Are Citation Metrics a Good Thing?

Abstract

Citation metrics are statistical measures of scientific outputs that draw on citation indexes. They purport to capture the impact of scientific articles and the journals in which they appear. As evaluative tools, citation metrics are mostly used in the natural sciences, but they are also acquiring an important role in the humanities, thereby affecting the development of research programs and institutions. While the strengths and weaknesses of citation metrics are extensively debated in a variety of fields, they have only recently started attracting attention in the philosophy of science. This paper takes a further step in this direction and presents an analysis of citation metrics from the perspective of a Kuhnian model for the development of science. To do that, it starts with an overview of citation metrics both at the general level and at the level of specific metrics, such as Impact Factor, h-index, and field-specific indicators. After that, it engages with Gillies' argument against the use of citation metrics for scientific research. According to Gillies (2008), citation metrics tend to over-protect *normal science* at the expenses of *revolutionary science*. This paper shows that, under certain conditions, citation metrics can in fact arbitrarily hinder the development of *normal science* and, in light of this, it cautions against using them for evaluative purposes.

1 Introduction

Publication metrics have become a dominant "currency" in science. Such metrics provide measures of research outputs by drawing upon citation analysis (Andersen 2019, van Raan 2019). They purport to capture the impact of scientific articles and the outlets in which they appear, viz. (peer-reviewed) journals. In the natural sciences especially, it is common to use them in assessing research, thereby guiding the development of research programs and the opportunities available to scholars and institutions. Besides the natural sciences, citation metrics are also acquiring an increasingly important role in the arts and humanities.

In recent years, however, there have been numerous calls to move away from citation metrics in favor of more exhaustive criteria (Hicks et al. 2015), although proposals to abandon such models have led to criticisms (Poot and Mulder 2021). In light of the current discussion, and given the stakes that are involved, it is plausible that citation metrics will at most be combined with qualitative criteria, rather than be replaced by them. Among the reasons to retain them, scientists raise concerns that if evaluation metrics were erased from specific contexts, such as particular institutions or even countries, those contexts would be at a disadvantage—for instance in university rankings or in competing for international grants—relative to those that have clear, albeit imperfect, measures of performance.

In this paper, I will weigh some of the main arguments for and against the adoption of quantitative indicators in the evaluation of scientific research. To structure the discussion, I shall divide the main arguments into the theoretical and the practical, placing particular focus on Donald Gillies's analysis of research assessment (2008).

One of the main arguments against the use of citation metrics, which is supported by Gillies, is that they provide a too quick assessment of scientific work, the proper evaluation of which actually requires more time. As an example, some metrics—for instance the Impact Factor—looks at a journals' citation patterns in the two- or five-year periods following an article's publication. However, the time it takes for a scientific community to recognize a discovery rarely corresponds to such a timeframe. As Gillies points out, had the evaluative system that we know today been in place historically, then scholars such as Wittgenstein and Frege, to name but two, would not have received proper academic support.

Gillies draws upon Thomas Kuhn's work on the development of science to argue that

a metrics-based evaluative system tends to over-protect *normal science* at the expense of *revolutionary science*. This is because the latter typically does not receive due credit from a particular scientific community (including in terms of citation metrics) until a new paradigm emerges. To Gillies's mind, the current system is highly exposed to the risk of overlooking "pink diamonds"—authors such as Wittgenstein or Frege—in favor of the status quo.

The discussion above reveals genuine limitations of the use of metrics as a reliable evaluative tool of scientific work. On the other hand, one of the strongest arguments in favor of their adoption is that, since we need an evaluative system, citation metrics offer just that. There is increased pressure from policymakers to provide evidence of scientific performance. This evidence is then used to justify public expenditure, to strengthen the accountability of scientists, and even—it is argued—to bolster trust in science (van Raan 2019). The main idea is that, while metrics are far from ideal, most of their biases can be corrected through more refined measures. By contrast, choices that are left to individual assessors are more subjective and prone to partiality.

In this paper, I will start by going back to the historical roots of citation analysis with the aim of shedding light on some of its main original aims and the conditions that necessitated its development (Sec. 2). I will show that citation analysis introduced a new criterion to the organization and retrieval of scientific literature, one based on the references that are contained in the literature. I will claim that this criterion provides a grounding for the use of citation metrics both for evaluative purposes and as a tool to navigate the literature (Sec. 2). Secondly, I will give an overview of some of the most common citation metrics, focusing on their properties and limitations as evaluative tools (Sec. 3). Finally (in Sec. 4), I will consider Gillies's argument about the risk of overlooking "pink diamonds". I will claim that, if we endorse Kuhn's analysis on the progress of science—as it is laid out in *The Essential Tension* (1959)—the risk of thwarting revolutionary science should not be the principal cause for concern that is raised by the metrics system. Rather, the risk of hindering the genuine development normal science is what is at stake. Therefore, I will focus on the conditions under which the system of metrics can support normal science and consider whether such conditions have so far been met. In light of this analysis, I recommend caution in using the metrics as evaluative tools and propose that they be combined with other evaluative methods.

