2020

Impact Metrics

John Gerring
University of Texas at Austin, jgerring@austin.utexas.edu

Sebastian Karcher
Syracuse University, skarcher@syr.edu

Brendan Apfeld
University of Texas at Austin, brendan.apfeld@utexas.edu

Recommended Citation

Gerring, John, Sebastian Karcher, and Brendan Apfeld. 2020. "Impact Metrics." In The Production of Knowledge: Enhancing Progress in Social Science, edited by Colin Elman, James Mahoney, and John Gerring, 371–400. Strategies for Social Inquiry. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108762519.015

15.Impact Metrics

John Gerring
Brendan Apfeld
Sebastian Karcher

Published as
Gerring, John, Sebastian Karcher, and Brendan Apfeld. 2020. "Impact Metrics." In *The Production of Knowledge: Enhancing Progress in Social Science*, edited by Colin Elman, James Mahoney, and John Gerring, 371–400. Strategies for Social Inquiry. Cambridge: Cambridge University Press.
https://doi.org/10.1017/9781108762519.015

Virtually every evaluative task in the academy involves some sort of metric (Elkana et al. 1978; Espeland & Sauder 2016; Gingras 2016; Hix 2004; Jensenius et al. 2018; Muller 2018; Osterloh and Frey 2015; Todeschini & Baccini 2016; Van Noorden 2010; Wilsdon et al. 2015). One can decry this development, and inveigh against its abuses and its over-use (as many of the foregoing studies do). Yet, without metrics, we would be at pains to render judgments about scholars, published papers, applications (for grants, fellowships, and conferences), journals, academic presses, departments, universities, or subfields.

Of course, we also undertake to judge these issues ourselves through a deliberative process that involves reading the work under evaluation. This is the traditional approach of peer review. No one would advocate a system of evaluation that is entirely metric-driven. Even so, reading is time-consuming and inherently subjective; it is, after all, the opinion of one reader (or several readers, if there is a panel of reviewers). It is also impossible to systematically compare these judgments. To be sure, one might also read, and assess, the work of other scholars, but this does not provide a systematic basis for comparison – unless, that is, a standard metric(s) of comparison is employed. Finally, judging scholars through peer review becomes logistically intractable when the task shifts from a single scholar to a large group of scholars or a large body of work, e.g., a journal, a department, a university, a subfield, or a discipline. It is impossible to read, and assess, a library of work.

For these reasons, quantitative metrics are an unavoidable component of gatekeeping in the academy. It is not a question of whether or not to employ metrics but rather which metrics to employ. We need to think carefully about what these evaluation metrics are and how they might, in turn, establish incentives for scholars, presses, departments, and universities. The metrics that gain traction in the social sciences today will influence the production of social science knowledge tomorrow.

Traditional metrics of quality are grounded in judgments of *reputation*. Sometimes, reputation is measured in a systematic fashion through surveys of professionals in a field. Other times, they are loosely understood norms, which everyone is assumed to share. In either case, journals, presses, and departments can be ranked by considering their overall reputation. At the top of a ranking, there is likely to be a high level of agreement. Toward the middle and bottom of the scale, however,

disagreement is more likely. Here, an evaluative system based on reputation breaks down, for middle- and lower-ranked units do not have clearly defined reputations. Another problem with reputational metrics is that elements under review are highly inter-dependent. It is difficult to separate the reputation of articles (or books) from the journals (or presses) they are published in, or the reputation of scholars from the departments and universities they are employed by. The reputation of institutions tend to overshadow the reputations of studies and scholars. Moreover, these reputations are sticky, and may reflect past glories more than current performance. A pecking order based solely on reputation is likely to perpetuate itself because there is no other metric by which performance can be judged. Harvard is tops not because of anything it does but because it is Harvard. Not only is this un-meritocratic, it also has a baleful effect on incentives – both for those at the top of the pecking order and for those at the bottom. No one has an incentive to improve their game if their position in the hierarchy is primarily a product of their title and institutional affiliation, especially if these factors are locked in (post-tenure).

In recent decades, another metric has come to the fore that takes a fundamentally different approach to evaluation. This approach measures *impact*, inferred from the citations accrued by a study.[1] Like reputation, impact may be measured at any level. An individual may be evaluated according to the citations accruing to her publications; a department may be evaluated according to the citations accrued by studies produced by all its faculty; a journal or press may be evaluated according to the citations accrued by all its publications; and so forth. Naturally, there is some inter-dependence among impact measures. A study's citation-count is apt to be influenced by a journal's citation-count, for example. Nonetheless, there is a much greater degree of independence among the units being assessed. An implication is that impact-based metrics are more meritocratic than reputational metrics.

We are cautiously optimistic about the increasing use of such metrics. However, this new standard of evaluation introduces several problems – as well as an entirely new set of incentives – that have not been widely appreciated. We begin by reviewing the shortcomings of current citation databases, which provide the basis for impact metrics and place limitations on the information contained in them. We view these problems as fundamental, but also temporary. In light of recent advances in information technology and library science it seems likely that these shortcomings will be overcome, perhaps in the near future. Next, we address potential objections to a citation-based system of evaluation. We argue that these objections are resolvable, for the most part. We conclude with thoughts on the possible impact of impact metrics.

Databases

Citation-based metrics have been around for a long while in the natural sciences (Garfield 1955). In the social sciences, they have come to the fore more recently – largely, it seems, because of the ready availability of citation data, which is now stored electronically and easily accessed on the web. The most widely used databases are Google Scholar (GS), Scopus, Web of Science (WoS), and more recently, Microsoft Academic (MA)[2]. Each has its specific limitations, but all struggle to achieve

[1] For wide-ranging discussions – many of them highly critical of this development – see Bornmann & Daniel (2008), Cronin & Sugimoto (2014), Gingras (2016), Hamermesh (2018), Sugimoto and Larivière (2018), Todeschini & Baccini (2016).

[2] A fairly recent initiative, the Open Citation Corpus (http://opencitations.net/), has the potential to provide a highly valuable, and completely open alternative. Currently its size and coverage – less than 1 percent of the number of publications in WoS – is too limited to warrant inclusion. Yet another database, dimensions.ai, published by the Digital Science group, which is co-owned with Springer/Nature, appears to provide similar services to Scopus/WoS and is

three essential goals: (a) correct identification of authors, (b) complete – or at least representative – coverage, and (c) open access.

Author identity

Publications must be linked correctly with their authors if citation metrics are to be used as an evaluative tool. A single junior faculty member with only a handful of publications may be able to track the citations of those pieces with manual searches of each one. But automation becomes essential when her publications grow in number or when one needs a systematic comparison of scholars across a department, university, subfield, or discipline. Impact metrics are useful if citations can be quickly aggregated by author, which in turn implies that they must be able to link publications with their authors, with a minimum of error.

Several attempts have been made to standardize author identification, most notably the International Standard Name Identifier (ISNI, for any creative work) and ORCID (specifically for academic work). However, their coverage at this time is too low to be of genuine use in mapping authors to works. We are, however, optimistic about the potential for ORCID, now commonly included in citation data for new works, to help solve this issue in the future. Importantly, both Scopus and WoS collaborate with the ORCID initiative.

Given the shortcomings (in coverage) of ISNI and ORCID, each of the existing databases has been obliged to create its own method for identifying authors.

GS links documents together by author only when a researcher intentionally creates a profile on the site. Once created, GS will attempt to identify new publications to add to the profile. However, the process is not flawless and requires authors to regularly "curate" their profile, confirming correct publications, deleting incorrect publications, and combining duplicates as appropriate. The most notorious – but far from only – example of this is the world's most prolific and most widely cited author "et al."[3] Moreover, many scholars have not created a GS profile and these absences are probably not random. As a result, GS profiles do not provide an unbiased measure of impact across groups of scholars (subfields, fields, departments, etc.).

Scopus generates scholar profiles automatically, creating a complete dataset but also one with greater opportunities for both false positives and false negatives. Scopus provides a way for authors to request corrections to their profiles. But we wonder whether any author would consider the benefits of corrections to outweigh even the minor costs of doing so – especially if the error was in the author's favor.

WoS does not create researcher profiles of the same kind as GS or Scopus. This service attempts to group publications by the same researcher together, but does not automatically create any citation statistics based on those publications. It is possible for users to generate such statistics with granular control over what is and is not included in these calculations, but additional steps are required and it results only in a temporary report.

