, in terms of the further production of new knowledge") and outward-facing goals ("the value of research for other social practices …. [such as] broader social efforts to protect biodiversity and preserve natural or undeveloped spaces").

These distinctions make it clear that bibliometrics — publication and citation counts, impact factors — are outputs and outcomes, rather than impacts. At best, these metrics are proxies for impacts, and indeed only proxies for inward-facing impacts. They provide at most limited insight into the scholarly impact of research, and none into social impact.

Despite being proxies, bibliometrics (incorrectly) appear to be useful for research assessment because impacts are often implicit, poorly defined, or difficult to measure, especially outward-facing social impacts. For example, in UC Merced's 2021-2030 strategic plan, the first goal is to "Engage Our World and Region Through Discovery and the Advancement of Knowledge." This is explained as including "interdisciplinary and transformational research that supports equity and prosperity globally and locally, with particular sensitivity for the San Joaquin Valley." "Support[ing] equity and prosperity" is an outward-facing impact. It is appealing, but also so vague that the proposed measures don't even try to operationalize it. Instead, the strategic plan lists various output measures (tenure rates, research spending) and a vague bibliometric item ("number of impactful papers," with "[specific] measures to be developed").

This is why the Leiden Manifesto states that "Programme goals should be stated at the start, and the indicators used to evaluate performance should relate clearly to those goals." Without

clear impacts it's impossible to develop measures of progress or achievement with respect to those impacts.

Incommensurability of research outputs and outcomes

In discussions at the SSHA department chairs and Executive Committee meetings, a repeated point has been that the outputs and outcomes are generally incommensurable. There is no meaningful, all-inclusive, unidimensional scale with which to compare a book written by a historian to a journal article written by a psychologist to a systematic review conducted by a team of biomedical researchers.

One aspect of incommensurability in this example is incommensurability by type of research output. Academic researchers produce numerous different kinds of outputs: monographs, anthologies, textbooks and other teaching materials; "original research" journal articles, reviews, theoretical papers, commentaries; white papers, policy reports; blog posts, opinion columns and letters, edutainment podcasts and videos, interactive websites, media interviews; conference presentations, keynotes and invited talks, flash talks, posters; artistic exhibitions and performances; workshops and conferences; public events.

These different kinds of outputs require not just different amounts of work, but different kinds of work and are anticipated to have qualitatively different kinds of outcomes and impacts. Even within a given field, an archive-based historical monograph is intended to do something very different from a short conference presentation or a public-facing website. It therefore does not make sense to ask how much "more" the monograph should "count" than the presentation or the website.

Research outcomes associated with a given type of output can also be incommensurable across fields. It is a commonplace in bibliometrics that the average citation rate differs across fields; for this reason, bibliometrics data providers such as Clarivate (Web of Science) and Elsevier (Scopus) provide "field-weighted" or "normalized" citation statistics.[5] One common explanation is differences in publication rates and citation practices. For example, in philosophy, a tenure-track assistant professor who publishes more than one paper per year (in "good journals") would be considered highly productive; while a typical mid-sized biomedical lab might publish a dozen papers a year. Some fields, such as history, are expected to have lengthy and comprehensive bibliographies; while others, such as biomedicine, typically have short and selective bibliographies (except for systematic reviews, which are expected to be comprehensive, etc.). In line with this, Radicchi, Fortunato, and Castellano (2008) argue that there is a "universal," cross-field lognormal distribution of *relative* citations, that is, the number of citations received by a paper divided by the average number of citations received by all papers published in the same field and year. However, subsequent work on the "universal" citation distribution has not produced a consensus on the functional form it should take — lognormal,

[5]For example: https://service.elsevier.com/app/answers/detail/a_id/14894/supporthub/scopus/related/1/

power-law, exponential, negative binomial, etc. — how it should be parameterized, and what underlying mechanisms might produce it (Golosovsky 2021). Indeed, Marx and Bornmann (2015) argue that data quality plays a major role in cross-field differences, and in particular the poor coverage of the humanities by databases such as Web of Science.

For these reasons, "field-weighted" adjustments to citation counts have, at best, speculative and controversial theoretical support.

The reference class problem

Attempts to adjust bibliometrics for differences across fields also run into a significant theoretical problem, which C. J. Lee (2020) calls "*the reference class problem for credit valuation in science*: to which of the agent's communities—which reference class—should credit valuations be indexed when determining the amount of credit the agent accrues …?" (1029, emphasis in original).

