




ARTICLE

Citation Metrics: A Philosophy of Science Perspective

Chiara Liscandra 

Heinrich Heine University, Düsseldorf, Germany
Email: chiara.liscandra@hhu.de

(Received 12 January 2024; revised 23 July 2024; accepted 19 August 2024)

Abstract

Citation metrics are statistical measures of scientific output that draw on citation indexes. They purport to capture the impact of scientific articles and the journals in which they appear. As evaluative tools, they are mostly used in the natural sciences, but they are also acquiring an important role in the humanities. 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 literature. 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. Starting from Gillies' argument against the use of citation metrics for scientific research (2008), this paper shows that citation metrics interfere with the development of normal science in certain fields or subfields. The main issue is that citation metrics do not take field-specific differences into account, thereby selectively favoring some fields and arbitrarily hindering the development of others. In other words, this paper shows that current citation metrics fail to “carve *science* at its joints”. In light of this, the paper cautions against their use for evaluative purposes.

Keywords: Citation metrics; field classification; Thomas Kuhn; philosophy of science policy

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 the assessments of research, and they thus influence 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 comprehensive criteria (Hicks *et al.*, 2015), although such proposals have themselves attracted criticisms (Poot and Mulder 2021). Among the reasons to retain them, science policymakers maintain that, despite their imperfections, citation metrics are still one of the best tools that we currently have to measure scientific

performance. Moreover, some scientists raise concerns that if citation 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 when competing for international funding – relative to those that continue to use quantitative indicators.

In this paper, I will weigh some of the arguments for and against quantitative indicators in the evaluation of scientific research from the perspective of the philosophy of science. The analysis offered here considers whether citation metrics contribute to or else interfere with the advancement of science as this is presented in a Kuhnian framework of scientific development (Kuhn 1959, 1970, 2000). To structure the discussion, I will divide the main arguments into the theoretical (Sec. 4) and the practical (Sec. 3), focusing in particular on Donald Gillies's analysis of research assessment (2008).

One of Gillies' arguments against the use of citation metrics is that they provide a too quick assessment of scientific work, the proper evaluation of which actually requires a more extended timeframe. As an example, some metrics – for instance the Impact Factor – study a journal's 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 the relevance or impact of a scientific 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 like Wittgenstein and Frege, to name but two, would not have received proper research 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 scientific *revolutions*. 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 argument above points to certain limitations in the use of metrics as a reliable evaluative tool of scientific work. On the other hand, one of the arguments in favor of their adoption is that citation metrics respond to our need for an objective – or at least intersubjective – evaluative system. There is increased pressure from policymakers to provide evidence of scientific performance (Csiszar 2020). This evidence is then used to justify public expenditure, to strengthen the accountability of scientists, and arguably to bolster trust in science (van Raan 2019). The key point 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 said to be more subjective and prone to partiality.

In this paper, I will start by going back to the historical roots of citation analysis in order to shed light on some of its original aims and the conditions that necessitated its development (Sec. 2). I will show that, besides evaluative assessment, citation analysis introduced a new criterion with which to organize and retrieve scientific literature, one based on the *references* contained in that literature. The analysis reveals that citation analysis could be used as a tool to navigate scientific literature, thus disclosing relevant work and accelerating the flow of scientific research. In Sec. 3, however, I argue that one of the current problems with citation metrics lies in how they take field-specific differences into account. Although it is widely acknowledged that citation metrics should be normalized across fields, the literature has not yet reached a consensus on how this should be achieved. This, I will argue, is currently a limitation of citation metrics, which leads to selectively favor certain fields over others. When fields are not adequately represented by the metrics system, such fields will not receive the support they need to

fully develop their potential and the path from normal science toward revolutions is compromised.

To illustrate, in Sec. 4, I will consider Gillies's argument about the risk of overlooking "pink diamonds". I will claim that, if one endorses Kuhn's analysis on the progress of science – as it is laid out in *The Essential Tension* (1959) – the risk of protecting the status quo should not be the principal cause for concern that is raised by the metrics system. Rather, the risk of hindering the genuine development of normal science is what is at stake. Therefore, I urge caution in using citation metrics as an evaluative tool and propose that they be combined with other evaluative methods.

A more general conclusion is that, regardless of whether the system based on evaluative metrics is endorsed, these indicators deeply affect the opportunities available to research programs and their scholars. Therefore, so long as the metrics remain in place, it is in the interest of scientists to be aware of their impact and understand how they function, as well as to recognize their strengths and weaknesses.