MA also offers user profiles akin to those in GS. Unlike GS, however, these profiles are created automatically and researchers then have the option to "claim" and make revisions to their own profile. MA profiles are less detailed than GS ones, offering only article and total citation counts. MA offers direct access to the underlying data (discussed in greater detail below) through Microsoft Academic Graph. This provides researchers with great flexibility, but also exposes directly the core problem entailed in automatically linking documents to author identifiers.

---

currently free of charge. Its launch was too late for inclusion in this article, but it may be another useful source for researchers.

[3] Their google scholar profile is available at https://scholar.google.nl/citations?user=qGuYgMsAAAAJ&hl=en and its backstory at http://ideophone.org/some-things-you-need-to-know-about-google-scholar/

While author identification is a particularly vexing problem for large-scale analysis, it can be solved with reasonable amounts of manual labor for medium-*n* analyses. Identifying false positives (i.e., studies falsely attributed to an author) is typically quick. Papers *missing* for a particular author are harder to identify, but the omission of some works is unlikely to cause significant measurement error: the citation counts of the papers for authors typically follows a power-law distribution (see e.g. Breuer and Bowen 2014; Brzezinski 2015), so that an accurate assessment of their total citation counts depends mainly on the inclusion of their most widely cited works, which (with some notable exemptions, as noted below) is typically the case.

Coverage

In addition to linking documents to their authors, databases must also contain a large portion of the scientific record. However, "larger" is not necessarily better. The coverage needs also to be representative of the scientific record. If one portion of the record is missing from a database then all metrics generated by that database will be biased in a systematic fashion. Finally, a database needs to avoid double counts (e.g. counting citations from a preprint and an identical journal article as if they were separately published documents).

Unfortunately, extant databases are limited in coverage and struggle with representativeness and double-counting. Moreover, there appears to be a trade-off between these goals. GS and MA have the largest corpus and do somewhat better on the inclusion of non-article items, but the quality of their data is lower. Scopus and WoS perform much better on data quality, but have poor coverage outside of traditional journal articles.

GS is the most extensive database, estimated around 140 million entries in 2014 (Sugimoto and Larivière 2018) and growing steadily. Its inclusivity with respect to contemporary sources might be regarded as a blessing – it includes working papers, papers posted on personal web sites, along with papers published in recognized journals. However, as for author identification, the quality of data in GS is so low that we do not believe it should be used without extensive manual curation. Many of these are "honest" errors, but the wide range of sources included in GS together with the lack of curation also offers an opportunity for malfeasance as demonstrated by Delgado López-Cózar, Robinson-García, and Torres-Salinas (2014), who were able to massively boost the *h*-index of every member of their working group by uploading 6 fake papers to a university website.

MA is the newest citation database of the four, but its growth has been rapid. Recent estimates suggest that MA grew from 83 million records in 2015 to 140 million in 2016, with current coverage equal to or exceeding that of WoS and Scopus (Hug, Ochsner, and Brändle 2017; Wade et al. 2016). Today, the MA frontpage suggests this number has continued to increase and now stands at over 173 million publications. However, coverage by subject matter varies greatly. Estimates range from less than 10% coverage in the humanities to over 90% in engineering and technology (Hug and Brändle 2017). MA suffers from quality issues akin to GS. For example, in using MA data to identify high impact papers, Wesley-Smith, Bergstrom, and West (2016) found that among their four highest impact articles was one duplicate, one reference to an entire journal (the *New England Journal of Medicine*) and one "In the press."

Scopus (ca. 70 million articles) and WoS (ca. 59 million articles) are significantly smaller than fully automated MA and GS, but both curate their sources, i.e. they select a subset of journals based on a specified set of standards such as peer review and ethical publication practices.[4] They argue that "the core literature for all scholarly disciplines may be concentrated in a relatively small number of journals" (Testa 2016) and that focusing on this core provides better representativeness by excluding

[4] See Testa (2016) for an in-depth discussion of the WoS journal selection process and https://www.elsevier.com/solutions/scopus/content for a summary of Scopus's practices.

meaningless citations. Curated indices are harder (though not impossible) to game. They also can more easily react to evidence of fraud by removing offending journals and/or authors from their indices. On the downside, the focus on formal publications introduces a significant lag. In some disciplines, the average delay between submission and acceptance can be as long as 30 months (see Ellison 2002, 951, for economics), i.e. citations *to* journal articles *in* journal articles can lag as much as five years.

While GS and MA cover more books than the article-focused Scopus and WoS, no database offers good coverage of books (Samuels 2013), and edited volumes constitute a particular source of confusion as editorship and authorship are often incorrectly noted in databases. Since books (including edited volumes) continue to comprise an important genre of academic production in the social sciences, this must be regarded as a major failing. Moreover, it penalizes certain sorts of work, namely, work that is broad in scope and work that employs qualitative methods, which generally require more space to articulate. Consequently, extant databases are systematically biased, giving an advantage to scholars who publish in journals, or whose work is more likely to be cited in journals.

Book production varies by subfield and by author — some subfields tend to share their scholarship in article form at greater rates than others within the same discipline and some authors tend to favor one type of publication over another. Thus even if book missing-ness was distributed in the same way as article missing-ness, there would still be bias in the citation counts.[5]

Gingras (2016) points out that while these databases do not include citations *within* books, they do contain citations *of* books (when books are cited by articles). Unfortunately, this does not overcome the problem. Note, first, that book and book section missing-ness is extremely high. Hug and Brändle (2017) estimate that MA, Scopus, and WoS contain, respectively, 14.3%, 9.7%, and 6.6% of book sections. Coverage of edited volumes is slightly better for MA (15.6%) but drops to a mere 2.6% and 3.9%, respectively for Scopus and WoS. These numbers are shockingly low compared with estimated coverage for articles in these databases — 73.0%, 77.5%, and 71.7%, respectively.

Book missing-ness is a problem for book authors who receive lower citation counts if their works are missing. The exclusion of citations within books affects all authors whose work is cited in these texts. This would not be a problem if citations within articles and books were highly correlated. However, books and articles do not cite the same sorts of studies. Authors' citation counts within articles is not highly correlated with their citation counts within books (Hicks 1999). The omission of book citations thus biases impact metrics in important ways.

Access

Given the importance of metrics, it is vital that whatever database(s) becomes the basis for constructing impact metrics for a discipline be open-access. Only in this fashion will full transparency be possible. And only in this fashion will it be possible to devise a variety of metrics to suit different purposes and concerns, as discussed in the next section. It is worrisome that the databases that provide fodder for scholarly metrics are controlled by privately held companies – Alphabet (the parent company of Google), Elsevier (the parent company of Scopus), Clarivate Analytics (the parent company of WoS) and Microsoft (the parent company of MA) (Weingart 2005).

Moreover, the business of constructing metrics has been taken over by a host of private consultancies such as Academic Analytics, iFQ, Sciencemetrix and CWTS. These companies sell their services to universities, governments, and other organizations and view their product as

[5] Note that we think it very unlikely that the distributions are the same given the different coverage rates across disciplines but that current studies have not addressed this question directly.

proprietary, which means that they make it available to administrators but not – except by special request – to researchers. It appears that these metrics are already affecting decisions about hiring, promotion, and funding without the full knowledge of stakeholders, i.e., researchers whose livelihoods and research is at issue. It is largely a behind-the-scenes operation. Equally worrisome, there are good reasons to worry about the quality of the data, and the metrics, that are driving these decisions (Basken 2018).

Of the databases considered here, access varies from completely open to both closed and behind a pay wall. Scopus and WoS are both subscription services, generally subscribed to by institutional libraries.[6] The cost is non-trivial; consequently, most universities choose to purchase access to one, but not both. This means that researchers with university connections do not generally have access to the same databases. Both services provide an API for which subscribers can request a key to access the database programmatically. Scopus provides a Python module so that users with a rudimentary knowledge of that language can run searches.

MA requires no subscription and an archived version of the underlying data is, for the time being, also publicly available. For those wishing to access the current version of the data, Microsoft provides tools to access the graph in several programming languages using the Microsoft Knowledge API. Open Academic Graph provides access to the underlying data (alongside the data for a similar but smaller citation graph, AMiner). We applaud this openness and recognize that it offers researchers opportunities unavailable with the other databases. However, we are concerned that this open-access policy could be changed at any time.

