**Scientometrics manuscript No.**
(will be inserted by the editor)

# What is meaningful research and how should we measure it?

**Sven Helmer · David B. Blumenthal · Kathrin Paschen**

This is a post-peer-review, pre-copyedit version of an article published in *Scientometrics*. The final authenticated version is available online at `http://dx.doi.org/10.1007/s11192-020-03649-5`.

**Abstract** We discuss the trend towards using quantitative metrics for evaluating research. We claim that, rather than promoting meaningful research, purely metric-based research evaluation schemes potentially lead to a dystopian academic reality, leaving no space for creativity and intellectual initiative. After sketching what the future could look like if quantitative metrics are allowed to proliferate, we provide a more detailed discussion on why research is so difficult to evaluate and outline approaches for avoiding such a situation. In particular, we characterize meaningful research as an essentially contested concept and argue that quantitative metrics should always be accompanied by operationalized instructions for their proper use and continuously evaluated via feedback loops. Additionally, we analyze a dataset containing information about computer science publications and their citation history and indicate how quantitative metrics could potentially be calibrated via alternative evaluation methods such as test of time awards. Finally, we argue that, instead of overrelying on indicators, research environments should primarily be based on trust and personal responsibility.

**Keywords** research evaluation · quantitative metrics · essentially contested concepts

S. Helmer
University of Zurich
Department of Informatics
Zurich, Switzerland
E-mail: helmer@ifi.uzh.ch
ORCiD: 0000-0002-9666-1932

D. B. Blumenthal
Technical University of Munich
Chair of Experimental Bioinformatics
Freising, Germany
E-mail: david.blumenthal@wzw.tum.de
ORCiD: 0000-0001-8651-750X

K. Paschen
Nephometrics GmbH
Zurich, Switzerland
E-mail: kathrin.paschen@gmail.com

## 1 Introduction

Quantitative metrics are used for managing education, evaluating health services, and measuring employee performance in corporations, e.g. see Austin (1996). This trend does not spare academia: there is a perception among researchers that appraisal of their work is focusing more and more on quantitative metrics. Participants in a series of workshops organized by the Royal Society reported that "current measures of recognition and esteem in the academic environment were disproportionately based on quantitative metrics" such as publication and citation count, h-index, i10-index, number of PhD students graduated, and grant income (Royal Society, 2017). These numbers are used to rank individuals in job applications or promotion procedures, departments in national research quality assessments (e.g. the Research Excellence Framework (REF) in the United Kingdom, the Valutazione della Qualità della Ricerca (VQR), or Research Quality Assessment, in Italy, and the Excellence in Research for Australia (ERA) in Australia), or even universities in international ranking lists (e.g. the Times Higher Education (THE) or Quacquarelli Symonds (QS) World University Rankings). This is because research funding needs to be managed, often by people unfamiliar with the research itself. Metrics provide the necessary simplification, and they promise to be impartial, deterministic, and decision-friendly.

However, there is rising recognition that these metrics do not adequately capture research excellence and are not effective at promoting it (Royal Society, 2017). In this paper, we provide arguments for this view. In Section 2, we argue that using metrics in performance management in an area as risky and as unclearly defined as research has an impact: researchers adapt to the metrics used to manage them and this adaptation changes the research practice. For instance, the most interesting research questions often tend to be challenging and therefore involve a high risk of failure. The current system discourages researchers from tackling these questions, as the short- and mid-term rewards under a metric-based scheme are low.