Academic fields are often thought of as clearly-bounded, discrete entities, reflecting the administrative division of universities into schools and departments. But reality is much messier than this: consider research areas such as quantum chemistry (operating across the boundaries between physics and chemistry), bioinformatics (molecular biology — itself biology and chemistry — and computer science), or behavioral economics (psychology and economics). Or indeed "interdisciplines" such as material science or cognitive science. Even within distinct fields, there can be radical differences in research questions, methods, and norms, as in anthropology (often divided into something like biological, cultural, archaeology, linguistics, and primatology), physics (experimental, theoretical, and cosmology), or philosophy (analytic, continental, and philosophy of science).

Because the boundaries within and between academic fields are so porous, a particular research project usually cannot be firmly located within a single research community, and an individual researcher or administrative unit will have a complex portfolio of projects spanning numerous different communities. For example, my Web of Science author profile[6] classifies me as working in History & Philosophy of Science; Philosophy; Public, Environmental & Occupational Health; Environmental Sciences & Ecology; and Science & Technology - Other Topics. This leaves out several areas under which I could be categorized, such as gender studies, science communication, science policy, and statistics. And there are potentially important differences at lower levels of classification; for example, publishing and citation practices between historians and philosophers of science are quite different. In practice, there are perhaps a dozen different fields — and thus *no* field — against which the citations to my research can be normed.

Bibliometrics databases attempt to avoid this problem using multicategory classifications, which are often imprecise. For example, in Scopus' All Science Journal Classification (ASJC)

[6] https://www.webofscience.com/wos/author/record/973913?authorIds=973913&state=%7B%7D

system[7], *Nature Human Behavior* is classified as Psychology: Experimental and Cognitive Psychology, Psychology: Social Psychology, and Neuroscience: Behavioral Neuroscience;[8] this journal also regularly publishes work in archaeology, cognitive science, and other fields across the social sciences. Across the 47,000 journals indexed by Scopus, the median journal has 2 ASJC codes, and more than 2,600 journals have 5 or more codes.

Tradeoffs between data coverage and data quality

A number of services provide bibliometric data. Web of Science and Scopus are commercial services; our campus subscribes to Web of Science. Google Scholar and Semantic Scholar are free. Web of Science and Scopus are not intended to cover every academic journal or book publisher, but instead the "highest impact journals" (https://webofscience.help.clarivate.com/Content/wos-core-collection/wos-core-collection.htm). Semantic Scholar and, especially, Google Scholar aim to be more comprehensive; users of Google Scholar will be familiar with search results including master's theses, conference programs, preprints, grey literature reports, and other documents that don't necessarily count as scholarly publications.

These differences in coverage partially explain why different services provide different values for standard bibliometrics such as publication count, citation count, and the h-index. Table 1 compares results for my author profile across these four platforms and my personally maintained CV.

Table 1: Comparison of publication counts, citation counts, and h-index from my author profile on four database services, and my CV, as of 2025-07-01. Web of Science lists 22 documents in the "core collection," and 2 additional "non-indexed documents." Publications on my CV are divided into "peer-reviewed publications," "other scholarly products" (primarily preprints and book reviews), and "general interest publications." Sources: Google Scholar profile; Scopus profile; Semantic Scholar profile; Web of Science profile; CV.

	publications	citations	h-index
Google Scholar	40	678	14
Scopus	25	309	10
Semantic Scholar	37	352	10
Web of Science	22+2	237	9
CV	23+15+4	-	-

For me, Google Scholar lists nearly twice as many publications as Web of Science, and 2.9 times as many citations. Scopus reports a few more publications than Web of Science (14%),

[7] https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/
[8] https://www.scopus.com/sourceid/21100838541

but 30% more citations. Semantic Scholar has almost as many publications as Google Scholar, but barely half as many citations. None of these include every publication listed on my CV.

The results for my author profile generalize: Martín-Martín et al. (2018) found that Google Scholar reported nearly all citations found by Scopus and Web of Science, along with a large number of other citations, especially from "non-journal sources (48%-65%), including theses, books, conference papers, and unpublished materials" (1). In addition, a companion piece by the same authors, Martín-Martín, Orduna-Malea, and López-Cózar (2018), finds evidence that Web of Science and Scopus inclusion criteria are biased against social sciences and humanities: "a large fraction of highly-cited documents [according to Google Scholar] in the Social Sciences and Humanities (8.6–28.2%) are invisible to Web of Science and Scopus. In the Natural, Life, and Health Sciences the proportion of missing highly-cited documents in Web of Science and Scopus is much lower."