GS is also a free service. It does not, however, support any kind of programmatic or automated data access or collection. Some unofficial and unsupported tools for automatic comparisons exist (e.g. Publish or Perish [Harzing 2007]), but their functionality is limited and attempts to use them in more expansive ways will quickly hit Google's (unpublished and unknown) rate limit.[7]

Conclusions

In this section, we have discussed limitations imposed by the databases that impact metrics draw upon. These limitations are considerable. Databases have not arrived at a consistent and universal identifier for authors. They are limited in coverage, especially with respect to books. And most are not open-access. Impact metrics cannot do the job envisioned for them until these obstacles are overcome.

Our guess is that most, if not all, of these limitations will be overcome in the coming years. Citation databases are evolving quickly, and one major player – MA – has only recently entered the field and other databases like *dimensions.ai* and OCC have recently appeared. The technical barriers to solving these problems are not insuperable (though there will doubtless remain some degree of error). And the payoff to the outfit that solves these issues is vast. Since there is already competition among four leading companies – each with vast resources – we assume that one of them will see fit to provide what the academy so desperately needs. Even so, impact metrics raise other issues, which may be considered intrinsic to the enterprise and hence more fundamental in nature.

[6] Scopus also provides free access to a limited set of its features.

[7] Publish or Perish now pulls data from both GS and MA, pointing to the relatively greater openness of the newer database.

Objections

The interpretation of "impact" based on citations rests upon assumptions about why scholars cite each other. Robert Merton (1968: 622) writes,

> The reference serves both instrumental and symbolic functions in the transmission and enlargement of knowledge. Instrumentally, it tells us of work we may not have known before, some of which may hold further interest for us; symbolically, it registers in the enduring archives the intellectual property of the acknowledged source by providing a pellet of peer recognition of the knowledge claim, accepted or expressly rejected, that was made in that source.

Implicitly, a citation is an acknowledgment of influence.[8] Accordingly, the cumulative citation count of a study, an individual, a journal, a department, a university, or any other unit of academic production is also a measure of influence, or impact.

However, using impact metrics to evaluate research and researchers is not without problems. In this section we review a series of objections to impact metrics: (a) types of citations, (b) non-citations, (c) fake citations, gaming, and citation quality, (d) genre bias, (e) non-equivalence across contexts, and (i) gender bias.

Types of citations

Not all citations are created equal. Some citations praise the cited paper, others point out fatal errors in it. Some citations stand alone, others engage deeply with the cited article. Some citations are a self-flattery or a courtesy to colleagues or friends, while others carry the influence of a work into an entirely new discipline. Citation counts grant all of these citations equal weight.

(A promising research agenda offers to categorize every citation using an ontology (see e.g. Shotton 2010). However, such categorizations currently rely largely on manual annotation and the chances of it reaching sufficient coverage for impact measurement in the near future are small.)

"Negative" citations seem particularly troublesome. Why should an author get credited for an article criticizing their work? At a closer look, the problem may be less critical than it first appears. For one, scientific progress often occurs in a dialectical fashion, with new work criticizing old work. Sometimes, the most important study in a field is wrong in its findings but nonetheless scopes out a new area of research, perhaps by posing a new question or a new method of exploring a question. We should recognize these sorts of contributions, and impact metrics provide a convenient way of doing so. A recent study finds evidence to support this optimism, noting that "negative citations concerned higher-quality papers, were focused on a study's findings rather than theories or methods" (Catalini, Lacetera, and Oettl 2015).

Similar concerns have been raised about "obligatory" citations, citations that come to dominate a topic even though other, less famous, studies are equally good. This is sometimes referred to as a bandwagon, herding, or "Matthew" effect (Hamermesh 2018, 129; Merton 1968). In some instances, the obligatory citation is to a work that is notably flawed. As with negative citations, this seems to call into question the value of a citation as a measure genuine scientific impact. Nonetheless, an important function of citations is as a placeholder, reminding the reader of a tradition of research or a particular position in a long-standing debate. Classics serve that function, even if they are no longer at the forefront of research (almost by definition, they are not). If a study

[8] For further discussion of the various functions and motivations underlying citations, see Bornmann and Daniel (2008), Cronin (1984), Erikson and Erlandson (2014), Nicolaisen (2007).

has managed to formulate a position in an especially persuasive fashion, and has come to be acknowledged as a classic, then it is performing a valued function in knowledge production.

Another issue with citation-based metrics is their origin. A citation that highlights one's own work, a "self-citation" (Hudson 2007), may be regarded as a stroke of vanity rather than evidence of impact. However, it is important to recognize that authors sometimes self-cite for perfectly good reasons (Gingras 2016, 24). After all, most knowledge comes from specialization, and this means that an advanced researcher will have honed a small body of knowledge over a long period of time. A recent study of citation patterns in economics seems to endorse this benign view of self-citation. Hamermesh (2018, 129) reports that self-citations and other-citations are highly correlated, meaning that authors who cite themselves are also likely to be cited by others. There are some known cases of abuse[9] (more on this below), but in the case of self-citation, these are easily mitigated through a change in metric, i.e. by discarding self-citations from citation-based metrics.

A related but thornier issue concerns author networks that include significant levels of self-citation. Authors are more likely to cite work that they are familiar with, and they are more likely to be familiar with work conducted by people within their network of friends and associates (Clements and Wang 2003). Of course, everyone has a network, and if these networks are of roughly equal size and status one would expect that network effects to cancel each other out. A recent study of economists shows that *relative* citation counts were unaffected when stripped of citations by coauthors and those judged to be within a researcher's network (Orazbayev 2017; discussed in Hamermesh 2018, 123). Networks become problematic where they are not equally available to all scholars or, worse, where they are put to strategic use as "citation cartels." We address these concerns in separate sections below.

Non-citations

A related problem is the *non*-citation, where previous work on a topic is not acknowledged, or not fully acknowledged (i.e., with a formal citation of the sort that might be counted in a citation database). This may occur (a) if there is insufficient space in a journal article to record the provenance of an idea, (b) if an author is trying to emphasize the originality of her own work (by omitting reference to previous work on a subject), (c) if previous work is judged inferior or outdated (and hence not necessary to cite), or (d) if previous work is judged so well-established that it now counts as factual and therefore need not be cited (aka obliteration by incorporation [Merton 1968: 622]).

Bibliometric research on this subject is not sufficiently advanced to discern the relative frequency of these four explanations for non-citations. Evidently, they have very different implications, and impact metrics do not currently provide any means of sorting them out, though it is possible that the development of artificial intelligence will provide an avenue for doing so in the not-too-distant future.

We trust that as impact metrics gain in prominence, authors, reviewers, and editors will come to regard citations as crucial aspects of a research publication, according it the space and seriousness which it deserves, subjecting non-citation to scrutiny, and thus mitigating the problem. We ought to encourage scholars to do a better job in constructing their literature reviews and a more complete job of citing relevant work. An essential component of that encouragement is granting authors sufficient space to cover the field (see Chapter 5).

[9] In a case that received significant attention recently, the editor of the high-impact *Perspectives on Psychological Science,* Robert Sternberg, cited his own papers 161 times (46 percent of all citations) in 7 articles published in the journal *under his own editorship* (Fried 2018).

Fake citations, gaming, and citation quality

A potential problem with GS and any other free-range citation database is that it does not distinguish the source of a citation. Anything defined vaguely as "scholarly" is in the catchment area, including papers posted on an author's own web site and working papers posted on other sites that may be unmonitored, or at any rate not subject to peer review. A citation is a citation is a citation.

This is unhelpful, at best, and an invitation to abuse, at worst. Note that authors can, in principle, boost their GS citation counts by posting dozens – or even hundreds – of papers that cite their work (Delgado López-Cózar, Robinson-García, and Torres-Salinas 2014; Gingras 2016: Kindle Location 1739). As noted above, databases that are open only to a curated set of formal publications (e.g., Scopus, WoS) do not face this issue. Where automated databases provide full datasets, such as MA, it is possible to restrict the citation sources included in citation counts to a subset of publications, ameliorating such problems.