2. The roots of citation analysis

Citation metrics are typically presented as an evaluative tool introduced to determine the allocation of resources in an increasingly competitive scientific market. In this section, I will argue that another aspect that is usually not reported is that citation metrics were also offered as a tool to organize scientific literature, one that arranges the literature according to the citations it includes.

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 citation 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 cite that publication.

One of the first indexes of academic literature is the *Science Citation Index* (SCI), now known as the *Web of Science*, which was compiled by Eugene Garfield at the end of the 1950s.¹ Garfield's index included most of the science and technology journals of his time (2,200 in 1971, a number that now stands at 21,000 if one considers the Web of Science Core Collection). Garfield's index ordered items by citations, i.e., it displayed them ranked according to the number of citations they have received.

The development of citation indexes is just one specific instance of a broader shift in various research fields that started in around the mid-twentieth century, prompted by newly available technological infrastructures and an increase in computing power. Standard examples include, for instance, the development of economic indicators such as the gross domestic product and unemployment and inequality indexes in economics (van Raan 2019). Similarly, information scientists developed measures to quantify various aspects of scientific work. These outputs responded to a variety of needs determined, for instance by the evolution of universities and academic publishing, budget constraints, an increasing demand for scientists' accountability, and the involvement of scientists in national and political agendas (Cherrier 2017).

In this context, 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

¹Before Garfield, Paul and Edward Gross (1927) compiled a *manual* citation index for the field of chemistry. 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.

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 publications more broadly. 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, co-citation analysis, and co-word analysis; and they have assisted the automatic indexing of search databases and, more recently, the development of academic 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.

With regard to policy assessment, the statistical approach to scientific production showed that one of its recurrent characteristics 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 fraction 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 used as a criterion to decide which periodicals to acquire and include in academic libraries operating under budget constraints (Gross and Gross 1927).

Besides policy assessment, however, there was a concern 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).² In response to this problem, citation analysis offered a new way of organizing and retrieving scientific literature. Previous classification systems, like the Dewey Decimal Classification System, were based on criteria such as alphabetical order and subject classification. They mostly relied on punch-cards catalogs that 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.

One instance of how citation analysis fundamentally changes the way in which scientists can search and select the literature is the so-called *forward-citation searching* strategy. This is a search option provided by citation indexes, which displays a list of all the papers that cite a particular work *after its publication*. Before its introduction, scientists were mainly limited to proceeding *backward* when searching new items, i.e., moving from one source to the references contained in that source. However, with

²He 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).

this search strategy, authors could for the first time expand their literature search beyond the references found directly in a text and look for the publications that cited that text after its publication. In this way, scholars could move *forward* in the literature and discover the stream of work triggered by a particular paper. This could, for instance, enable them to see whether the original results of the paper remained solid or if it was already outdated by more recent scientific work (Garfield 1955). This illustrates one way in which citation indexes could help researchers move through the literature more quickly and keep abreast of its more recent developments.³

In a nutshell, 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 ensure it was based on clear and shareable criteria. Citation metrics are nowadays principally associated with this evaluative role, which is also the aspect that most often stays on the radar, for political reasons.

On the other hand, the expansion of scientific production called for efficient and systematic tools to process an increasing volume of literature. In this respect, citation analysis offered a new method with which to search for and select the relevant research. The method represents a fundamental shift in the design of information organization, which now draws on digital catalogues that filter results on the basis of citations. In either case – whether it is adopted by scientists or by policymakers – citation analysis starts from the assumption that citations are “information-carriers” (Bornmann and Daniel 2006).⁴

The idea that citation analysis can be used to navigate the literature offers an initial counterargument to the claim that citation metrics tend to block the advancement of science because they protect the status quo. However, the question of whether citations enable quality assessment or information retrieval is at the heart of the entire debate between the supporters and the critics of citation metrics. To shed light on this, the next section focuses on a specific aspect of this discussion, namely the problem of differences in citation practices across fields and relatedly, the issue of field classification. I will then claim that the problem of field specificity has greater repercussions for evaluation than for information retrieval.

3. Citation metrics normalized

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

³On top of policy evaluations and literature navigation, yet another role that Garfield envisioned for the citation index was that of an “association of ideas” index (Garfield 1955). He proposed that this 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 to observe the development of ideas in science or the communication structure of a research program; in the former, however, citations work as a *literature selection device* for scientists to navigate the literature and retrieve important work from it.