However, the broad coverage of Google Scholar comes with a critical tradeoff: poor data quality. My informal understanding is that Google has never employed professional librarians in the Scholar project, and relies almost entirely on machine learning to parse citations. While authors can edit the metadata on publications on their profile, there is no way to flag erroneous or questionable citations. Simonton (n.d.) chronicles severe metadata errors and erratic changes in citation counts on their Google Scholar profile. In a small study, Sauvayre (2022) found that only two out of 281 purported citations to two source articles did not contain errors. Ibrahim et al. (2024) show that citation counts on Google Scholar can easily be inflated by using an LLM to generate fake papers and posting them on preprint servers.

Semantic Scholar is not as widely-used as Google Scholar, and has not received the same kind of attention from scientometricians as a potential bibliometrics source. However, it also relies heavily on machine learning, may or may not employ librarians, and also has noticeable error rates. For example, while working on this document I realized that Semantic Scholar had conflated *Philosophy of Science* — the flagship journal in the field, published by CUP on behalf of my professional association, the US-based Philosophy of Science Association[9] — with *Philosophia Scientiæ* — a small journal, primarily in French, and published by the Université de Lorraine[10].

All together, to my knowledge there is no source of bibliometrics that can be considered to have both reliable coverage of the humanities and social sciences and reliable citation data.

Gender and racial/ethnic bias

There is a substantial literature examining gender and racial/ethnic bias in bibliometric and other measures of research productivity (e.g., grant application and funding rates). This

[9] https://www.cambridge.org/core/journals/philosophy-of-science
[10] https://journals.openedition.org/philosophiascientiae/635

literature consistently shows that women and racial/ethnic minorities are systematically disadvantaged from resource allocation through the publication process and on to citation rates, across fields. Davies et al. (2021) provide a relatively recent review covering 43 different studies and reviews (pp 3-5). They argue that problems of underrepresentation by gender and race/ethnicity are "self-perpetuating due to reliance on citation metrics, which reflect deeply entrenched biases and exclusionary networks that disadvantage systemically marginalized groups, and these citation metric biases continue to rise globally" (4). Some specific notable findings include:

- "the citation gap between genders was found to be as large as 30% across 13 Science, Technology, Engineering, Mathematics, and Medicine (STEMM) disciplines"
- "women receive more manuscript rejections … are less likely to be published in prestigious journals (which typically have high citation rates) …, and are less likely to be invited to write commentaries"
- "racially and/or ethnically diverse scientific teams … experience more than 5% lower acceptance rates and fewer citations than less diverse author teams"
- "Citational segregation—where authors prefer citing authors from the same racial/ethnic group(s)—has been demonstrated with white authors citing other white authors more frequently"
- "gender and racial citation biases remain stable or have even worsened over the last half century" (Davies et al. 2021, 3–5)

Ginther et al. (2011) found that Black and Asian applicants for NIH funding are significantly less likely to receive awards, even after controlling for factors such as educational background, prior awards, and publication records. They attributed some of this disparity to topic choice, with lower award rates for community and population-level research vs. "basic" or "fundamental" research. Hofstra et al. (2020) find that "underrepresented groups produce higher rates of scientific novelty" but "novel contributions by gender and racial minorities are taken up at lower rates" than majorities. By combining text analysis with metadata and bibliometrics, Hengel (2022) compares preprints to published versions of journal articles, arguing that women authors are held to higher standards than men and that this explains numerous gendered disparities in academic publishing: longer review times, lower acceptance rates, lower publication rates in the most prestigious journals, and lower submission rates. Masters-Waage et al. (2024) find similar traces of bias in promotion and tenure decisions. Looking across fields, Sugimoto and Larivière (2023) argues that "Disciplines with parity (or those dominated by women) tend to be lower cited and relegated to the bottom tiers of the academic hierarchy."

An important proviso with virtually all of these studies is that they use automated demographic imputation, that is, imputing authors' gender and race/ethnicity based on their names. Lockhart, King, and Munsch (2023) shows that these methods are less accurate for both gender and racial/ethnicity minorities, which generally results in misclassifying members of minority groups as members of majorities. In general, measurement error on the independent variable produces *regression dilution*, in which estimated correlations/regression coefficients are biased

towards 0. That is, all else being equal, automated demographic imputation would be expected to yield underestimations of the magnitude of gender and racial/ethnic bias.