As the importance of metrics to scholarly careers increases, so will attempts to "game" the system (Campbell 1979; Muller 2018). That is, scholars will try to achieve the specified target but not in a way that achieves the underlying goals that the target was intended to encourage. "Teaching to the test," is an example of this sort in primary and secondary school systems; but there are many others (Lazear 2006).

Gaming is likely to arise with respect to any metric that is used for purposes of evaluation. If the metric is quantity (number of published papers), canny scholars will divide up their publishing efforts into "minimal publishing units" or re-publish the same ideas in multiple venues (first in a peer review article, then in a book, and subsequently in edited volumes).

Relative to gaming that occurs with other metrics, impact metrics are fairly transparent. We know who is citing whom and where. This means that citations leave a trail that can be investigated. Suppose one suspects that there are "citation cartels" in a field or subfield, i.e., quid pro quo agreements among scholars or journals to cite one another.[10] A weak cartel, one that involves just a few errant citations or a few publications, will probably be impossible to detect, but will also have little effect on anyone's impact metric. A strong cartel, by contrast, should be fairly easy to detect, and we can imagine algorithms developed explicitly for this purpose. Consider a network analysis that maps citation patterns across members of a discipline, measuring the strength of each dyadic relationship. Such an analysis can render an expected relationship for every dyad (every pair of members). This can then be evaluated against the actual strength of the relationship. Those dyads that are considerably stronger (e.g., more than one standard deviation) than predicted, might be regarded as suspect, i.e., probable cartels.

In this respect, impact metrics are a bit like insider trading in the stock market. We cannot stop such behavior from occurring, but we can frequently detect it after the fact – especially if the behavior is egregious. WoS, the most influential aggregator of impact metrics to date, is already implementing such sanctions on a regular basis (see e.g. Van Noorden 2013). However, their process is only partially transparent, focused on journals rather than individual researchers, and specific to the WoS index. An openly accessible and transparently constructed list of offenders would be invaluable.

Genre bias

Certain types of work are more likely to be cited than others, and these patterns may not always accord with our sense of their original contributions to scientific knowledge. For example,

[10] This flouts the intended function of a citation, i.e., to recognize important work in a field (see discussion at the outset of this section), and renders impact metrics less meaningful.

methodological studies and reviews of the literature are apt to be more widely cited than works that attempt to make a substantive contribution (Gingras 2016: Kindle Location 1112).

Some might not be alarmed by these well-established citation patterns. If a methodological innovation, or a cogent summary of a complex method, offers useful guidance to a field this might be regarded as an important contribution. Likewise, if a summary of the literature on a subject allows scholars to mark progress, to define the current state of a field, and to point the way to future research, this is surely a valuable contribution. Both of these activities deserve to be recognized and rewarded.

For those who feel that the citation-based rewards for these kinds of contributions are too great it ought to be possible to handicap this sort of work in an impact metric. This requires a classification system that can effectively distinguish between (a) methodological studies, (b) reviews of the literature, and (c) all other work (assumed to be substantive in some fashion). Such classifications might be derived from the orientation of journals (e.g., all articles in *Annual Review* would be easy to classify as literature reviews and all articles in a methods journal would be easily classified as methodological), from the keywords and abstracts that accompany nearly all journal publications, or from discipline-specific classifications (e.g., in economics).

Non-equivalence across contexts

Disciplines and sub-disciplines observe different norms with respect to citation practices. Some fields encourage comprehensive citation (law journals are probably the most extreme example) while others insist upon minimal citations, presumably to improve readability or to shorten the length of the text (see Chapter 5). Additionally, some fields are larger – with more practitioners and more journals – than others. Both factors contribute to widely varying citation counts across fields and (sometimes) subfields.

The problem of non-equivalence is invidious if one is comparing across fields without correction. However, impact metrics offer the possibility of defining impact in a variety of ways. For example, one may calculate citations according to percentiles within a field or subfield. Translated into percentiles, it is then possible to compare across fields. Fields might be broadly defined ("economics") or narrowly defined ("field experiments in development economics"). One might even define specific topics, allowing one to distinguish topics that are widely studied (e.g., "economic growth") from those that are less studied ("fertility") within a discipline.

Of course, there will always be questions about how to define fields and subfields, opening the way for arbitrary manipulation. However, most of the comparisons that are drawn between individual journals, scholars, or departments are within the same field, or closely related fields. Economists are compared to other economists, and development economists to other development economists. So non-equivalence across contexts may not be a very important problem, in practice.

Gender Bias

Citations, like many other parts of academia, perpetuate existing gender biases. In what scholars have dubbed the "Matilda effect" (Rossiter 1993), the female counterpart to the "Matthew effect," women consistently receive less credit for scientific discoveries in both public discourse and among their peers. This effect has been widely demonstrated for citation patterns in the fiels of communication (Knobloch-Westerwick and Glynn 2011), international relations (Malinak, Powers, and Walter 2013), and across many disciplines (Larivière et al. 2013) – though perhaps not in economics (Hammermesh 2018).

The causes of these disparate citation patterns are hard to assess, but two issues mentioned previously may contribute. First, researcher networks are not gender-neutral. For example, a study of

co-authorship suggests that male political scientists are more likely to coauthor with other men (Teele and Thelen 2017). And such male-dominated networks will, of course, produce male-dominated citation patterns. Self citations are another factor contributing to gender bias in citations. Using large corpora, two different teams of researchers have found that men cite themselves more often than women do, in some disciplines as much as twice as often (Ghiasi, Larivière, and Sugimoto 2016; King et al. 2017). These patterns are troubling for impact-based metrics and there is no simple fix. With effects varying across time and disciplines, there is no single adjustment factor that can be applied, though a reasonable point of departure in a model-based impact factor is a fixed-effect for author gender.

Evidently, impact metrics need to be used self-consciously, with an eye to their potential biases. These biases likely go beyond gender and may extend to researcher's race or sexual orientation (we are not aware of studies of the latter). In defense of citation metrics, we would note that there is reason to believe they are *less* biased than other commonly used measures. In a remarkable article, Weisshaar (2017) shows that gender is a strong predictor for whether researchers receive tenure even controlling for various impact measures, including citation counts.[11] The fact that we can use such – biased – measures of impact and productivity and still detect gender discrimination in current evaluation systems indicates the magnitude of biases inherent in the status quo. This should not distract from impact metrics' weaknesses, but those who oppose them on these grounds need to reckon with the reality that current assessment methods, which are often highly subjective and personalized, may fare even worse.

Clarifications

In this section, we address two dimensions that are commonly conflated with impact metrics. The first is *productivity* and the second is *journal impact*.

Productivity

Impact metrics do not attempt to measure the quantity of scholarly production, i.e., aka *productivity*. A citation count of 200 might reflect citations gathered by a single study or a hundred studies.

One might take the position that the quantity of production – of an individual scholar, a journal, a department, a university, or a subfield – is irrelevant. Arguably, a single study that garners 200 citations is equivalent in impact to one hundred studies that garner two citations apiece. Even so, for some purposes it is surely important to measure productivity as well as impact.

One approach is to combine these two dimensions into a single metric, such as the well-known *h*-index. Another approach is to regard quality (proxied by impact) and quantity as separate dimensions, to be measured independently. We generally prefer the latter approach, as there is no easy way to combine these two bits of information into a single index without considerable information loss.

The *h*-index has several additional flaws that have been often noted but, somehow, do not seem to discourage its proliferation in the academic world. A person's *h*-index can go up but not down; it is monotonic. This means that it is a poor gauge of over-time performance. A scholar who has not been cited in 30 years will have the same *h*-index as she did 30 years ago. Second, the h-index rewards individuals whose publication/citation ratio is close to 1, a rather arbitrary feature of academic performance. Third, and relatedly, the *h*-index tends to reward those who publish

[11] In fact, while number of publications and publications in a "top journal" were strong predictors, accumulated citations was not.

incessantly. By contrast, an individual with a few extremely influential publications will have a very low h-index, even if the latter are Nobel worthy (Gingras 2016, 42f.).

In any case, impact measures do not introduce any further complications into what is, at heart, a problem of multi-dimensionality. Reputation-based metrics, for example, face the same dilemma of how to combine quality and quantity.