A more general conclusion is that, regardless of the whether the system based on evaluative

metrics is endorsed, these indicators deeply affect the opportunities available to research programs and their scholars. Thus far, they have played only a minor role in the humanities in comparison to the natural sciences, but they have already become relevant and it is likely that they will become even more so in the foreseeable future. Even only indirectly, these metrics are already relevant in the humanities. To give an example, university rankings measure research and teaching performance partly on the basis of quantitative indicators and, while controversial, have become influential in attracting students and researchers, and in the allocation of funding. Therefore, so long as the metrics remain in place, it is in the interest of scientists to be aware of them and understand how they function, as well as recognizing their strengths and weaknesses.

2 The roots of citation analysis

The aim of this section is to retrace the roots and original purposes of citation analysis. Citation metrics are often presented as an evaluative tool that has been introduced to determine the allocation of resources in an increasingly competitive scientific market. In this section, I will argue that one aspect that is typically not as well reported is that citation metrics were also advanced as a tool to help scientists navigate a rapidly developing mass of literature. Once it becomes clear that these two applications of citation metrics are two sides of the same coin, it is easy to see, firstly, that the use of citation metrics as assessment tools is grounded in the use that scientists make of citations in their own work; and, secondly, that certain objections apply to both sides.

To start with, citation analysis is a quantitative method for the examination of various features of citations of publications—for instance, their number, patterns, and graphs. The dataset on which citation analysis is based is a citation index, that is, a bibliographic index that lists publications and, for each one, all the publications included in the index that refer to that publication.

The founding father of citation analysis is widely considered to be Eugene Garfield (1925–2017). Garfield compiled the Science Citation Index (SCI), now known as the *Web of Science*, at the end of the fifties. Before him, one of the first indexes of academic literature was provided for the field of chemistry by Paul and Edward Gross (1927) who compiled it *manually*. They

took one of the most representative journals in chemistry in their time, *The Journal of the American Chemical Society*; they noted all the journals that were cited by articles published in that journal over a certain period of time, and then ranked them according to the number of times they were cited.

Aided by computers, Garfield's citation index included most of the science and technology journals of his time. Garfield's method was thus different than the one used by Gross and Gross: he did not consider only the citations from articles published in a particular representative journal ("top-down"), but rather the citations that all the journals in the index received from each other ("bottom-up"). The figures mounted up immediately: in 1933, the Gross and Gross' chemistry index included 247 different journals, while, in 1971, the SCI collected 2,200 science and technology journals. As of today, the Web of Science Core Collection contains around 21,000 peer reviewed-journals.

Garfield's index played a crucial role in opening up an entirely new body of statistical work on scientific production. It was instrumental in the establishment of research programs such as scientometrics and bibliometrics, which have been extremely prolific since. Scientometrics aims at measuring the growth and development of science via mathematical models and statistical analysis, while bibliometrics focuses in particular on statistical measures of articles, journals, books, and so forth. Over time, both fields have achieved a variety of results: to give some examples, they have developed citation metrics based on increasingly advanced statistical techniques; they have created bibliometrics indicators such as bibliographic coupling (Kessler 1963), co-citation analysis (Small 1973), and co-word analysis (Callon et al. 1983); and they have assisted the automatic indexing of search databases and, more recently, online search engines in information science (Polonioli 2020).

At the very outset, the statistical approach to scientific production provided, among other things, the first quantified measure of the exponential growth of science, in terms of publications rate and number of scientists. The increase in scientific production raised crucial questions about how such an expansion should be handled. Such issues concerned both policymakers, who were tasked with establishing criteria to determine budget allocation (Csiszar 2020); and scientists themselves, who faced the problem of processing a growing amount of literature in a limited amount of time. As we shall see below, citation analysis could apparently serve both aims.

As regards policy assessment, the statistical approach to scientific production showed that a recurrent characteristic is that it is not uniformly distributed. This feature was then taken as evidence to ground some of the early policies based on citation metrics. It was observed, for instance, that most scientific output typically comes from a small group of scientists (Lotka 1926); that citations are driven by a small number of papers (de Solla Price 1963); and that the relevant literature is scattered between a few crucial publications (de Solla Price 1963). As an example of how these factors were used for policy, citation ranking was proposed as a criterion to decide which periodicals to include in academic libraries operating under budget constraints (Gross and Gross 1933). For instance, when assessing the collection of journals owned by a library, it might be decided to acquire new periodicals that have attracted many citations and, vice versa, to exclude journals of lesser impact.

Besides policy assessment, however, authors were concerned with the impact of scientific growth on scientists themselves, as they had to process an ever-growing amount of prior work and keep up with a rapid inflow of new publications. In the words of Margolis (1967): "As a result of the recent expansion of scientific literature, more time and effort are being devoted to *the selection* of what is to be read than to the actual reading" (p.1213, italics added). He continued: "New information is *accumulating faster that it can be sorted out.* [...] A new scale of values based on citations is by no means infallible, or, in many cases, even fair, but at least it provides an alternative to the existing one [quantity of publications], which is at the root of the crisis." (p. 1219).