⁴On the role of citations in scientific work, see also David Hull’s *Science as a Process* (1988), in particular Ch. 8, pp. 134–135. In this work, Hull defends an evolutionary account of the development of science, drawing an analogy between evolutionary processes in biology and the advancement of science. In a nutshell, the idea is that, in a similar way to how organisms behave to increase their inclusive fitness, scientists behave to increase conceptual inclusive fitness. On this view, citations represent a form of cooperative behavior in science, where citing a piece of work signals the use of that work to the scientific community.

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 the ranking of publications reflects their underlying merits.

The literature on the relation between metrics and quality (or influence, or scientists' productivity) is extensive (Andersen 2019, Heesen 2017). The issue is highly debated because it requires the establishment of criteria of "good science" and raises questions on who should define them. Technically, the main providers of citation metrics – Clarivate Analytics for the Web of Science and Elsevier for Scopus – typically sidestep this discussion by clarifying that their rankings merely report scientists' own judgments: those who *do the citing* are ultimately scientists themselves and the metrics simply leverage the scientists' own assessments. In particular, Clarivate and Scopus purportedly vet their sources via citation databases that include only peer-reviewed work. This sets them apart from providers like Google Scholar, which includes entries and references beyond peer-reviewed sources.

Nonetheless, the question remains whether citations genuinely indicate the properties they intend to capture. On this point, the literature identifies two main views on citing behavior. According to the so-called "normative theory" (see, for instance, Merton 1973), scientists cite other scientists to inform the reader about previous results, on which their own work builds, and to give credit to the authors who produced those results. Moreover, references provide the coordinates that other readers need to retrieve the sources through library catalogues. On the other hand, the "social constructivist" view (see, for instance, Knorr-Cetina 1991) maintains that the motivations behind citing behavior are complex and often depend on disciplinary, context-specific norms. Citations can be used as rhetorical devices to persuade readers – or at least some groups of them – for instance by showing that one masters the "authoritative" literature in their field. Although it is not always clear how to separate these two views neatly, typically those who defend the normative view tend to agree that citations can be used for evaluative purposes, while proponents of the social constructivist view cast doubts on this.

For instance, to support their claim, the critics highlight that citations *per se* are no signs of quality, as scientists might cite a piece of work not only to acknowledge its merit but also to criticize it or to point to it as an example of scientific misconduct. Moreover, there are certain aggregate effects that can undermine the metrics' validity as a whole. In this respect, one issue that is often discussed is the "Matthew effect" for citations, 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 (Halevi 2019). According to the advocates of citation metrics, however, most of these factors can be accounted for with appropriate corrections for self-citations, publication type, number of co-authors, language, gender, and so on. Moreover, it is often argued that, while citations can be made for a variety of reasons, they are not "randomly" distributed so that they cannot provide a background for research or give intellectual credit.

Yet one problem that remains unsolved across a range of metrics concerns differences in citation cultures, which affect the metrics themselves. In the remainder of this section, I will show that citation metrics have so far failed to account for these differences. In the next section, I will then argue that this limitation has crucial repercussions on scientific progress within a Kuhnian model of scientific development. According to Kuhn,

scientific progress requires support for normal science. In the context of evaluative systems based on citation metrics, support for normal science would require that scientific work be appropriately evaluated based on merit. However, since citation metrics fail to evaluate scientific work appropriately in certain fields or subfields, they also fail to support scientific progress in those areas.

A vivid illustration of the problem related to citation cultures comes when looking at the impact factor (IF). To see this, we need to make a brief detour to describe the basic features of the IF.

To start with, 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 (Osterloh and Frey 2020). Garfield's aim was to rank journals on the basis of the citations they received over citable items, i.e., over the number of articles that such journals published in a given period of time.

Garfield believed that several factors influence whether a journal will be cited, for instance, the reputation of the authors or the controversiality of the topic and he was aware 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 (say, e.g., 2021) to papers published in the previous two years (2019–2020), over the total number of articles that it published in those two years.

One might ask why Garfield has settled on this temporal framework. This question is particularly relevant in the context of this paper because it reveals that field-specific considerations affect the definition of some of the parameters of the metrics. To see this, Garfield picked two years as the timeframe for citations because from his database he observed 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 even within broad areas such as science and technology, for various reasons.