Journal impact

Commonly, the quality of a scholar's publications is inferred from the quality of the journals that she publishes in. Indeed, a recent experimental study discovered that the addition of "low-quality" publications to a CV downgrade the perception of a scholar's work, even when the rest of the publications on the CV are the same (Powdthavee et al. 2018).

Nowadays, the most common metric for journal quality is a journal's *impact factor*, derived from citations to articles published in that journal over a period of time (typically 2 years). While this may be a useful device for librarians needing to make decisions over which journals to subscribe to, it is by no means a reliable proxy for *article* impact (see Larivière et al. 2016 for an excellent summary). While there is a correlation between journal impact factor and citations received by an article in the journal, the correlation is not particularly strong and appears to be decreasing since the 1990s, presumably due to improvements in search technology which effectively separate articles from the journals they appear in (Lozano, Larivière, and Gingras, 2012). Relatedly, the distribution of articles by received citations is strongly left-skewed (the majority of articles even in highly-ranked journals receive few citations) so that mean citation counts per journal is a misleading indicator, sensitive to the inclusion/exclusion of one or two widely cited articles.

Even more worrisome, there is strong evidence suggesting that journals have been attempting to "game" their impact factor (Heneberg 2016) as well as the exact criteria used in constructing such measures, in particular what counts as an "article" for the purpose of constructing the denominator of the journal impact factor.

Skepticism of journal impact factors as a tool for research assessment has become more mainstream recently. In 2012, a group of high-impact journals and funders passed the "Declaration on Research Assessment" (DORA 2012), which explicitly discourages the use of journal-level metrics to evaluate scholarship. So far as we can see, the demonstrable weakness of such aggregate measures and related heuristics such as "publication in a top journal" strengthens the case for article and/or author-level impact metrics.

The Impact of Impact Metrics

If impact metrics are here to stay (and, by all accounts, they are growing in influence), what sort of impact are they likely to have on the production of knowledge? We must consider this matter carefully before making a recommendation in favor of a new standard of excellence, which may have unintended consequences on the activity of scholars. We shall argue that six factors need to be considered: (a) the quest for impactful work, (b) the publication process, (c) efficiency, (d) flexibility, (e) meritocracy, and (f) democracy and reliability.

Impactful work

Insofar as citation counts drive the academic business, scholars will presumably seek to write articles and books that have broad impact. What sort of work might this be?

We presume that narrowly pitched empirical exercises are unlikely to have much impact on a field, even if published in a top venue – unless they have a surprising finding. The effect of this reorientation of scholarship is likely to further discourage replication, a problem that we have discussed at length in this section of the book. And it may prompt scholars to push their arguments further than the evidence allows or to engage in shenanigans to pass arbitrary hurdles of statistical significance. These must be counted on the negative side of the ledger, though they could be countered by other proposals discussed in this volume – to subject confirmatory studies to pre-registration (see Chapters 7, 9), to develop mechanisms for incentivizing and publicizing replication studies (see Chapters 10-11, 13), and so forth.

On the positive side of the ledger, exploratory work (i.e., work that propounds new theories or approaches) should be more widely appreciated. Work that is synthetic – theoretically and/or empirically – ought to be more widely appreciated. Finally, work that requires enormous investments of time and/or money, perhaps involving large teams of researchers and a long time-line, is more likely to be appreciated. In these respects, we can expect that highly ambitious work will be rewarded, and perhaps undertaken with greater frequency.

Likewise, one may anticipate that work that is highly redundant – overlapping with extant work – will be avoided. Note that insofar as impact displaces quantity as a measure of academic success, one synthetic paper that receives 1,000 citations is as valuable as ten papers receiving 100 citations each. Thus, there is no incentive to reduce ideas to the smallest publishable unit or to re-publish work in near-identical form in multiple venues.

The Publication process

In a world where post-publication success (judged by impact metrics) is valued over the status of the journal or press in which a manuscript is published one can imagine that the vetting process for these gatekeeping venues might be adjusted.

The current process is long (often lasting a year or more for one venue and several years if the first submission is unsuccessful), time-consuming (for authors, reviewers, and editors), stressful (for authors), and intrusive (insofar as it imposes upon the author a particular way of presenting her work, so as to please editors and reviewers). None of these features is desirable, on its face, though all might be defended if they serve to produce a better product.

However, even if the product is improved by all of this back and forth (a case that is by no means clear), it matters less if the ultimate arbiter of success is impact metrics rather than the status of the venue in which an article or book appears. In this scenario, the main job of journals and presses – considered as a whole – is to distinguish between work that is publishable and work that is not sufficiently worthy to be publishable. A secondary, but much less important, question is to decide upon the particular journal or press that a manuscript appears in, thus serving as an initial signal of its quality. Insofar as post-publication impact displaces journal or press reputation in the evaluation of scholarly work, the secondary goal is diminished. Scholars will care most about getting their work out, and obsess less about the imprimatur of the venue.

This, in turn, should encourage journals and presses to render up or down decisions fairly quickly about whether or not to publish a manuscript so that endless hours of precious time are not wasted in the vetting process and so that work reaches the public in a timely fashion. Authors already consider time-under-review as a criterion of journal selection; we can anticipate that this will become an even more important consideration in a world where journal prestige has less value in the academic marketplace.

Efficiency

Developing new impact metrics is a task that a few individuals (presumably, well-versed in bibliometrics) can perform for the entire academy. Learning how to use them is a one-time investment that everyone will need to make. Collecting the data necessary for calculating these impact metrics (assuming a well-developed open-source citation database is developed, as discussed in the previous section) is automatic, as are periodic updates. (By way of contrast, other metrics that have come to govern academic life are often extraordinarily time-consuming to collect and to maintain, feeding the growth of administrative staff that suck up university resources and faculty time [Muller 2018: ch 7].) Finally, the application of impact metrics to gatekeeping tasks is fairly easy.

Impact metrics are therefore a highly efficient mode of evaluation. This, in turn, should relieve burdens on decisionmakers – researchers, administrators, government officials – so they can get on with other tasks (e.g., research and teaching). Too often, debates about academic production are carried on without reference to opportunity costs. Evaluation, like any other activity, takes time, and the time spent on this activity is not directly productive. Insofar as impact metrics save us time, enhancing efficiencies in the production of knowledge, this should be counted as a blessing.

Flexibility

Readers should appreciate the enormous flexibility of impact measures, which can be weighted, aggregated, and denominated in various ways, providing a tool that can be adapted for a wide variety of purposes. One can discount the number of citations received by an individual, a study, or an academic institution by the quality of the source (the journal/press or publication) – which, itself, can be benchmarked by its citation count or by its network centrality (West and Vilhena 2014). One can discount the number of citations by their total, e.g., by a logarithmic transformation. (This helps to overcome the extreme skewness of citation counts (Hamermesh 2018), though it may be wondered whether, on substantive grounds, the 10^{th} citation should be viewed as less valuable than the 1000^{th}.) One can discount citations by the number of coauthors of a study. One can transform raw citations into percentiles based on overall citations for a field, subfield, discipline, country, or some other unit, generating an adjusted index that is (arguably) more equivalent across diverse contexts. On the assumption that there is a persistent gender bias in academe, one may construct model-based impact metrics that include a gender dummy.

In this fashion, many potential biases can be dealt with in statistical models. Of course, these models are only as strong as the assumptions that go into them. Nonetheless, they offer a practical recourse to problems that are widely recognized and measurable (through metrics).

One can also move beyond traditional academic sources to gauge the impact of a work *outside* the academy by registering comments, ratings, re-Tweets, Facebook likes, Pinterest shares, bookmarks, and microblogging from social media or views and downloads from repositories such as Mendeley, Academia and ResearchGate – a new area known *alternative metrics*, or *altmetrics* (Costas et al. 2014; Priem 2014).

Ultimately, one would hope to go further, to measure the impact of academic work on decisionmakers. A recent report points out,

> Evidence of external impacts can take a number of forms – references to, citations of or discussion of an academic or their work; in a practitioner or commercial document; in media or specialist media outlets; in the records of meetings, conferences, seminars, working groups and other interchanges; in the speeches or statements of authoritative actors; or via inclusions or referencing or weblinks to research documents in an external organisation's websites or intranets; in the funding, commissioning or contracting of research or research-based consultancy from university teams or academics; and in the direct involvement of

> academics in decision-making in government agencies, government or professional advisory committees, business corporations or interest groups, and trade unions, charities or other civil society organisations (Wilsdon et al. 2015: 46).