With respect to this point, citation analysis offered a new way of organizing and retrieving scientific literature. Previous classification systems were based on criteria such as alphabetical order and subject classification, which were less and less manageable as the mass of publications increased (Svenonius 2000). Citation indexes introduced a new kind of academic library, one which arranges and returns the literature on the basis of citations. At the center of this shift, is the crucial role given to *references*, which become the core of a *signaling system* that scientists can use to navigate the literature.

A concrete example of a new feature of citation indexes is known as *forward-citation searching*. Search engines like Google Scholar typically show a "cited by" link under each entry, which displays a list of all the papers that cite that paper. This feature opens up new possible search strategies. Before citation indexes, scientists typically sought new literature

by moving from one source to the references that that source contained. In other words, their searches could only proceed *backwards*. With citation analysis, authors could for the first time expand their literature searches beyond the references found directly in a text, by looking at the publications that cited that text *after its publication*. Thus, rather than proceeding only *backwards*, that is, moving from a paper to its prior sources, scholars could now move *forwards* in the literature and check if a paper was a solid reference or if it was already outdated by more recent scientific work (Garfield 1955).

Yet another role that Garfield envisioned for the citation index was that of an "association of ideas" index (Garfield 1955). He believed that the index would give scientists a way to follow the dissemination of a piece of work in the literature by providing a map of the scientific landscape based on the citation network of the papers in circulation (Biagioli 2018). This point is similar but not identical to the application of citation analysis that I have illustrated above. The main difference is that the latter refers to maps of citation networks that can be used, for instance, to observe the development or the communication structure of a research program, while the former concerns the role of citation indexes as what I refer to as a *literature selection device*.

To sum up, citation analysis flourished at a time when science was advancing at a faster pace than ever before. During this period, new statistical methods became available to collect data on scientific production and analyze them quantitatively. On the one hand, policymakers adopted citation metrics to make the assessment of scientific work faster and ensured it was based on clear and shareable criteria. These days, citation metrics are mostly associated with this evaluative purpose, which is also the aspect that most often flies "under the radar", for political reasons.

On the other hand, the expansion of scientific production required efficient and systematic tools to process an increasing volume of literature. In this respect, citation analysis offered a new method with which search and select the relevant research.

In either case—whether it is used by scientists or by policymakers—citation analysis tracks the use of publications in the literature. One of the arguments in favour of using citation metrics to assess scientists, is precisely that the metrics reflect what scientists consider to be relevant and worth *citing*. However, whether usage licenses quality assessment is a question at the center of the entire debate between the supporters and the critics of evaluative citation

metrics.

To shade light on this, the next section critically analyses citation metrics as evaluative tools both at a general level and at the level of a number of standard indicators. The survey begins with Impact Factor (IF) and then considers some of the alternatives that have been developed specifically to compensate for some of the IF's shortcomings, namely h-index, journal influence weight indicators, and field-specific indicators (e.g., FWCI).

3 An overview of citation metrics

Citation metrics are statistical measures that combine citation data with other variables, for instance citations over periods of time or citations over quantity of publications. The rankings published on the basis of these criteria—for individual scientists, articles, journals, departments, all the way up to entire universities—rest on the assumption that citations track certain positive aspects of scientific production. In other words, that ranking of publications reflects certain underlying merits.

The literature on the relation between metrics and quality is extensive (Andersen 2019, Heesen 2017) and the issue is highly debated because it requires the establishment of criteria of "good science" and raises questions on who should define them. That said, if citation metrics are used for evaluative purposes, they should ideally capture values that we deem important in science, such as the quality of scientific work.

With respect to the issue of quality, there are certain considerations that concern the providers of citation metrics and their products. Firstly, at the time of writing, the leading data analytic companies working on citation metrics are Clarivate and Elsevier (since 2004), whose databases are respectively known as the Web of Science and Scopus. These products are available as a subscription based service and typically can be consulted via a university library account. The citation metrics that Clarivate and Elsevier develop draw on databases that only include only peer-reviewed work. This fact is often used to claim that they merely report scientists' own judgments on the basis that those who cite are ultimately the scientists themselves.

However, an argument that criticizes the previous point is that citation metrics *interpret* citations as if they were a sign of distinction, even though scientists may cite for other reasons

than that, including to criticize a piece of work. Nonetheless, it might be said that scientific criticism is part of the advancement of science and that scientists build their work on what they consider to be worth improving and criticizing. That said, it is not always the case that highly cited contributions deserve praise: in fact, a work may receive attention as an example of fraud or scientific misconduct and it would be misleading to reward it and its authors, merely because it occupies a high position in citation rankings. This is to say, the gap between citation metrics and quality to some extent remains open. The content of scientific work must still be assessed on other grounds other than citation metrics, which at most help to track the amount of *use* that is made of it.

This observation brings us back to the point of using the metrics to navigate the mass of scientific literature. Scientists may use citation metrics to guide their literature searches together with a variety of other criteria to select what to read. They will ultimately be unable to judge a work only by the number of citations it has received, but in the end must judge the quality of its content. Similarly, given the multiple types and reasons for including them (scientists might cite for reasons of quality and/or usefulness, but also to raise criticisms, or for social influence) quality assessments cannot be reduced to the metrics only.