On the one hand, the problem with shorter time spans is that they tend to favor disciplines that cite more quickly than others; similarly, they favor so-called “shooting star” publications over the “sleeping beauties,” where the latter are publications that remain dormant for some time after publication before suddenly being discovered and attracting numerous citations (van Raan 2019). On the other hand, the problem with a longer time span is that it encompasses both newer and older articles; it aggregates those whose citations may differ substantially for reasons of time rather than impact. Extensive work on the aging of scientific literature is being carried out with the aims of identifying the optimum citation window and when citations reach their peak (see, e.g., Moed *et al.* 1998). The debate is ongoing and suggests that the definition of a time period is constrained by measurement and statistical issues that intertwine with field-specific characteristics.

The length of the citation window is in fact central to explaining why 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. Quality assessments in the humanities often eschew quantitative indicators, but in practice, whether or not an IF is available for journals in the humanities depends on the index where a journal is listed.⁵ Once a journal enters a citation index, its position in the ranking will be evaluated next to journals from fields that may have different citation cultures that have an effect on the results. Moreover, the discussion on citation windows and the recognition delay has centered mostly on the natural sciences. In this context, citation metrics respond to certain distinguishing features of the natural sciences that might not apply to the humanities or social sciences. In other words, citation metrics developed with certain fields as benchmarks do not automatically transfer to other fields.

This relates to the other point for this section, concerning the issue of the classification of fields. As soon as one compares the IF across different fields, one can observe that its order of magnitude varies considerably. 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, including time lag between publication and citations, how quickly research is published in different fields, and the fact that not all fields are covered by the indexes equally. A further point refers to *citation density*, i.e., the fact that certain fields cite more than others, resulting in differences in the average number of cited papers across fields. Because of these variations, it is clear that one should not compare fields 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. There are various ways in which this has been done, and the development of new techniques is a central area of study in scientometrics (see e.g., Waltman 2016). The driving principle here is to consider the citations that a certain journal or publication receives in a specific span of time over the citations that an average journal or publication *in that field* receives in the same period.

One of the most intricate aspects of this type of indicator is the classification of a field. Fields play a crucial role in a citation metric, and yet, the issues of how to classify them is far from straightforward, for conceptual and technical reasons. Some metrics rely on the classification of fields from the providers, and in fact both Clarivate's Web of Science (WoS) and Elsevier's Scopus have their own systems. In these cases, the providers define fields on a top-down basis and link journals to such fields. The IF relies on Clarivate's classification system, which is based on *categories* within fields (Wang and Waltman 2016): in total, the WoS has about 20 fields and 250 categories. Examples of fields include biology, chemistry, medicine, economics, and history. Examples of categories in, for instance, physics include astronomy and astrophysics, mechanics, optics, particle physics, and thermodynamics. Economics includes general economics, development studies, business, management, and urban studies, among others.

One of the problems with this approach is that even within one category – for example, general economics – there are mainstream areas, such as macroeconomics and microeconomics, and smaller areas, such as for instance economic history or the history of economic thought, whose IF cannot be expected to match. And even within a

⁵Philosophy 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. Only recently, journals that belong to the Arts & Humanities Citation Index (AHCI) have started receiving an IF too. Note also that different rankings than the IF exist for journals in the arts and the humanities, which may be used by scholars (see, on this, Polonioli 2016).

mainstream area, such as macroeconomics, there are subdivisions, like theoretical macroeconomics and applied macroeconomics, whose IF will also be different, and so on.

The literature discusses the problematic aspects of field classification extensively, highlighting the difficulties of the current approaches. For instance, a systematic study of the field of medical research (van Eck *et al.* 2013) has shown that, within a selected subset of the WoS categories, there are within-field differences in citation practices that make the use of the IF as a quality indicator inappropriate in that context. Similar conclusions have been reached in the fields of library and information science (Leydesdorff and Bornmann 2016). For an example drawn from philosophy, consider that the WoS categories include ethics, history and philosophy of science, and logic, among others. A cursory look is enough to realize that these are rather broad categories that will very likely encompass areas with different citation cultures that will affect the metrics accordingly. The upshot is that the categories used by Clarivate – but the argument also applies to Scopus (see Waltman and Wang 2016) – often fail to take into account field-specific differences in a reliable manner.