Our investigation of research evaluation processes was triggered by a discussion in which we tried to imagine academia as a topic of a "Black Mirror" episode. "Black Mirror" is a dystopian science-fiction series that originated on British television in 2011, often describing scenarios in the near future employing technology that is already available or might soon be reality. According to Singh (2014) "Black Mirror" stands in the tradition of the best science-fiction works, which "study life as it is now, in our time, through the speculative lens of changing technology." We found this setting a fitting way to illustrate what being an academic could look like in a dystopian environment in the near future if the current development of relying increasingly on quantitative metrics continues. In such a scenario (depicted in Section 3), the process of conducting and publishing research would be entirely gamified. In Section 4, we look at the notion of research from the point of view of essentially contested concepts. We argue that people disagreeing on how to define high-quality research is not due to inconsistent use of of terminology or the failure to understand each other's definitions, but is an inherent property of complex and abstract concepts, such as research. This implies that we cannot expect to find an all-encompassing definition, but need to engage in an ongoing discussion. We come back to quantitative metrics in Section 5, arguing that the most widely used metrics for research quality are proxy indicators, i.e., indirect measures. We understand that it is very difficult to come up with direct measures, but it would already help if the people designing indicators are aware of this fact and put mechanisms in place to evaluate the quality of the indicators at regular intervals to see if they are fit for purpose. Additionally, we argue that using fewer performance measures and avoiding to over-regulate an environment can have beneficial effects. In Section 6 we look at concrete numbers and

analyze citation numbers for a computer science bibliography dataset. Although our findings are inconclusive, we think that contextual information, such as the test of time awards we investigate, can help in designing better indicators. We give an overview of several initiatives arguing for a more conscientious use of indicators and show how they relate to our work in Section 7. Finally, in Section 8, we conclude by recommending evaluation practices based on trusting the researchers and discussing the goals of research explicitly to improve the management of research.

## 2 Current Trends

Since evaluation metrics affect who will have a successful career in science, as well as which research gets funded, individuals and institutions adapt to these metrics, in order to receive a better evaluation. Parnas (2007) points out that measuring research output by numbers rather than peer-assessed quality has a profound impact on research itself: there is a tendency to produce more publications, as this will increase the measured values. This increase is mainly accomplished by writing shallower papers without investing a lot of time in carefully conducted research. Common techniques are doing small empirical studies based on a few observations, specifying systems and languages without actually implementing them, going for the "smallest publishable unit" (already mentioned in Broad (1981)), and working in large groups, adding the names of group members without actual contributions to publications. For instance, Ioannidis et al. (2018) have identified scientists who publish a paper every five days, calling them hyperprolific authors. Between 2001 and 2014 the number of hyperprolific authors increased from 4 to 81, while the total number of authors increased by a factor of 2.5.[1] The methods for gaming the system go even further: Weingart (2005) mentions that standards for PhD candidates have been lowered, and that there is a tendency towards unambitious but safe research proposals.

There are also cases of highly questionable behavior. Hvistendahl uncovered practices involving researchers buying author slots on papers written by others or buying papers from online brokers (Hvistendahl, 2013). Another practice is researchers turning to predatory publishers to get low-quality work published, i.e., by paying dubious publishers and bypassing the peer-review process (Beall, 2012). Plagiarism and duplicate publication is also on the rise; Steen et al. (2013) report increasing numbers for these types of offenses. Gewin (2012) discusses forging experimental data, creating fake data, and similar misconduct: according to the Thomson Reuters Web of Science database (formerly known as Web of Knowledge), 381 journal articles were retracted in 2011, up from 22 in 2001. Noorden (2011) reports a ten-fold rise in the number of retractions between 2001 and 2010 in Thomson Reuters' Web of Science, even though the number of published papers only went up by 44%.

## 3 A Dystopian Future

While attending a Re-coding Black Mirror workshop organized by Troullinou et al. (2018), we identified an underlying pattern of many "Black Mirror" episodes: take a scenario involving human interactions that is awkward, annoying, or creepy when encountered in a face-to-face situation and use technology to amplify the effect and/or scale it up to millions

[1] Publications done by large international teams, which tend to have a large number of authors, as well as Chinese and Korean authors whose names could not be disambiguated were excluded.

of people. Below, we sketch a few scenes extrapolating the impact of technology and methods from not so far in the future on evaluating research work. Let us look at the typical day of a researcher, called Dr X, in the near future.