Secondly, and related to the previous point, the main competitor of Clarivate and Elsevier is Google Scholar, which has been available since 2004. Unlike the former, Google Scholar provides freely available citation rankings. Typically, it is the first search engine that scientists turn to to check their statistics, mainly because it is easy to access and read. However, it is considered by some to be less reliable than Clarivate and Elsevier, in part because it includes entries and references beyond peer-reviewed journals. Citations may also come from sources like preprint archives, presentation slides, dissertations, and also blogs, tweets, and web pages. This explains why citation counts on Google Scholar are typically higher than those from Clarivate and Elsevier, which in some cases may indicate that a scholar's work has also been picked up by non-academic sources (Andersen 2019, Delgado López-Cózar, Orduña-Malea, and Martín-Martín 2019).[1] While attracting attention from the wider world might be deemed positive, it is open to debate whether scientists should be assessed in terms of the success of their work both within and outside the scientific community. There may be disciplines or subfields whose content is so technical that it cannot be accessed by non-peers,

1. On the "democracy" of citation metrics from non-peer reviewed sources, see Heesen and Bright 2021

and scientists working in them should not be at a disadvantage because of this.[2]

A further consideration about Google Scholar is that the algorithm that it uses to rank papers is not publicly disclosed: we only know that documents are weighed in accordance with a number of variables, which also include the quantity and recency of citations (Beel and Gipp 2009). This raises the issue that if scientists cannot know how their publications are ranked, the way they are subsequently evaluated is not transparent.

A final point about the providers is that some of the metrics I discuss below typically come either from Clarivate (e.g., the IF), from Elsevier (e.g., the FWCI), or from them both along with Google Scholar (h-index). When the same metric is compiled by more than one provider, differences may be caused by the reliance on different datasets or by the methods used to calculate the indicators. Awareness of such differences is important because when such metrics determine the evaluation of a scientist's work variations in the outcome may not necessarily due to the work itself, but rather to the specifics of the index or the providers's way of calculating the metric.

With these clarifications in mind, I will now give a short description of some of the most common citation metrics in the literature, in particular the Impact Factor (IF) and some of its alternatives. For a more comprehensive analysis, see Moed (2019).

3.1 Specific Citation Metrics

The literature on the specific features of the metrics is not primarily concerned with what citations exactly track. Rather, it starts from the assumption that citations track certain properties of scientific work, whatever it may be, and then considers how to build a statistical measure, for instance the Impact Factor, that captures that property and compares it against other statistical measures, for instance the h-index.

- **Impact Factor** – The IF is the first metric that has been developed in the literature by Garfield himself. It has given rise to an entire research stream on bibliometric indicators and remains the gold standard of citation metrics, in spite of its limitations being widely discussed. Garfield's idea for the IF was to rank journals on the basis of the citations they receive over citable items, i.e., over the number of articles that they published in a given

2. A further development in the direction, are the so-called *altmetrics* indicators, which track data concerning scholarly objects from online social media platforms (Wouters, Zahedi and Costas 2019).

period of time. Garfield knew that several factors besides merit influence whether a journal will be cited, for instance the reputation of the authors or the controversiality of the topic and that such factors are difficult to express quantitatively. However, he also noted that the more articles a journal publishes in a given period of time, the higher the likelihood of that journal being cited, other things being equal (Garfield 1972). In light of this, the IF measures the number of citations that a journal receives in a certain year to papers published in the previous two years, over the total number of articles that it published in the previous two years.

One reason why Garfield picked two years as the timeframe for citations is that he observed from his database that science and technology articles typically receive the majority of their citations in the first two years after publication. To allow for some variation in time, a five-year IF is now also available. However, determining the appropriate citation window is far from trivial, for various reasons. On the one hand, the issue with shorter time spans is that they tend to favor so-called "shooting star" publications over the "sleeping beauties"; and also to favor disciplines that cite more quickly. On the other hand, the problem with a longer time span is that it includes both newer and older articles; it aggregates those whose citations may differ substantially for reasons of time rather than impact. There is extensive work on the ageing of scientific literature which aims to identify the optimum citation window and the years during which publications reach their citation peaks (see, e.g, Moed et al. 1998). The debate is still open and shows that the definition of a time period should be sensitive to different citation cultures.

The length of the citation window is in fact central to explaining why, for instance, the humanities use citation metrics much less than the natural sciences. In the humanities, a citation window of two or five years is often considered unable to capture significant citation data. This is partly because contributions in the humanities are thought to remain "valid" for a longer time than papers in the sciences. For instance, in philosophy it is standard to cite papers that are older than just two or five years, without this implying that the research is outdated. That said, whether or not an IF is available for journals in the humanities ultimately depends on the index where a journal is listed. For instance, philosophy journals that appear on the Science Citation Index or on the Social Science Citation Index (SSCI),

such as some journals in the philosophy of science, do receive an IF. Conversely, journals that belong to the Arts & Humanities Citation Index (AHCI) alone do not.[3]

Generally speaking, the IF is a journal indicator: it is not an article indicator or a scholar indicator. As such, the IF, as has been repeatedly pointed out, should not be taken as a proxy for the merit of an article: it only gives information about *the journal* in which an article appears (Osterloh and Frey 2020). A certain article may have received many citations and another none, yet they will have the same IF. This issue is particularly pressing in light of the statistical properties of the data on which citation metrics are based. As mentioned above, citation data is not uniformly distributed, but rather is highly skewed around certain clusters, in other words a few papers are cited much more frequently than the majority. This makes metrics based on simple statistical averages, such as the IF, rather unrepresentative of the merit of an article.[4] Similarly, it has been said that the IF should also not be taken as a proxy for the merit of a scholar: this is because it may be the case that in the same timeframe a scientist has one publication with a high IF and another scientist has ten with a high IF, which is something that the IF would not convey.