To better accommodate these issues, other metrics, such as the Field-Weighted Citation Index (FWCI from Scopus), classify publications on the basis of their “similarity,” which is identified by means of shared citations and key terms. These metrics do not rely on a predefined field classification but cluster papers in a bottom-up manner. In the case of the FWCI, 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.⁶ Once again, however, identifying the correct level of similarity is not straightforward. References and keywords are an indirect indication of a field, and one of the crucial problems here is that papers often get classified in clusters that do not adequately match their field (Janssens *et al.* 2017).

Given the difficulties with field classification, authors are exploring approaches that sidestep classifications altogether. An example is that of indicators that rely on *source* normalization: rather than comparing the citations a paper receives to those of similar papers, in this approach citations are normalized over the length of the reference list from citing papers. The advantage of this method is that it does not depend on a classification of fields since it only includes citing papers. The literature, however, is mixed on whether normalizing sources of citations outperforms normalizing cited papers over similar ones (Waltman 2016).

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). However, at the time of writing, a solution to the problem of citation metrics that allow for cross-group comparison has not yet been found.⁷ As we will see in the next section, the difficulty of accounting for differences between fields can have serious repercussions for fields that are evaluated alongside others when they should instead be assessed in accordance with their own standards.

⁶Alternatively, 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.

⁷In this article, I am not providing an argument showing that there is an in-principle issue at stake; the main point discussed here is that, more than two decades of scientometrics work has not yet led to a satisfactory solution to the problem of field-specific citation metrics.

4. Citations, normal science, and revolutions

The previous section highlights certain problems connected with field normalization and how they affect citation metrics accordingly. This section grounds the debate about citation metrics in recent literature from the 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 a concept from statistics to argue that, depending on the evaluative method we choose, we may run into two kinds of error: false positives (type I errors), should the system reward science that is in fact bad science; or false negatives (type II errors), 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. 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 occur when funding is given to flawed research that will prove unproductive. As we know from statistics, type I and type II errors typically pull in opposite directions: when one tries to reduce false positives by making the evaluative system stricter, the chance of excluding false negatives increases. Vice versa, by enlarging the pool of results, the probability of false positives increases.

Gillies believes that the current system of scientific assessment is tilted toward avoiding type I errors rather than type II errors, and that this is a mistake. His principal objection is that, given that the progress of science is largely unpredictable, the more research programs are pursued, the more likely it is that some of them will prove successful. In other words, an overly conservative approach is counterproductive in scientific contexts, where new discoveries and developments often emerge from unexpected places. This, according to Gillies, is why we should favor type II errors, as the benefits that can come from scientific discoveries largely outweigh the costs.

To support his claim, Gillies refers to Thomas 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 revolutions. 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 cumulatively and systematically as if they were solving puzzles. Phases of normal science are the typical state of science and are usually long lasting; however, they are sometimes interrupted by periods of crisis, which can lead to the emergence of a new paradigm that replaces the previous one, either partially or entirely. In phases of revolutionary science, scientists explore new research avenues, articulate possible alternatives, and new methods and questions proliferate. Revolutions 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 favor normal science and the status quo, and to discourage revolutions or innovations. This, he argues, is because citation rankings favor more traditional research over novel programs; and where this is the criterion used to

decide the allocation of funding and resources, the former benefits to the detriment of the latter.

However, is it actually the case that citation metrics overprotect normal science? After all, the argument could be challenged by the consideration that citations do not always reflect approval for a piece of work, but often also its limitations. Papers that offer objections to earlier work may pave the way for further results that advance the state of the art; and where their contributions are significant such papers will receive more citations as more and more scientists build on them. In principle, citations can help to reveal the limits of a paradigm and increase the pace at which it reaches its full potential. However, as this paper shows, one crucial problem with current citation metrics is that they can even challenge and interfere with the development of normal science.

To address this issue in more detail, I will consider Gillies' argument in light of Kuhn's phase model of the development of science, as illustrated in *The Essential Tension: Tradition and Innovation in Scientific Research* (1959). My main contention is that, within the Kuhnian framework, one of the most pressing problems with the current metrics system is not so much that citation metrics are excessively geared towards normal science and this is what might obstruct or hold back innovations, but rather that the system may undermine the development of normal science itself. In other words, it may in fact not protect normal science enough.