Scene 1: At the desk with a PhD student

DR X: Morning Anna, how's your first month?
[*Dr X doesn't wait for Anna's answer.*]
DR X: So, your draft ...
ANNA: Did you like my ideas?
DR X: Well, I ran your draft through A4 ...
ANNA: Sorry, "A" what?
DR X: Academia 4.0. A tool to help you polish your paper, it's connected to a huge database of publications and their stats.
[*Turns towards monitor, points at a window on the screen.*]
DR X: Here you see some general stuff about the paper: probabilities of getting accepted at different venues, likelihood of attracting citations, just everything. And even better: it comes with a recommender system and uses machine learning to improve your stats.
ANNA: Um ...
DR X: As you can see, these numbers don't look too great at the moment. So, let's get started.
DR X: You see, when I hover over this paragraph, it suggests a few ways to increase impact. Use active voice, cut down on adverbs, that sort of thing. It can also generate text passages for you.
DR X: Here, for example, this is a pretty weak statement. "Possible to achieve a speedup"? Your work is better than that, so say it!
ANNA: Well, um, I guess ...
DR X: Well, as it stands, this section has a very low likelihood of attracting citations, see that gauge up there? Can you make the statement stronger?
[*Dr X highlights the section; a gauge at top of the page flutters briefly and settles to a low reading.*]
ANNA: So, the speedup, it's, in some cases, it's as much as 8.
DR X: Ah. So, write "we've observed a speedup by as much as one order of magnitude"!
ANNA: Eight, not ten ...
[*Dr X gives Anna a look, Anna shuts up. Dr X changes the text, highlights the modified section. The gauge at the top flutters again and settles at a higher reading.*]
DR X: See, now that's much better. Ok, ...
[*Dr X scrolls through the text, stopping at parts that the tool is highlighting.*]
DR X: ... ah, right, definitely need to be citing Prof. W's group here. Make a note of that.
ANNA: I read their work, it's not really relevant for our work.
DR X: Well, we owe them.
[*Anna makes a note on her tablet. Dr X notes a red warning on the screen.*]
DR X: I almost missed that. Need to improve your RGE factor.
ANNA: RGE ...?
DR X: Respectable graphs and equations, throw a few more fancy mathematical formulas into the text.
ANNA: I tried to keep the paper accessible...

DR X: You do want an A* paper, don't you? So, think of something.
[*Anna hesitates for a moment.*]
ANNA: Um, one more thing ... I know it isn't really related to my PhD, but I had a look at this new algorithm ...
DR X: Not now, first you have to hit your targets. Look:
[*Dr X brings up another screen.*]
DR X: This is your current likelihood of continued funding.
DR X: And besides, not hitting your target will make me and the department look bad in the next evaluation.
[*Dr X's smartphone chimes.*]
DR X: Got to go, see you later!

Scene 2: Hiring committee meeting

PROF Y: Thanks for joining me to discuss the application of Dr M. What are your opinions?
[*A group of academics sits around a table in a meeting room to discuss a job applicant. A pan around the room shows everyone using a tablet or a laptop (or both), usually with some form of dashboard showing some metrics.*]
PROF Z: I had a look at her work, seems very interesting. I also liked her presentation and she has three years of industry experience. Our students could really profit from someone like this.
DR X: That's all well and good, but A4 tells me that her h-index is rather low.
[*Dr X clicks on a button.*]
DR X: And the amount of funding she has attracted so far, let's not even go there.
PROF Z: Still, the topics she is working on...
[*Prof Y, who has been typing on his laptop, interrupts.*]
PROF Y: I just ran the numbers through A4. If we hire her, our average quality level will drop by a few points, meaning our department will lose two ranks in the national ranking.
PROF S: I double-checked our planned application to the new excellence framework with A4 yesterday. Department-wide we're short four A* publications ... Not pointing any fingers here, but I guess everyone knows their own numbers... Anyway, Dr M does not have enough of those. So this will jeopardize our application.
DR X: How come the target was 20, though? Wasn't it 15 last year?
PROF S: We opened two new researcher positions, they do come with obligations.
[*Prof Y looks at his colleagues.*]
PROF Y: So, I think you'll all agree that hiring Dr M will not work out.