We have so far considered only a single article-level metric: the count of citations it has received. This is not for a lack of options. Looking at author impact, Wildgaard et al. (2018) identify 108 different indicators and the number has likely grown significantly since. Several of the concerns with impact measures noted above, such as the quality of citations and concerns about gaming the system, are particularly sensitive to crude measure such as absolute citation counts. Even simple corrections, such as excluding self-citations, can help to alleviate some of the problems, and more sophisticated measures can provide significant improvements.

The most promising measures, in our opinion, are *network-based*. These metrics assess the position of an article (or an author) in the network of all academic citations. In this fashion, they are able to register not just citation counts but also whether a study is cited by other influential studies and whether it is cited outside a narrow set of scholars working on a very specific topic. This methodology will be most familiar to readers from the PageRank application that underlies Google search engine. One promising implementation, built with data from MA is the Author Level Eigenfactor, or ALEF (Wesley-Smith, Bergstrom, and West 2016). Particularly intriguing is their suggestion in a subsequent paper (Portenoy, Hullman, and West 2016) to graph authors' influence networks, which depict a multi-dimensional picture rather than a single score. The authors provide a website (scholar.eigenfactor.org) allowing users to generate such graphs.

Granted, the flexibility of impact metrics also complicates their usage. There is no single statistic that will serve all purposes and overcome all difficulties. Instead, there are a variety of statistics – and many more that are likely to be developed in the coming years – that serve a variety of purposes. End-users must pick and choose carefully, and this opens the way for abuse. One such abuse is the over-reliance on individual metrics such as the *h*-index, whose flaws we have touched upon. Even so, we regard flexibility as a virtue, and we expect that, over time, a small number of metrics will come to be widely used, their strengths and weakness will be well-understood, and hence less liable to abuse.

Meritocracy

One of the flaws of reputational metrics is that evaluations of studies, individuals, departments, universities, journals, and presses bleed into one another. The reputation of one feature affects the reputation of another. For example, the reputation of a scholar is, in part, derived from her university, while the reputation of a university is, in part, derived from the scholars who teach there. The same is true for journals, presses, and subfields. Everything is endogenous to everything else. That is, the reputation of each element of a discipline is dependent on the reputations of all the other elements of a discipline.

Consequently, reputational evaluations are highly ambiguous. It is not clear what they represent, aside from an overall judgment of reputation. Additionally, because institutions are generally more well-known than individuals, institutional reputations tend to overshadow individual reputations. We are familiar with the top departments and journals in our field; we are less familiar with the individuals who work in the top departments or the articles that appear in top journals. Consequently, the latter are evaluated – reputationally – by their location. Scholar A must be tops since she is in a top department; Article B must be tops since it appears in a top journal. A similar logic applies to those individuals and articles who are not highly placed; they are downgraded by their association with institutions that have a lesser reputation.

Finally, these institutional reputations are sticky, for there is nothing – other than overall reputation – to cause them to rise or fall. Reputations are self-perpetuating. Harvard is tops because it is Harvard, and whatever it does, and whomever is associated with it, becomes tops.

Impact metrics, by contrast, allow for *independent* judgments of each unit – studies, individuals, departments, universities, journals, presses, and so forth. Each can be judged by its impact on a subfield, a discipline, or the social sciences at-large. Each varies independently, which means that a study, a scholar, a department, a university, a journal, or a press may rise or fall in impact over time and their fate is not linked (or at least not closely linked) to other institutions and individuals.

In these respects, an impact-based system of evaluation is considerably more meritocratic than a reputational system. High-performing individuals within low-performing departments receive their due. High-performing articles within low-performing journals receive their due. And so forth. It means that individuals and institutions at the top of the pecking order cannot sit on their laurels, and individuals and institutions at the bottom of the order need not remain forever in the basement.

If impact metrics are highly valued, emphasis should shift from the rank of the journal or press that a study was published in to the impact that study has achieved. Ultimately, who cares if an article (book) was published in a top journal (press) or a bottom journal (press)? Likewise, when evaluating an individual scholar we should be concerned primarily with what she has published and the impact that work has had – not where the work was published or what department or university (or country) she belongs to. Impact should trump reputation. Alternately framed, impact should drive reputation.

Of course, impact metrics do not mean that units – studies, individuals, departments, universities, journals, and presses – are *entirely* independent. The imprimatur of a publication venue and a scholar's institutional affiliation will always have some effect on the impact of a published work. Nonetheless, over time, one can hope that the arguments and evidence contained in a strong publication will outweigh the reputational effects of its venue and its author. Hamermesh (2018, 151) finds that "many economists at lower-ranked faculties…are cited more than the median faculty members at higher-ranked schools," suggesting a high degree of independence between the fate of an article and the institutional affiliation of its author. Likewise, a study of journal impact factors and the impact of articles published in those journals shows that the relationship began to weaken at the end of the twentieth century in response – one presumes – to the accessibility of articles on-line, which effectively removes the journal as a publishing unit allowing readers to access articles individually (Lozano, Larivière, and Gingras 2012).

Insofar as we value meritocratic principles in the academy this ought to be pleasing. And insofar as meritocratic principles affect incentives – for those at the top as well as for those at the bottom – we should anticipate that this more meritocratic method of assessment will boost academic production overall.

Democracy and Reliability

Intertwined with the question of meritocracy is the question of democracy and reliability. In any discipline, there are a few top journals, presses, departments, and funders. These top units play a critical role in the production of knowledge, serving as gatekeepers for the most desirable positions, publications, awards, grants, and fellowships. To a remarkable extent, membership in these elite institutions overlaps, or rotates, among a small elite such that decisions affecting everyone in the academy are monopolized by a few.

To some extent, this can be justified as the product of a thoroughgoing meritocracy, where the most capable members rise to the top. Surely, no one would advocate that evaluations of

manuscripts, grant and fellowship proposals, and candidates for awards, jobs and promotion be divvied out to members chosen by lot or surveyed collectively. It is the opinion of top experts that should matter most in these determinations, and not everyone is equally informed – especially in areas of academic work that are highly specialized.

However, the current system of scholarly evaluation is top-heavy in a way that cannot be entirely justified on the basis of expertise. Positions play an important role. For example, in seeking outside reviewers for academic promotion and tenure universities will often stipulate that evaluators be situated in "peer or peer plus" institutions. Position trumps achievement, in other words. One suspects that similar considerations apply, albeit in more informal ways, to other gatekeeping tasks. On editorial boards, one finds a familiar collection of names – *eminenti* whose role is primarily to enhance the status of the journal or press, but who are nonetheless in a position to determine publication decisions of an important journal or press. Grants and fellowships are meted out to the already-famous, presumably to enhance the status of the grant or fellowship rather than to advance scientific progress. Meanwhile, the talents of scholars at lesser institutions are under-employed.

It is difficult to defend such a system as being in the long-run interests of knowledge production. It is, on the face of it, neither efficient nor fair. Nor does it offer encouragement to scholars who are outside the top elite to improve their game, as one's position in the hierarchy is likely to be determined quite early in one's professional career, after which it is difficult to alter.

Even if the current system were entirely meritocratic it is highly susceptible to stochastic error. When evaluating a study (e.g., for a journal or press) or a candidate (e.g., for a job, fellowship, or grant), opinions among social scientists are apt to vary. Reviews of the same manuscript are often at odds with one another; reviews of the same applicant, e.g., by members of a search committee, are often at odds with one another. This is why we often cite the "luck of the draw" when a paper, grant/fellowship proposal, or a job application is accepted or rejected.

These outcomes are stochastic not simply because there is disagreement among vetters but also, perhaps more importantly, because the pool of vetters is extremely small. Most decision-points in an academic career depend upon a small number of gatekeepers, who render a decision at a particular point in time. Decisions about hiring are made by the hiring committees of a handful of departments who have positions to fill in one's field. Decisions about publication are made by an editor along with a few reviewers. Likewise, decisions about grants and fellowships are made by a small panel, perhaps supplemented by an external review(s).