To illustrate, in *The Essential Tension*, 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 convincingly solved within the paradigm itself. The scientists' persistent attention to and concentrated effort on the paradigm in which they are working eventually lead them to acknowledge that the paradigm may not have the resources to address such problems. But it is by pursuing normal science that researchers pave the way for the advancement of science. Innovation is a natural step in the unfolding of normal science, and anything that facilitates normal science will eventually lead to revolutionary science as well. This is a claim that Kuhn also advances in the *Structure* (1970): "[I]t is only through normal science that the professional community of scientists succeeds, first, in exploiting the potential scope and precision of the older paradigm and, then, in isolating the difficulty through the study of which a new paradigm may emerge." (p. 152).

As I argued above, however, the previous situation in which normal science paves the way for revolutions only occurs if paradigms are considered in their own terms, that is, if paradigms are allowed to develop through the pursuit of normal science. However, due to the problems related to field operationalization discussed before, citation metrics often fail to support normal science in the first place. By not rewarding scientific work, for reasons related to the classification of such work, rather than merit, citation metrics interfere with the development of normal science in certain fields.

To see the argument again, let us focus only briefly on the notion of paradigm, or disciplinary matrix, which is central to Kuhn's work. The literature has discussed this concept extensively and partly as a reaction to the debate it generated, Kuhn has refined

it throughout his scholarship (1970, 2000). One of the main criticisms is that Kuhn uses the term in two different ways, both as a sociological and an epistemic unit (Masterman 1965). The former identifies the members of a scientific community, and the latter refers to the set of resources that the community adopts – resources such as models, laws, symbolic generalizations, values, and exemplars. According to Kuhn, however, this does not lead to a vicious circularity, where a scientific community is defined by the epistemic tools that its members adopt, and where the use of such epistemic tools defines who belongs to the scientific community. It is possible first to identify scientific communities in isolation – i.e., independently of their epistemic content – and then to explore the epistemic activities of their members, and in this way establish a link between the sociological and the epistemic side.

Kuhn leaves the task of identifying scientific communities to the sociology and history of science. He even foresees that advances in empirical techniques would at some point allow scientists to classify paradigms on the basis of citations analysis, by focusing on “communication networks including [...] the linkages among citations.” (1970, p. 178). However, as shown above, current advances in scientometrics show that it is not yet clear how to establish cohesive units on the basis of citation linkages; in particular, categories may be too broad and include several fields or subfields and thus interfere with the development of some of them. It follows that if science policymakers adopt citation metrics to provide incentives and support for scientists, they may inadvertently suppress scientific work that is developing according to the standards of normal science.

In other words, if a certain work is not assessed against the appropriate citation culture, it is not possible to infer whether its citations reflect its quality or the lack of it, as there is no reference set to which it can be compared: research could be praised or disregarded irrespective of its quality. And if the point of normal science is to develop cumulatively by building on the foundations of significant work, the metrics may simply fail to reflect the relevance of such a work. If funding is distributed according to citation metrics, scientists may be discouraged from pursuing research pathways regardless of criteria of quality that may pertain to them.

Given that some domains, depending on their citation culture, attract fewer citations than others, a policy that insists on rewarding publications whose metrics fall above a certain threshold risks ignoring research paradigms that do not reach that threshold. This clearly favors certain work over others that carry out high-quality research within their domains. Therefore, as long as we do not have adequate solutions to a fields' operationalization problem, there may be evaluative distortions due to classification issues.

Notice that this is a problem not only for citation metrics like the IF, but is one that extends to all the metrics that do not even take field-specific differences into account. Some popular metrics, such as the *h-index*, which measures the productivity of scientists, do not offer field normalization; the indicators published by Google Scholar are also not field specific. Although the analysis developed in this paper refers to specific metrics, the problem discussed here extends more broadly, as no citation metrics have thus far addressed this issue adequately.

Notice also that the problem of field classification has a greater effect on citation metrics used for evaluative purposes than citation metrics used as literature selection devices. Although some of the problems of field classification have repercussions for the results of literature searches, authors have a more active role in scrutinizing the results of their searches in order to separate relevant results from irrelevant ones. This is why, although literature search and evaluation are two sides of the same coin, the criticism set out in this paper applies more strongly to the latter than to the former.

In conclusion, 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 revolutions, the problem of protecting normal science is preliminary to 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 ideal 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 2021, Heesen and Bright 2021, Lee 2021, Kitcher 2001, Polonioli 2019, Shaw 2021).