- **h-index** – The so-called h-index, named after its creator Jorge Hirsch, is a citation metric that couples productivity and citation record. A torrent of literature has followed its publication (Hirsch 2005), and has praised, criticized, and contributed to its refinement (see Schubert and Schubert 2019 for a review). In brief, if a scientist has a h-index of h, that scientist has at most h publications that have been cited at least h times—and no fewer times. The higher the h-index, the higher the number of publications which have all received at least h citations.

  The attractiveness of the h-index lies, among other things, in its simplicity. Even so, it is not free of problems, only two of which I will mention here. First, the h-index tends to increase with academic age and, therefore, favours senior over junior scholars. To correct for this, a five-year h-index is now available, which considers only publications in that timefame. A second problem is that if one scientist has, for example, one publication

3. Note also that different rankings than the IF exist for journals in the arts and the humanities, which may be used by scholars that need to resort to rankings (see, on this, Polonioli 2016).

4. In this respect, Frey and Osterloh (2020) report that authors benefit from the high citations received by a tiny minority of papers in a journal and suggest that this may be one of the reasons why scientists are in favor of the metrics regardless of their inaccuracies.

with a thousand citations and another scientist has one publication with one citation, both would nevertheless have an h-index of one. This is because the h-index does not convey information about the quantity of citations that a paper receives if that paper is outside the set that the h-index encompasses. These two features provide some indications of what makes the h-index problematic for the evaluation of individual scientists and of productivity generally.

- **Journal Influence Weight Indicators** – A controversial feature of the IF is that it treats all citations equally. In other words, what matters for the IF is the amount of citations that a journal receives over published items, without further qualifications. However, it has been argued that citations should not all be treated equally, since some sources are of a better quality than others. In other words, citations from more prestigious sources count more than citations from less prestigious ones. In light of this, new rankings, such as the *SCImago Journal Ranking* (SJR, based on Scopus), the *Eigenfactor* and the *Article Influence Score* (both based on Clarivate) weigh citations according to their prestige. But an issue with these metrics is that there is no independent way of measuring a journal's prestige. Thus, in order to establish whether a source is prestigious, these metrics look once again at citations. Those from the most cited journals count more than those from less cited ones. In other words, these metrics use citations recursively: prestigious journals are those that are highly cited by prestigious journals, which are those that are highly cited. Besides the issue of prestige, most of the problems that apply to the IF as an evaluative also tool apply to weighted indicators.

- **Field-Specific Indicators** – The IF order of magnitude varies considerably across fields. For instance, in philosophy of science the highest IF in 2020 was around 4, in economics 15, in psychology 24, and in medicine 90. The variation is usually attributed to differences in citation culture, time lag between publication and citations, and to the fact that not all fields are equally covered by the indexes. Because of this variation, it is plain to see that fields should not be compared on the basis of their IF alone—which is to say, without taking the disciplinary context into account.

  To better enable cross-field comparisons, field-normalized citation metrics normalize citations across scientific fields. To do this, they consider the citations that a certain journal

or publication receives in a specific time period over the citations that an average journal or publication in that field receives in that time period. A value equal to 1 indicates that the publication has been cited as often as might be expected from the literature, whereas a higher value indicates that it has received more citations than average. Alternatively, it is also possible to use percentiles and rank publications according to their standing—the top 1%, 10%, 25% and so on—in their field.

One of the most intricate aspects of this type of indicator is the operationalization of a field. Fields play a crucial role in a citation metric, and yet, drawing a line between them is far from straightforward. Some metrics rely on the classification of fields from the providers, and both the Web of Science and Scopus have their own systems. In these cases, the providers define fields on a top-down basis and link journals to such fields. The problem is that even within one field—for example, economics—there are mainstream fields, such as macroeconomics and microeconomics, and smaller fields, such as for instance economic history or history of economic thought, whose IF cannot be expected to match. And even within a mainstream field, such as macroeconomics, there are theoretical macroeconomics and applied macroeconomics whose IF also do not match, and so on.

To better accommodate this issue, other metrics, such as the Field-Weighed Citation Index, classify publications on the basis of their "similarity", where similarity is identified by means of shared citations and key terms. Once again, however, arriving at the right level of similarity is not an easy task; references are an indirect indication of a topic, and one of the crucial problems here is that papers often get classified in clusters that do not adequately match their field. All in all, while establishing fields from the top-down entails some arbitrary decisions, bottom-up classifications are to a certain extent also contestable. Field-normalized citation indexes are one of the most recent developments in the literature and are currently a subject of debate among bibliometricians (for an overview, see Waltman and van Eck 2019).