Scene 3: Back at home

SPOUSE: How was your day, honey?
[*Dr X is seen at the dinner table with their partner. Dr X sighs.*]
DR X: Nothing special really ...
SPOUSE: Oh, ok, ... do you think we could ...
[*The smartphone of Dr X makes a sound.*]

DR X: Sorry, got to take this.

[*Dr X unlocks phone and selects the Academia 4.0 app. Suddenly a fanfare sounds and Dr X punches the air in triumph.*]

DR X: Yes, A4 tells me that I've been promoted to researcher level 2.4.1!

SPOUSE: Congratulations, I guess ...

DR X: I can finally apply for positions of category C3.

[*A window pops open in the app stating the targets for the next level: must have published at least ten more papers, must have published at least three more papers in A* venues, must have submitted grant applications for a total of $2,000,000, must have secured grants for a total of $500,000.*]

## 4 Meaningful Research as an Essentially Contested Concept

As exemplified in the story about Dr X, one striking feature of purely metric-based research evaluation practices is that both researchers and evaluators have less and less autonomy and personal responsibility. On the one hand, researchers are institutionally mistrusted in that they are denied the capacity to decide autonomously whether a research project is worth pursuing. On the other hand, evaluators tend to assess research by blindly checking if it fulfills a set of predefined criteria, and hence abdicate their own authority to actually judge the content of research.

We do not propose to abolish all forms of research evaluation and to let researchers pursue whichever paths they wish to follow without any accountability. The resources for funding research are limited after all, so someone has to decide which research is worth financing (Armstrong, 2012). However, we believe that before developing evaluation tools for decision-makers, we have to think about what we value. What should be the role of universities and academic research? In "The Guardian", Swain (2011) quotes Stephen Anderson, director of the Campaign for Social Science, saying that while the government has created a market economy in higher education it is not yet clear how that constantly moderated market will work. He suggests that potentially far-reaching changes are being made for reasons of financial expediency, without any thought of what their wider effect will be and goes on to state "what we are looking for is a greater vision for what the end product might look like" and asking "what is it we are all trying to do?"

### 4.1 Essentially Contested Concepts

We believe that the goal of managing research should be to make sure that meaningful research is done. However, it can be argued (Ferretti et al., 2018) that the concept of meaningful research is *essentially contested* in the sense introduced by Gallie (1955). According to Gallie, a concept is essentially contested if its "proper use [...] inevitably involves endless disputes about [its] proper uses on the part of their users" (Gallie, 1955), and he mentions the concepts of art, democracy, and social justice as typical examples. It is important to note that the root cause of the disputes does not lie in an inconsistent use of terminology or the failure of people to understand someone else's definitions, but is intrinsic to the concept. In the following, we also refer to an extensive discussion of Gallie's work by Collier et al. (2006). According to Collier et al., the goal of Gallie was "to provide a rigorous, systematic framework for analyzing contested concepts;" although this framework is sometimes discussed controversially, it offers important tools for making sense of complex concepts.