This paper takes a further step in that direction by focusing on citation metrics as a science policy tool that is increasingly used within academia. In particular, it considers citation metrics, and the IF in particular, from the perspective of a Kuhnian model for the development of science.

First, I have argued that the uses of citation metrics for evaluative purposes and for navigating the literature rest on similar ground that is the citations included in the literature. This shows the limits of the argument according to which citation metrics tend to be conservative; on the contrary, they can in principle facilitate the spread of scientific work, disclosing its strengths and problems and thus motivating further research.

Second, the paper shows that, in order to support the development of the wide array of fields that exists in science, citation metrics should be normalized against field-specific citation practices. However, the scientometric literature is currently discussing the limitations that the current approaches to field operationalization exhibit.

All in all, when looking at citation metrics from the perspective of Kuhn’s philosophy of science, it emerges that if metrics are used indiscriminately across paradigms, they may interfere with the development of normal science in some of these fields. Assuming that we do not wish to suppress promising paradigms, 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 may 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. This shows that citation metrics may still have an important role to play and that there is evidently more work to be done in order to overcome their limitations.

This paper has attempted to discuss one way in which the philosophy of science can significantly engage with topics in science policy, 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 of 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. It also tells us that scientific

progress builds on normal science and thus, according to this model, we need to ensure that the conditions for normal science are in place. 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 H. (2013). 'The Second Essential Tension: On Tradition and Innovation in Interdisciplinary Research.' *Topoi* **32**, 3–8.
- Andersen H. (2019). 'Can Scientific Knowledge Be Measured by Numbers?' In *What is Scientific Knowledge?*, pp. 144–159. London: Routledge.
- Biagioli M. (2018). 'Quality to Impact, Text to Metadata: Publication and Evaluation in the Age of Metrics.' *KNOW: A Journal on the Formation of Knowledge* **2**, 249–275.
- Bornmann L. and Daniel H.-D. (2006). 'What do Citation Counts Measure? A Review of Studies on Citing Behavior.' *Journal of Documentation* **64**, 45–80.
- Cherrier B. (2017). 'Classifying Economics: A History of the JEL Codes.' *Journal of Economic Literature* **55**, 545–579.
- Csiszar A. (2020). 'Gaming Metrics Before the Game.' In M. Biagioli and A. Lippman (eds), *Gaming the Metrics: Misconduct and Manipulation in Academic Research*. Cambridge: MIT Press.
- de Solla Price D. (1963). *Little Science, Big Science*. New York, NY: Columbia University Press.
- Douglas H. (2021). *The Rightful Place of Science: Science, Values, and Democracy*. Tempe, AZ: Consortium for Science, Policy & Outcomes.
- Garfield E. (1955). 'Citation Indexes for Science: A New Dimension in Documentation Through Association of Ideas.' *Science* **122**, 108–111.
- Garfield E. (1972). 'Citation Analysis as a Tool in Journal Evaluation: Journals can be Ranked by Frequency and Impact of Citations for Science Policy Studies.' *Science* **178**, 471–479.
- Gillies D. (2008). *How should Research be Organised?* London: College Publications.
- Gross P. and Gross E. (1927). 'College Libraries and Chemical Education.' *Science* **66**, 385–389.
- Halevi G. (2019). Bibliometric studies on gender disparities in science. In G. Wolfgang, U. Schmoch H. F. Moed, and M. Thelwall (eds.), *Springer Handbook of Science and Technology Indicators*, pp. 563–580. New York: Springer.
- Heesen R. (2017). 'Academic Superstars: Competent or Lucky?' *Synthese* **194**, 4499–4518.
- Heesen R. and Kofi Bright L. (2021). 'Is Peer Review a Good Idea?' *The British Journal for the Philosophy of Science* **72**, 635–911.
- Hicks D, Wouters P, Waltman L, De Rijcke S. and Rafols I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, **520**(7548), 429–431.
- Hull D. (1988). *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago, IL: University of Chicago Press.
- Janssens C., Goodman M., Powell K. and Gwinn M. (2017). 'A Critical Evaluation of the Algorithm Behind the Relative Citation Ratio (RCR).' *PLoS Biology* **15**, e2002536.
- Kitcher P (2001). *Science, Truth, and Democracy*. UK: Oxford University Press.
- Knorr-Cetina K. (1991). 'Merton's Sociology of Science: The First and the Last Sociology of Science?' *Contemporary Sociology* **20**, 522–526.
- Kuhn T. (1959). 'The Essential Tension: Tradition and Innovation in Scientific Research.' In C.W. Taylor and F. Barron (eds), *Scientific Creativity: Its Recognition and Development*, pp. 341–354. New York: Wiley.
- Kuhn T. (1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Kuhn T. (2000). *The Road Since Structure: Philosophical Essays, 1970–1993, with an Autobiographical Interview*. Chicago, IL: University of Chicago Press.
- Lee C. (2021). 'Certified Amplification: An Emerging Scientific Norm and Ethos.' *Philosophy of Science* **89**, 1–24.
- Leydesdorff L. and Bornmann L. (2016). 'The Operationalization of "Fields" as WoS Subject Categories (WC s) in Evaluative Bibliometrics: The Cases of "Library and Information Science" and "Science & Technology Studies".' *Journal of the Association for Information Science and Technology* **67**, 707–714.
- Lotka A.J. (1926). 'The Frequency Distribution of Scientific Productivity.' *Journal of the Washington Academy of Sciences* **16**, 317–323.