A core axiom of probability theory is that small samples produce unreliable results. We experience this unreliability at virtually every step of an academic career. In this light, a principal blessing of impact metrics is that they enlarge the pool of vetters, i.e., the number of experts who render a judgment about a given study. By enlarging the pool, impact metrics thus democratize the process of academic evaluation while enhancing its reliability.

Importantly, participants are not chosen randomly; they are restricted to those who can be presumed to know something about the work in question. They are experts, by virtue of having published on a subject. In this fashion, the fate of a study, a researcher, or any other unit of interest rests with the entire academic community – delegated to those who are experts in the relevant area – rather than with a handful of individuals who happen to control top journals, presses, and departments at a particular point in time.

Conclusions

In the first section of this chapter, we pointed out that extant citation databases – GS, Scopus, WoS, and MA – are prone to error (especially with respect to identifying authors), under-sourced (books,

in particular, are often not included), and for the most part closed-source (and thus not fully transparent or fully adaptable to the many functions they might serve). We believe that these flaws are likely to be corrected in the near future as these databases (or at least one of them) expand their coverage, improve their algorithms, and shift to open-source. Until that point, we have reservations about the use of impact metrics as an evaluative tool.

In the second section, we carefully reviewed other objections to the use of impact metrics. We acknowledged that there are many problems with the way impact metrics are currently employed. For example, the two most common metrics, citation totals and *h*-index scores, reveal some important information but can also mislead, especially if one is comparing scholars from different fields and subfields or at different points in their career. We need a broader array of impact metrics, suitable for different purposes, and the community of end-users needs to be acquainted with the uses and possible abuses of each metric. These are not insurmountable obstacles, but they should not be taken lightly.

In the third section, we clarified the distinction between impact metrics and other subjects that are frequently conflated such as productivity (the number of studies a scholar or institution has produced). Likewise, metrics that attempt to combine impact and productivity – such as the *h*-index – are merging dimensions that are often poorly correlated and thus difficult to combine without bias.

In the final section of the paper, we discussed the probable impact on scholarship of the increasing role of impact metrics. We reasoned that the rise of this new metric should enhance incentives to produce impactful work, which means that highly ambitious projects should receive their due. By the same token, scholarship that aims for incremental gains, e.g., through replications of extant work, might be given short shrift. We reasoned that the role of publishers as principal gatekeepers in the production of knowledge would be downplayed. An implication is that the publishing process might become more efficient and less stressful. We reasoned that the use of impact metrics might enhance the efficiency of judgments about scholarly achievement, freeing up time to spend on other matters such as research and teaching. We reasoned that impact metrics offer enormous flexibility in the sorts of comparisons that can be drawn and the sorts of impacts – extending, ultimately, to the "real world" – that might be measured. We reasoned, finally, that the use of impact metrics would make the academic world somewhat more meritocratic and more democratic.

Before concluding, it is important to recognize that impact metrics share the strengths and weaknesses of all statistics. A measure of democracy, of corruption, of gross domestic product, or any other concept is useful if it is used responsibly, with an understanding of its limitations. Too often, these metrics are treated as precise or are given an overly broad interpretation. Too often, statistics serve as a substitute for deliberation.

Fourcade and Healy (2017, 294), argue that even "[a]mateurish or barely defensible data collection and ranking schemes turn out to have the capacity to control the status order of professional fields, partly just in virtue of their quantitative character." They also point out that such measures allow for and drive the increasing commodification of a sector; by assigning "value", they allow for "Seeing like a market" (Fourcade and Healy 2016). An oft-cited example is the effect of the *US News* ranking on the behavior of law schools, which went to considerable lengths to improve their relative ranking through activities guaranteed to be ineffectual in improving instruction or research (Espeland and Sauder 2016). Lawrence (2007) makes this argument explicitly about the incentives created by a reliance on impact factors in the evaluation of science ranging from overhyping of research results to an emphasis on networking over research. While we disagree with some of his concerns (and have addressed them throughout this paper), we share concerns about over-reliance on metrics or their mindless application.

In any case, to say that metrics are abused is not to say that we would be better off without them. The question to be asked of any tool is not whether it is good or bad but whether it assists in accomplishing the tasks at hand, taking into account other tools that might be used for that purpose and the overall efficiency of the enterprise.

With respect to metrics of academic performance, the principal alternatives are grounded in *reputation* (surveys of scholars about the standing of particular journals, departments, or schools), *quantity* (the number of published articles and books), or some admixture of the two. We find these metrics informative and perhaps essential for some purposes, but also limited in applicability and insight. Reputational measures may be applied to institutions but not to individuals. Measures of quantity may be applied to individuals but only to some institutions. (One cannot judge a journal by the quantity of articles published.) Neither approach offers a direct measure of scholarly impact. The other alternative is *peer review*, i.e., reading the studies produced by an individual, department, journal, or field, and reaching conclusions based on that reading. This approach, while absolutely essential in many contexts, is time-consuming, subjective, and incapable of rendering systematic comparisons. All existing methods of evaluation are prone to abuse, e.g., gaming (by researchers) or bias (of evaluators). Reputation and peer review seem especially prone to biases grounded in gender, race, or school networks.

When considered in light of the alternatives, impact metrics have much to offer. Our first proposal, therefore, is that social scientists make full use of this class of metrics. Insofar as they can assist in decisions about hiring and promotion, the allocation of resources, and other gatekeeping functions they ought to be a part of our toolkit – as a supplement, not a replacement, for other methods.

Our second proposal is to push for the development of better citation databases, and a wider class of impact metrics, so that flaws in currently available metrics can be overcome. The major technical obstacle is the database, discussed in the first section. Once that issue is solved – ideally by an open, freely available citation corpus rather than a proprietary black box – it should be possible to construct new metrics that overcome some of the oft-noted deficiencies of extant metrics.

Our third proposal is to use impact metrics appropriately. If, for example, one is interested in evaluating the impact of a journal it is appropriate to examine journal impact metrics (aka factors). If one is interested in evaluating the impact of an article, it is appropriate to examine article impact metrics. And so forth. By contrast, it does not make sense to assume that *journal* impact factor is a proxy for *article* impact, or that a scholar's impact can be inferred from the impact of the journals she has published in.

Our fourth proposal is to make more use of network-based citation networks such as the author-level eigenfactor. These algorithms incorporate not just citation counts for a study but also the importance and breadth of the locations in which a given study is cited. This provides a deeper measure of impact and one less prone to gaming.

References

Basken, Paul. 2018. "UT-Austin Professors Join Campaign Against Faculty-Productivity Company." Chronicle of Higher Education (January 24).

Bornmann, Lutz, Hans Dieter Daniel. 2008. "What do citation counts measure? A review of studies on citing behavior." Journal of Documentation 64(1), 45–80.

Breuer, Peter T., and Jonathan P. Bowen. 2014. "Empirical Patterns in Google Scholar Citation Counts." ArXiv:1401.1861 [Cs], April, 398–403. https://doi.org/10.1109/SOSE.2014.55.

Brzezinski, Michal. 2015. "Power Laws in Citation Distributions: Evidence from Scopus." Scientometrics 103 (1): 213–228. https://doi.org/10.1007/s11192-014-1524-z.

Campbell, Donald T. 1979. "Assessing the impact of planned social change." Evaluation and Program Planning 2(1): 67–90.

Catalini, Christian, Nicola Lacetera, Alexander Oettl. 2015. "The incidence and role of negative citations in science." Proceedings of the National Academy of Sciences, 112(45), 13823–13826.

Clements, Kenneth W., Patricia Wang. 2003. "Who Cites What?" Economic Record 79 (245): 229–44.

Costas, R., Z. Zahedi, P. Wouters. 2014. "How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications." Scientometrics 101(2): 1491–1513.

Cronin, Blaise. 1984. The Citation Process. The Role and Significance of Citations in Scientific Communication. London: Taylor Graham.

Cronin, Blaise, and Cassidy R. Sugimoto, eds. 2014. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, Massachusetts: The MIT Press.

Delgado López-Cózar, E., N. Robinson-García, D. Torres-Salinas. 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.

DORA. 2012. "San Francisco Declaration on Research Assessment." 2012. https://sfdora.org/read/.

Elkana, Y., J. Lederberg, Robert K. Merton, A. Thackray, H. Zuckerman. 1978. Towards a metric of science: The advent of science indicators. New York: Wiley.