As this section makes clear, there is no such thing as *the* perfect metric: each one has certain limitations that are intrinsic to their very definition and mathematical properties. On the one hand, it might be argued that when metrics are taken together, one will compensate for some of another's shortcomings. In other words, while one metric alone is typically a poor

predictor, looking at a set of them may present a more complete, albeit imperfect, picture. On the other hand, the fact that the set of available metrics is increasingly growing in number and sophistication makes them less effective and more difficult to use. If tailor-made metrics can be given for every individual case, they cease to be indicative of the target they aim to meet, whether it is quality, impact or productivity.

More generally, some criticisms of citation metrics apply to them as a whole. One common example of this is the Matthew effect, whereby authors tend to cite papers that have already been cited (Strevens 2006). Other problems include self-citations, articles by numerous co-authors, and the type of publication, with review papers, for instance, tending to attract more citations than other types of articles. Gender and language can also skew results: male authors are cited more often than female authors (Halevi 2019) and publications in English more often than those in other languages. Some of these factors, however, can be corrected for, for instance, by excluding self-citations, or by taking account of publication type, number of co-authors, gender, and so on. That said, even the strongest advocates of citation metrics are well aware of the limitations of these measures of scientific performance and suggest ways that they can be used "responsibly" (see on this, the Leiden Manifesto by Hicks et al. 2015).

Nevertheless, one problem that remains unresolved across the range of metrics concerns the proper classification of fields that was mentioned above. As we will see in the next section, this can have serious repercussions for research programs that are evaluated alongside others when they should instead be assessed in accordance with their own standards.

4 On Gillies's analysis of research assessment

The previous sections examined citation metrics and a number of their limitations as assessment tools. This section grounds the debate about citation metrics in recent literature from the history and philosophy of science. It focuses in particular on Gillies's book *How Should Research Be Organized?*(2008), in which the author opens with the question of how a system of research assessment should be set up so that it promotes good science and encourages high-quality research.

Gillies borrows an example from statistics to show that, depending on the evaluative method we choose, we may run into two kinds of error: false positives, should the system

reward science that is in fact bad science; or false negatives, if the evaluative system fails to recognize and reward good science.

Gillies reminds us of cases such as Wittgenstein or Frege, who are examples of false negatives (or type II errors). Frege's scholarship was largely rejected by his contemporaries, even though it laid the groundwork for modern mathematical logic. Wittgenstein did not publish during his 17 years at Cambridge, during which he collected material for his *Philosophical Investigations*. Gillies also recalls that Semmelweis's research on puerperal fever was not supported by his peers, even though once it was accepted, it reduced dramatically the main cause of death of women in childbirth. Even Copernicus did not gain much acceptance from the astronomers of his time. All of these, according to Gillies, are examples of "pink diamonds", which is to say precious pieces of research that would have been lost had science been funded according to the criteria in place now.

Conversely, false positives (type I errors), occur when funding is given to flawed research that will prove unproductive. Gillies believes that the current system is skewed towards avoiding type I errors rather than type II errors.

To demonstrate this, Gillies refers to Kuhn's work on the development of science. In *The Structure of Scientific Revolutions* (1970), Kuhn famously distinguishes phases of so-called *normal science* and phases of *revolutionary science*. In normal science, scientists work within a paradigm that provides research questions and methods for problem-solving. Scientists unfold the paradigm, answer the questions that it generates, and proceed systematically as if they were solving puzzles. Phases of normal science are the typical state of science and are usually long-lasting; however, at times they are interrupted by phases of revolutionary science, in which a new paradigm emerges and replaces the previous phase.

During revolutionary science, scientists explore new research avenues; they challenge the methods and the questions of the previous paradigm and propose possible alternatives. Phases of revolutionary science are exceptional, but can lead to innovation and breakthroughs.

According to Gillies, the current evaluative system based on citation metrics, and—relatedly, on peer review—tends to favour normal science and the status quo, and to discourage revolutionary science or innovations.But is it indeed the case that the metrics protect normal science and prevent revolutionary science? To answer these questions I consider Gillies' argument in light of Kuhn's phase-model of the development of science. My main claim is that

if one considers the Kuhnian framework one of the most pressing problems with the current metrics system is that it might undermine the development of normal science.

To illustrate, it is helpful to turn to Kuhn's work *The Essential Tension: Tradition and Innovation in Scientific Research* (1959). In this, Kuhn analyses the dynamics between tradition and innovation in the development of science and discusses an apparent tension between these two factors: scientists working within a tradition tend to follow their paradigm strictly and disregard alternative explanations; and yet "the ultimate effect of this tradition-bound work has invariably been to change the tradition" (p.234) (see, on this, also Andersen 2013). The path to innovation is rooted in normal science, in other words, in the meticulous, painstaking work within a paradigm: "New theories and, to an increasing extent, novel discoveries in the mature sciences are not born *de novo*. On the contrary, they emerge from old theories and within a matrix of old beliefs about the phenomena" (p.234).