Quantification as depoliticization

Thus far, I have offered "technical" critiques of bibliometrics, showing limitations in their fit for purpose as measures of research quality. But it is important to recognize that, in the context of decisions such as how to allocate faculty lines, bibliometrics also have a broader function, as what we might call a *depoliticization strategy*.

STS scholars argue that quantification or *scientization* — replacing explicitly political decision-making processes with expert or "data-driven" processes — is a key legitimizing strategy used by modern states and other political and social authorities (Jasanoff 1990; Porter 1995; Sarewitz 2000). This strategy was on display in 2020, when political leaders presented themselves as "following the science." Even Covid skeptics made their arguments in terms of quantified evidence and data visualizations (C. Lee et al. 2021). On an individual level, many students in my Critical Thinking course come in thinking that using numbers is enough to make an argument valid — without regard to how those numbers were produced and whether they have anything to do with the conclusion the argument is attempting to support.

Broadly, in our society quantified evidence is treated as extra-political — existing outside or beyond (political) controversy — and therefore super-political — taking precedence over and even shutting down (political) controversy. Quantification is thereby appealing when political controversies seem to be heated and intractable, appearing to give us (that is, whoever has the power to impose a particular system of quantification) a way to end debate and resolve the issue, moving it outside or beyond politics. Thus, "depoliticization."

Allocating extremely scarce resources such as faculty lines is already a touchy issue. Doing so under conditions of existential uncertainty for American higher ed, impending demographic decline, and pre-existing structural deficits only makes the controversy more fraught. In this context, bibliometrics appear to provide an "objective," depoliticizing solution.

However, the "technical" critiques I have offered show that the promise of using bibliometrics to depoliticize decisions about faculty lines is a mirage. Any attempt to operationalize bibliometrics for decisionmaking requires prior decisions, about how to compare incommensurable research outputs, how to assign departmental research portfolios to fields and adjust for differences between fields, which source(s) of bibliometrics data to use and how to manage the corresponding tradeoffs between coverage and data quality, and how to mitigate rather than exacerbate the influence of deep structural biases. Rather than depoliticizing, bibliometrics would simply become a new arena for political controversy (Daniel J. Hicks 2017; D. J. Hicks 2018), as different interests across campus push for different ways of resolving these prior decisions that happen to benefit their particular group.

Recommendations

Bibliometrics are neither fit for purpose as measures of research quality nor strategically useful for cooling heated controversies. As an alternative, I would recommend adopting a process with the following features, based on general best practices in research assessment/evaluation and points from the Leiden Manifesto in particular:

1. Start by articulating aims

   Faculty hires can serve many different kinds of goals, both inward-facing (scholarly) and outward-facing (social or community): growing (or maintaining) enrollments, developing a unit's existing research strengths, broadening a unit's coverage, fostering interdisciplinary research, attracting external funding, addressing a significant social issue, supporting a certain local community, and so on. The campus should develop specific versions of these or other goals, as the primary basis on which hiring decisions will be made.

2. Approach hiring requests as program impact plans

   Program impact plans — also known as theories of change and logic models — describe how an organization believes that a certain course of action does (or will) lead to certain outputs, and thereby downstream outcomes and impacts that ultimately promote the organization's aims. For example, given an aim of strengthening ties between engineers and social scientists in robotics research, a NSF funding program might devote 20% of its funds to a solicitation that requires two PIs, one from each area. The program impact plan would identify, as an output, more funding to interdisciplinary teams; as outcomes, funded researchers publishing in a wider variety of journals than they would have otherwise and an increased presence of social scientists at (engineering-oriented) robotics conferences; and as an impact, incorporation of social science research findings into engineering research.

   Departments wishing to hire faculty could take a similar approach, preparing short statements (say, capped at 1,000 words) that explain, for example, how a new hire in a particular area would enable them to add an attractive emphasis track to their major and regularly offer a popular elective course for another major (outputs); how this would make the campus more competitive for potential students in both majors (outcomes); and thereby promote enrollment growth (or counteract trends in declining enrollment; impacts).

   These program impact plans should follow the warning of the Leiden Manifesto, and avoid misplaced concreteness and false precision: quantitative enrollment estimates are much more uncertain than they might appear to be. It is also important to recognize the possibility of bias in qualitative assessment; for example, STEM majors as such are not necessarily more popular with students than humanities majors. Nonetheless, both

faculty and campus administrators can assess whether a given hiring proposal might plausibly lead to the anticipated outputs, outcomes, and impacts.