Ellison, Glenn. 2002. "The Slowdown of the Economics Publishing Process." Journal of Political Economy 110 (5): 947–93. https://doi.org/10.1086/341868.

Erikson, Martin G., Peter Erlandson. 2014. "A Taxonomy of Motives to Cite." Social Studies of Science 44:4, 625–637.

Espeland, Wendy Nelson, Michael Sauder. 2016. Engines of Anxiety: Academic Rankings, Reputation, and Accountability. New York: Russell Sage Foundation.

Fourcade, Marion, and Kieran Healy. 2016. "Seeing like a Market." Socio-Economic Review, December, 9–29. https://doi.org/10.1093/ser/mww033.

Fourcade, Marion, and Kieran Healy. 2017. "Categories All the Way Down." Historical Social Research 42 (1): 286–96. https://doi.org/10.12759/hsr.42.2017.1.286-296.

Fried, Eiko. 2018. "7 Sternberg Papers: 351 References, 161 Self-Citations." March 29, 2018. http://eiko-fried.com/sternberg-selfcitations/.
Garfield, E. 1955. "Citation indexes for science. A new dimension in documentation through association of ideas." Science 122(3159), 108–111.
Ghiasi, Gita, Vincent Larivière, and Cassidy R Sugimoto. 2016. "Gender Differences in Synchronous and Diachronous Self-Citations," 8.
Gingras, Y. 2016. Bibliometrics and research evaluation: Uses and abuses. Cambridge, MA: MIT Press.
Hamermesh, Daniel S. 2018. "Citations in Economics: measurement, uses and impacts." Journal of Economic Literature 56(1), 115–156.
Harzing, A.W. 2007. Publish or Perish, available at http://www.harzing.com/pop.htm
Heneberg, Petr. 2016. "From Excessive Journal Self-Cites to Citation Stacking: Analysis of Journal Self-Citation Kinetics in Search for Journals, Which Boost Their Scientometric Indicators." PloS One 11 (4): e0153730. https://doi.org/10.1371/journal.pone.0153730.
Hix, Simon. 2004. "A global ranking of political science departments." Political studies review 2.3: 293-313.
Hudson, John. 2007. "Be Known By the Company You Keep: Citations—Quality or Chance?" Scientometrics 71 (2): 231–38.
Hug, Sven E., Martin P. Brändle. 2017. "The Coverage of Microsoft Academic: Analyzing the Publication Output of a University." Scientometrics 113: 1551-1571.
Jensenius, Francesca R., Mala Htun, David J. Samuels, David A. Singer, Adria Lawrence, Michael Chwe. 2018. "The Benefits and Pitfalls of Google Scholar." PS: Political Science & Politics (forthcoming) 1-5.
Knobloch-Westerwick, Silvia, and Carroll J. Glynn. 2013. "The Matilda Effect—Role Congruity Effects on Scholarly Communication: A Citation Analysis of Communication Research and Journal of Communication Articles." *Communication Research* 40 (1): 3–26. https://doi.org/10.1177/0093650211418339.
King, Molly M., Carl T. Bergstrom, Shelley J. Correll, Jennifer Jacquet, and Jevin D. West. 2017. "Men Set Their Own Cites High: Gender and Self-Citation across Fields and over Time." Socius: Sociological Research for a Dynamic World 3 (December): 237802311773890. https://doi.org/10.1177/2378023117738903.
Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. "Bibliometrics: Global Gender Disparities in Science." Nature News 504 (7479): 211. https://doi.org/10.1038/504211a.
Larivière, Vincent, Veronique Kiermer, Catriona J. MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, and Stephen Curry. 2016. "A Simple Proposal for the Publication of Journal Citation Distributions." BioRxiv, July, 062109. https://doi.org/10.1101/062109.
Lawrence, Peter A. 2007. "The Mismeasurement of Science." *Current Biology* 17 (15): R583–85. https://doi.org/10.1016/j.cub.2007.06.014.
Lazear, Edward P. 2006. "Speeding, terrorism, and teaching to the test." The Quarterly Journal of Economics 121 (3): 1029–61.
Lozano, George A., Vincent Larivière, and Yves Gingras. 2012. "The Weakening Relationship between the Impact Factor and Papers' Citations in the Digital Age." Journal of the American Society for Information Science and Technology 63 (11): 2140–45. https://doi.org/10.1002/asi.22731.
Maliniak, Daniel, Ryan Powers and Barbara Walter. 2013. "The Gender Citation Gap in International Relations." International Organization 67 (4): 889-—922.
Merton, Robert K. 1968. Social Theory and Social Structure. Enlarged ed. New York: The Free Press.
Muller, Jerry Z. 2018. The Tyranny of Metrics. Princeton: Princeton University Press.

Nicolaisen, Jeppe. 2007. "Citation Analysis." Annual Review of Information Science and Technology 41:1, 609–641.
Orazbayev, Sultan. 2017. "Diversity and Collaboration in Economics." Unpublished, UCL SSEES.
Osterloh, M., B.S. Frey. 2015. "Ranking games." Evaluation Review, 39(1), 102–129.
Portenoy, Jason, Jessica Hullman, and Jevin D. West. 2016. "Leveraging Citation Networks to Visualize Scholarly Influence Over Time," November. https://doi.org/10.3389/frma.2017.00008.
Powdthavee, Nattavudh, Yohanes E. Riyanto, Jack L. Knetsch. 2018. "Lower-Rated Publications Do Lower Academics' Judgments of Publication Lists: Evidence from a Survey Experiment of Economists." Journal of Economic Psychology (forthcoming).
Priem, Jason. 2014. "Altmetrics." In B. Cronin & Cassidy R. Sugimoto (eds), Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact (Cambridge, MA: MIT Press) 263-87.
Rossiter, Margaret W. 1993. "The Matthew Matilda Effect in Science." Social Studies of Science 23 (2): 325–41.
Samuels, David. 2013. "Book Citations Count." PS: Political Science & Politics 46(4), 785-790.
Shotton, David. 2010. "CiTO, the Citation Typing Ontology." Journal of Biomedical Semantics 1 (1): S6. https://doi.org/10.1186/2041-1480-1-S1-S6.
Sugimoto, Cassidy R. and Vincent Larivière. 2018. Measuring Research: What Everyone Needs to Know. Oxford University Press.
Teele, Dawn and Kathleen Thelen. 2017. "Gender in the Journals: Methodology, Coauthorship, and Publication Patterns in Political Science's Flagship Journals." PS: Political Science & Politics: 50(2): 433-447.
Testa, James. 2016. "Journal Selection Process." Clarivate Analytics. https://clarivate.com/essays/journal-selection-process/.
Todeschini, R., A. Baccini. 2016. Handbook of bibliometric indicators: Quantitative tools for studying and evaluating research. Weinheim: Wiley-VCH.
Van Noorden, Richard. 2010. "A profusion of measures." Nature 465.7300: 864-867.
Van Noorden, Richard. 2013. "New Record: 66 Journals Banned for Boosting Impact Factor with Self-Citations: News Blog." Nature News Blog (blog). June 19, 2013. http://blogs.nature.com/news/2013/06/new-record-66-journals-banned-for-boosting-impact-factor-with-self-citations.html.
Wade, Alex D., Kuansan Wang, Yizhou Sun, and Antonio Gulli. 2016. "WSDM Cup 2016: Entity Ranking Challenge." In Proceedings of the Ninth Acm International Conference on Web Search and Data Mining, 593–94. WSDM '16. New York, NY, USA: ACM.
Weisshaar, Katherine. 2017. "Publish and Perish? An Assessment of Gender Gaps in Promotion to Tenure in Academia." Social Forces 96.2: 529-560.
Wesley-Smith, Ian, Carl T. Bergstrom, and Jevin D. West. 2016. "Static Ranking of Scholarly Papers Using Article-Level Eigenfactor (ALEF)." ArXiv:1606.08534 [Cs], June. http://arxiv.org/abs/1606.08534.
West, Jevin D., Daril A. Vilhena. 2014. "A Network Approach to Scholarly Evaluation." In B. Cronin & Cassidy R. Sugimoto (eds), Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact (Cambridge, MA: MIT Press) 151-65.
Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. 2015. The metric tide: Report of the independent review of the role of metrics in research assessment and management. London: SAGE.