The main idea is that by doing normal science, science advances and eventually runs into an increasing set of problems—*anomalies*—that struggle to be solved within the paradigm itself. The persistent attention and concentrated effort of scientists on the paradigm in which they are working are key to acknowledging that the paradigm may not have the resources to address such problems. But it is by pursuing normal science that we pave the way for the advancement of science. Innovation is a natural step in the unfolding of normal science, and what facilitates normal science will eventually lead to revolutionary science as well.

As I argued above, however, the previous situation occurs only as long as research programs are considered in their own terms. In this regard, one harm that citation metrics might cause is that of suppressing research programs that are developing according to the standards of normal science. As argued in Sec. 3, this can happen if evaluation based on citation metrics is carried out irrespective of the differences between research programs, that is, if the metrics are not calibrated to the specificities of each field or subfields of inquiry. This goes back to the need to set a benchmark for each individual research program. Given that certain research programs, depending on their publication and citation culture, attract fewer citations than others, a policy that insists on rewarding publications whose metrics fall above a certain threshold risks ignoring research areas that do not reach that threshold. This clearly favors mainstream work over research fields that carry out high-quality research within their domains. Therefore, as long as we do not have an adequate solution to a fields' operalization problem,

there may be evaluative distortions due to classification issues.

According to Gillies, the emergence of new fields is typically discouraged by the metrics system. Nonetheless, novel and revolutionary science—in a Kuhnian sense—does emerge, provided that normal science develops to its full potential. Given that without normal science there can be no revolutionary science, the problem of protecting normal science takes priority over that of missing "pink diamonds", that is, revolutions and innovations in science.

5 Conclusions

The question of how to design institutions for scientific research is crucial for various reasons. Ideally, science-policy measures should provide an optimum framework for fulfilling scientific aims: they should support high-quality research, encourage scientific breakthroughs, and provide the conditions for scientists to excel. And yet, the question of whether they actually do so has only recently started receiving attention in the philosophy of science literature (e.g., Douglas 2010, Heesen and Bright 2017, Lee 2021, Kitcher 2003, Polonioli 2019, Shaw 2021).

This paper takes a further step in that direction by focusing on citation metrics as a science policy that is increasingly used within academia. It first considers the problem of citation metrics at a general level, then at the level of certain concrete metrics, and finally from the perspective of a Kuhnian model for the development of science.

I have argued that the uses of citation metrics for evaluative purposes and for the navigating of the literature rest on similar ground, that is on the need to come to terms with the accelerating growth of scientific production. In both cases, citation metrics work as *heuristics* that provide some guidance for the task at hand. That said, I have shown that there are some general problems with their application, and also some specific problems raised by individual metrics. Finally, I have shown that one important limitation of citation metrics emerges from the fact that we lack an adequate way to operationalize scientific fields. When looking at this problem from the perspective of Kuhn's philosophy of science, it emerges that if the metrics are used indiscriminately across fields, they may interfere with the development of normal science in some of these fields. Assuming that we do not wish to suppress promising research programs, the analysis above would seem to invite caution when using evaluative systems based on citation metrics alone.

Clearly, one question that remains open is that of what alternative we are left with, since the need for criteria with which to assess scientific work still remains. In this respect, one of the most valuable features of citation metrics is that they offer intersubjective criteria of evaluation. Indeed, when intersubjective criteria are lacking, room may be left for informal and implicit standards, for conflict of interest and for negative biases. This makes scientists vulnerable to the subjective opinions of the evaluators and highly dependent on their assessments. That said, intersubjectivity need not be achieved only by means of actual citation metrics. There is therefore no major reason that prevents the current system from being improved; there is just important work to be done in that direction.

In this respect, this paper has attempted to reveal one way in which the philosophy of science can significantly contribute to science policy analysis, by focusing on the role that citation metrics play in science. The philosophy of science brings normative considerations into the picture, and these are central to the task establishing evaluative criteria. For instance, as this paper shows, the philosophy of science provides models that identify the conditions for the advancement of science; it tells us that evaluative systems need to provide room for flexibility, since promising results can sometimes turn out to be flawed, while the relevance of others may only emerge in the future. All this testifies to the importance of philosophers of science engaging in science policy debates, lest we run the risk of overlooking counteracting factors in scientific inquiry.

References

Andersen, Hanne. "Can Scientific Knowledge Be Measured by Numbers?" In *What Is Scientific Knowledge?*, 144–159. Routledge, 2019.

———. "The second essential tension: On tradition and innovation in interdisciplinary research." *Topoi* 32, no. 1 (2013): 3–8.

Beel, Jöran, and Bela Gipp. "Google Scholar's ranking algorithm: an introductory overview." In *Proceedings of the 12th international conference on scientometrics and informetrics (ISSI'09)*, 1:230–241. Rio de Janeiro (Brazil), 2009.

Biagioli, Mario. "Quality to impact, text to metadata: Publication and evaluation in the age of metrics." *KNOW: A Journal on the Formation of Knowledge* 2, no. 2 (2018): 249–275.

Bornmann, Lutz, and Rüdiger Mutz. "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references." *Journal of the Association for Information Science and Technology* 66, no. 11 (2015): 2215–2222.

Bradford, Samuel Clement. *Documentation.* Crosby Lockwood& Son Ltd, London, 1948.