3. Faculty provide peer review

   While it is impossible to completely depoliticize high-stakes resource allocations decisions, it is worth considering peer review — especially for research funding — as a familiar model of political decisionmaking that does not often erupt into intractable, heated controversy. I have seen firsthand how SSHA's Curriculum Committee throughly but constructively scrutinizes proposals for new majors. UC Merced might take a similar approach to hiring faculty: departments could prepare hiring proposals — in the form of program impact plans — which would be reviewed by school Executive Committees, and then by CAPRA and/or other Senate committees. Reviewing committees could provide both qualitative evaluations and rubric-based assessments of each proposal, along the lines of NSF and NIH.

   While far from perfect, distributing power over faculty hires and incorporating accountability/transparency would be preferable to the current system, which often relies heavily on patronage relations among individual faculty, deans, and the provost.

References

Cagan, Ross. 2013. "The San Francisco Declaration on Research Assessment." *Disease Models & Mechanisms* 6 (4): 869–70. https://doi.org/10.1242/dmm.012955.

Davies, Sarah W., Hollie M. Putnam, Tracy Ainsworth, Julia K. Baum, Colleen B. Bove, Sarah C. Crosby, Isabelle M. Côté, et al. 2021. "Promoting Inclusive Metrics of Success and Impact to Dismantle a Discriminatory Reward System in Science." *PLOS Biology* 19 (6): e3001282. https://doi.org/10.1371/journal.pbio.3001282.

Funk, Russell J., and Jason Owen-Smith. 2016. "A Dynamic Network Measure of Technological Change." *Management Science* 63 (3): 791–817. https://doi.org/10.1287/mnsc.2015.2366.

Ginther, Donna K., Walter T. Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L. Haak, and Raynard Kington. 2011. "Race, Ethnicity, and NIH Research Awards." *Science* 333 (6045): 1015–19. https://doi.org/10.1126/science.1196783.

Golosovsky, Michael. 2021. "Universality of Citation Distributions: A New Understanding." *Quantitative Science Studies* 2 (2): 527–43. https://doi.org/10.1162/qss_a_00127.

He, Jiangen, and Chaomei Chen. 2018. "Predictive Effects of Novelty Measured by Temporal Embeddings on the Growth of Scientific Literature." *Frontiers in Research Metrics and Analytics* 3. https://doi.org/10.3389/frma.2018.00009.

Hengel, Erin. 2022. "Are Women Held to Higher Standards? Evidence from Peer Review." *The Economic Journal*, May, ueac032. https://doi.org/10.1093/ej/ueac032.

Hicks, D. J. 2018. "The Safety of Autonomous Vehicles: Lessons from Philosophy of Science." *IEEE Technology and Society Magazine* 37 (1): 62–69. https://doi.org/10.1109/MTS.2018.2795123.

Hicks, Daniel J. 2016. "Bibliometrics for Social Validation." *PLOS ONE* 11 (12): e0168597. https://doi.org/10.1371/journal.pone.0168597.

———. 2017. "Scientific Controversies as Proxy Politics." *Issues in Science and Technology*, January 2017. https://www.jstor.org/stable/24891967.

———. 2021. "Productivity and Interdisciplinary Impacts of Organized Research Units." *Quantitative Science Studies* 2 (3): 990–1022. https://doi.org/10.1162/qss_a_00150.

Hicks, Daniel J., David A. Coil, Carl G. Stahmer, and Jonathan A. Eisen. 2019. "Network Analysis to Evaluate the Impact of Research Funding on Research Community Consolidation." *PLOS ONE* 14 (6): e0218273. https://doi.org/10.1371/journal.pone.0218273.

Hicks, Daniel J., Carl Stahmer, and MacKenzie Smith. 2018. "Impacting Capabilities: A Conceptual Framework for the Social Value of Research." *Frontiers in Research Metrics and Analytics*. https://doi.org/10.3389/frma.2018.00024.

Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. 2015. "The Leiden Manifesto for Research Metrics." *Nature* 520 (7548): 429. https://doi.org/10.1038/520429a.

Hicks, D., and R. Simmons. 2019. "The National Robotics Initiative: A Five-Year Retrospective." *IEEE Robotics Automation Magazine* 26 (3): 2–9. https://doi.org/10.1109/MRA.2019.2912860.

Hofstra, Bas, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. 2020. "The Diversity–Innovation Paradox in Science." *Proceedings of the National Academy of Sciences*, April, 201915378. https://doi.org/10.1073/pnas.1915378117.