- Margolis J.** (1967). '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**, 1213–1219.
- Masterman M.** (1965). 'The Nature of a Paradigm, from the Book: Criticism and the Growth of Knowledge.' In *Proceedings of the International Colloquium in the Philosophy of Science, London*. Cambridge: Cambridge University Press.
- Merton R.K.** (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL: University of Chicago Press.
- Moed H.F., Van Leeuwen T. and Reedijk J.** (1998). 'A New Classification System to Describe the Ageing of Scientific Journals and their Impact Factors.' *Journal of Documentation*. **54**, 387–419
- Neuhaus C. and Daniel H.-D.** (2009). 'A New Reference Standard for Citation Analysis in Chemistry and Related Fields Based on the Sections of Chemical Abstracts.' *Scientometrics* **78**, 219–229.
- Osterloh M. and Frey B.S.** (2020). 'How to Avoid Borrowed Plumes in Academia.' *Research Policy* **49**, 103831.
- Polonioli A.** (2016). 'Metrics, Flawed Indicators, and the Case of Philosophy Journals.' *Scientometrics* **108**, 987–994.
- Polonioli A.** (2019). 'A Plea for Minimally Biased Naturalistic Philosophy.' *Synthese* **196**, 3841–3867.
- Polonioli A.** (2020). 'In Search of Better Science: On the Epistemic Costs of Systematic Reviews and the Need for a Pluralistic Stance to Literature Search.' *Scientometrics* **122**, 1267–1274.
- Poot R. and Mulder W.** (2021). *Banning Journal Impact Factors is Bad for Dutch Science*. London: Times Higher Education.
- Shaw J.** (2021). 'Feyerabend's Well-Ordered Science: How an Anarchist Distributes Funds.' *Synthese* **198**, 419–449.
- Strevens M.** (2006). 'The Role of the Matthew Effect in Science.' *Studies in History and Philosophy of Science Part A* **37**, 159–170.
- Svenonius E.** (2000). *The Intellectual Foundation of Information Organization*. Cambridge, MA: MIT Press.
- Van Eck N.J., Waltman L., van Raan A., Klautz R. and Peul W.** (2013). 'Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research.' *PLoS One* **8**, e62395.
- Van Leeuwen T.N. and Calero-Medina C.** (2012). 'Redefining the Field of Economics: Improving Field Normalization for the Application of Bibliometric Techniques in the Field of Economics.' *Research Evaluation* **21**, 61–70.
- van Raan A.** (2019). 'Measuring Science: Basic Principles and Application of Advanced Bibliometrics.' In G. Wolfgang, U.S.H.F. Moed and M. Thelwall (eds), *Springer Handbook of Science and Technology Indicators*, pp. 237–280. New York, NY: Springer.
- Waltman L.** (2016). 'A Review of the Literature on Citation Impact Indicators.' *Journal of Informetrics* **10**, 365–391.
- Waltman L. and Jan van Eck N.** (2019). 'Field Normalization of Scientometric Indicators.' In G. Wolfgang, U.S.H.F. Moed and M. Thelwall (eds), *Springer Handbook of Science and Technology Indicators*, pp. 281–300. New York, NY: Springer.
- Wang Q. and Waltman L.** (2016). 'Large-Scale Analysis of the Accuracy of the Journal Classification Systems of Web of Science and Scopus.' *Journal of Informetrics* **10**(2), 347–364.