Csiszar, Alex. "Gaming Metrics Before the Game." In *Gaming the Metrics: Misconduct and Manipulation in Academic Research*, edited by Mario Biagioli and Alexandra Lippman. Cambridge: MIT Press, 2020.

Delgado López-Cózar, Emilio, Enrique Orduña-Malea, and Alberto Martín-Martín. "Google Scholar as a data source for research assessment." In *Springer handbook of science and technology indicators*, 95–127. Springer, 2019.

Douglas, Heather. "The Rightful Place of Science: Science, Values, and Democracy," 2021.

Garfield, Eugene. "Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies." *Science* 178, no. 4060 (1972): 471–479.

———. "Citation indexes for science: A new dimension in documentation through association of ideas." *Science* 122, no. 3159 (1955): 108–111.

Gillies, Donald. *How should research be organised?* College Publications. London, 2008.

Gross, Paul, and Edward Gross. "College libraries and chemical education." *Science* 66, no. 1713 (1927): 385–389.

Heesen, Remco. "Academic superstars: competent or lucky?" *Synthese* 194, no. 11 (2017): 4499–4518.

Heesen, Remco, and Liam Kofi Bright. "Is peer review a good idea?" *The British Journal for the Philosophy of Science*, 2021.

Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. "Bibliometrics: the Leiden Manifesto for research metrics." *Nature* 520, no. 7548 (2015): 429–431.

Hirsch, Jorge. "An index to quantify an individual's scientific research output." *Proceedings of the National academy of Sciences* 102, no. 46 (2005): 16569–16572.

Kessler, Maxwell Mirton. "Bibliographic coupling between scientific papers." *American documentation* 14, no. 1 (1963): 10–25.

Kuhn, Thomas. "The essential tension: tradition and innovation in scientific research." In *Scientific creativity: its recognition and development*, edited by Taylor CW and Barron F, 341–354. Wiley, New York, 1959.

———. *The structure of scientific revolutions*. Vol. 111. Chicago University of Chicago Press, 1970.

Lee, Carole J. "Certified Amplification: An Emerging Scientific Norm and Ethos." *Philosophy of Science*, 2021, 1–24.

Lotka, Alfred J. "The frequency distribution of scientific productivity." *Journal of the Washington academy of sciences* 16, no. 12 (1926): 317–323.

Margolis, J. "Citation Indexing and Evaluation of Scientific Papers: The spread of influence in populations of scientific papers may become a subject for quantitative analysis." *Science* 155, no. 3767 (1967): 1213–1219.

Moed, Henk F, Thed Van Leeuwen, and Jan Reedijk. "A new classification system to describe the ageing of scientific journals and their impact factors." *Journal of Documentation*, 1998.

Osterloh, Margit, and Bruno S Frey. "How to avoid borrowed plumes in academia." *Research Policy* 49, no. 1 (2020): 103831.

Polonioli, Andrea. "A plea for minimally biased naturalistic philosophy." *Synthese* 196, no. 9 (2019): 3841–3867.

Polonioli, Andrea. "In search of better science: on the epistemic costs of systematic reviews and the need for a pluralistic stance to literature search." *Scientometrics* 122, no. 2 (2020): 1267–1274.

———. "Metrics, flawed indicators, and the case of philosophy journals." *Scientometrics* 108, no. 2 (2016): 987–994.

Poot, Raymond, and William Mulder. *Banning journal impact factors is bad for Dutch science*. Retrieved from: https://www.timeshighereducation.com/opinion/banning-journal-impact-factors-bad-dutch-science, 2021.

Price, Derek de Solla. *Little science, big science*. Columbia University Press, 1963.

Schubert, András, and Gábor Schubert. "All along the h-index-related literature: a guided tour." In *Springer handbook of science and technology indicators*, edited by Glaenzel Wolfgang, Ulrich Schmoch H. F Moed, and Mike Thelwall, 301–334. Springer, 2019.

Shaw, Jamie. "Feyerabend's well-ordered science: how an anarchist distributes funds." *Synthese* 198, no. 1 (2021): 419–449.

Small, Henry. "Co-citation in the scientific literature: A new measure of the relationship between two documents." *Journal of the American Society for information Science* 24, no. 4 (1973): 265–269.

Strevens, Michael. "The role of the Matthew effect in science." *Studies in History and Philosophy of Science Part A* 37, no. 2 (2006): 159–170.

Svenonius, Elaine. *The intellectual foundation of information organization*. MIT press, 2000.

Van Raan, Anthony. "Measuring science: basic principles and application of advanced bibliometrics." In *Springer handbook of science and technology indicators*, edited by Glaenzel Wolfgang, Ulrich Schmoch H. F Moed, and Mike Thelwall, 237–280. Springer, 2019.

Waltman, Ludo, and Nees Jan van Eck. "Field normalization of scientometric indicators." In *Springer handbook of science and technology indicators*, edited by Glaenzel Wolfgang, Ulrich Schmoch H. F Moed, and Mike Thelwall, 281–300. Springer, 2019.

Wouters, Paul, Zohreh Zahedi, and Rodrigo Costas. "Social media metrics for new research evaluation." In *Springer handbook of science and technology indicators*, 687–713. Springer, 2019.