Ibrahim, Hazem, Fengyuan Liu, Yasir Zaki, and Talal Rahwan. 2024. "Google Scholar Is Manipulatable." arXiv. https://doi.org/10.48550/arXiv.2402.04607.

Jasanoff, Sheila. 1990. *The Fifth Branch: Science Advisers as Policymakers*. Harvard University Press.

Kojaku, Sadamori, Giacomo Livan, and Naoki Masuda. 2021. "Detecting Anomalous Citation Groups in Journal Networks." *Scientific Reports* 11 (1): 14524. https://doi.org/10.1038/s41598-021-93572-3.

Lee, Carole J. 2020. "The Reference Class Problem for Credit Valuation in Science." *Philosophy of Science* 87 (5): 1026–36. https://doi.org/10.1086/710615.

Lee, Crystal, Tanya Yang, Gabrielle Inchoco, Graham M. Jones, and Arvind Satyanarayan. 2021. "Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online." *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May, 1–18. https://doi.org/10.1145/3411764.3445211.

Lockhart, Jeffrey W., Molly M. King, and Christin Munsch. 2023. "Name-Based Demographic Inference and the Unequal Distribution of Misrecognition." *Nature Human Behaviour*, April, 1–12. https://doi.org/10.1038/s41562-023-01587-9.

Martín-Martín, Alberto, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. 2018. "Coverage of Highly-Cited Documents in Google Scholar, Web of Science, and Scopus:

A Multidisciplinary Comparison." *Scientometrics*, June, 1–14. https://doi.org/10.1007/s11192-018-2820-9.

Martín-Martín, Alberto, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado López-Cózar. 2018. "Google Scholar, Web of Science, and Scopus: A Systematic Comparison of Citations in 252 Subject Categories." *Journal of Informetrics* 12 (4): 1160–77. https://doi.org/10.1016/j.joi.2018.09.002.

Marx, Werner, and Lutz Bornmann. 2015. "On the Causes of Subject-Specific Citation Rates in Web of Science." *Scientometrics* 102 (2): 1823–27. https://doi.org/10.1007/s11192-014-1499-9.

Masters-Waage, Theodore, Christiane Spitzmueller, Ebenezer Edema-Sillo, Ally St. Aubin, Michelle Penn-Marshall, Erika Henderson, Peggy Lindner, Cynthia Werner, Tracey Rizzuto, and Juan Madera. 2024. "Underrepresented Minority Faculty in the USA Face a Double Standard in Promotion and Tenure Decisions." *Nature Human Behaviour*, October, 1–12. https://doi.org/10.1038/s41562-024-01977-7.

Petersen, Alexander M. 2022. "Evolution of Biomedical Innovation Quantified via Billions of Distinct Article-Level MeSH Keyword Combinations." *Advances in Complex Systems* 25 (1). https://doi.org/10.1142/S0219525921500168.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, N.J: Princeton University Press.

Radicchi, Filippo, Santo Fortunato, and Claudio Castellano. 2008. "Universality of Citation Distributions: Toward an Objective Measure of Scientific Impact." *Proceedings of the National Academy of Sciences* 105 (45): 17268–72. https://doi.org/10.1073/pnas.0806977105.

Sarewitz, Daniel. 2000. "Science and Environmental Policy: An Excess of Objectivity." In *Earth Matters: The Earth Sciences, Philosophy, and the Claims of Community*, edited by Robert Frodeman, 79–98. Prentice Hall. http://www.cspo.org/_old_ourlibrary/ScienceandEnvironmentalPolicy.htm.

Sauvayre, Romy. 2022. "Types of Errors Hiding in Google Scholar Data." *Journal of Medical Internet Research* 24 (5): e28354. https://doi.org/10.2196/28354.

Simonton, Dean Keith. n.d. "Google Scholar Citations: Serious Errors." Accessed July 1, 2025. https://simonton.faculty.ucdavis.edu/research/google-scholar-citations-serious-errors/.

Sugimoto, Cassidy R., and Vincent Larivière. 2023. *Equity for Women in Science: Dismantling Systemic Barriers to Advancement*. Harvard University Press.

Zeng, Xiao Han T., Jordi Duch, Marta Sales-Pardo, João A. G. Moreira, Filippo Radicchi, Haroldo V. Ribeiro, Teresa K. Woodruff, and Luís A. Nunes Amaral. 2016. "Differences in Collaboration Patterns Across Discipline, Career Stage, and Gender." *PLOS Biology* 14 (11): e1002573. https://doi.org/10.1371/journal.pbio.1002573.