Let us now have a closer look at the framework. Gallie defines seven conditions that a concept has to satisfy to be considered essentially contested. First, it has to "accredit some kind of valued achievement." In the words of Collier et al., this means that a concept "generally implies a positive normative valence." In the case of meaningful research, we claim that it has value in itself, by being intellectually stimulating, and/or that it has a positive impact on society in the form of practically useful research, so this clearly meets the first criterion. Second, the concept has to be internally complex, i.e., it is made up of multiple components. This is also fulfilled by (meaningful) research, as it is not a simple and straightforward task, but made up of many interrelated activities. Third, the concept has to be describable in various ways, which is closely related to the second criterion: if a concept comprises many different components, it is likely that users put different emphases on these components. This also holds for research: the methodologies used in physics differ from the ones used in social sciences or literary research. Even in a single field, the approaches taken by different researchers vary, as research is made up of many different interrelated activities. Fourth, the meaning of a contested concept is open and may evolve. Gallie states that "accredited achievement must be of a kind that admits of considerable modification in the light of changing circumstances" or as Collier et al. formulate it: they are "subject to periodic revision in new situations" and that this "revision cannot be predicted in advance." Many of the ground-breaking results in research have triggered a paradigm shift, changing the way that other researchers conduct their work. Gallie states that the first four conditions are the most important and necessary ones for a concept to be called essentially contested, but that they do not provide a sufficient definition yet. He goes on to describe three additional conditions. The fifth condition asserts that different persons or groups not only have their own opinions about the correct use of a concept, but that they are aware of other uses and defend their way of doing things against these alternatives. This is true for meaningful research as well: researchers have their reasons for applying certain methodologies, or for pursuing certain theories, and will justify their choices. Gallie introduces the final two characteristics to distinguish essentially contested concepts from situations in which a dispute is caused by confusion about terminology. The sixth characteristic assumes that the concept originates from an authority, or exemplar, that is acknowledged by all the users of a concept, even by groups that disagree on its proper use. According to Collier et al. (2006), "the role of exemplars in Gallie's framework has generated much confusion. This is due, in part, to his own terminology and to inconsistencies in his presentation." In its narrow interpretation, this refers to a single, original exemplar. In its wider interpretation, Gallie asserts that it can include "a number of historically independent but sufficiently similar traditions." The important point here is that the concept under question has one underlying idea as a common core and is not rooted in multiple different ideas. We believe that this point is covered by a long list of well-known researchers and role models in history, who have employed a wide range of methodologies, some of them outdated by now, and who may have also propagated some erroneous views, but there is consensus about those researchers having conducted meaningful research that has profoundly advanced their field or even society as a whole. Finally, the seventh characteristic maintains that different uses of a concept competing against and acknowledging each other can advance and improve the use of a concept as a whole. Collier et al. also call this *progressive cooperation* and allege that this may (but does not have to) lead to an eventual decontestation by initiating more meaningful discussions. We believe that this is true of the concept of meaningful research: in order to properly understand what meaningful research is, we continuously have to engage in discussion with others and adapt our current understanding to a changing sociocultural environment.

As we can see, the notion of meaningful research satisfies all the conditions formulated by Gallie. Next, we discuss some of the implications (Gallie states them as outstanding questions in Gallie (1955)). The most obvious consequence is that once we have identified an essentially contested concept, we know that it may not be possible to find a single general principle or best use of it. However, that does not mean that it is impossible to do meaningful research on an individual level. Gallie asserts that for an individual there can very well be rational arguments to use a certain variant of a concept and even switch to another variant when the circumstances change. In an optimistic setting, a participant recognizing a concept as essentially contested allows them to respect a rival use rather than discrediting it. In turn, this could help raise a discussion about different aspects of such a concept to a higher level, acknowledging that different strands may actually help in advancing the whole concept. In a pessimistic setting, this may lead towards more aggressive behavior of a participant who may try to sideline other approaches after realizing that they cannot convince their adversaries by reasoning. This may result in a situation in which different parties revert to political campaigning to gain influence.

### 4.2 Concepts and Conceptions

There is also a different school of thought postulating that the dispute identified by Gallie is actually caused by superimposing two different meanings in the term concept. On the one hand, there is the concept itself, which is an abstract and idealized notion of something, while, on the other hand, there are different conceptions, or instantiations, of this concept. Dworkin (1972, 1978) uses "fairness" as an example and goes on to explain the important distinction between concept and conception: "members of [a] community who give instructions or set standards in the name of fairness may be doing two different things." If they ask people to treat others fairly and do not give specific and detailed instructions, they use the idealized concept and it is up to each person to decide how to actually act fairly. On the other hand, they could formulate specific instructions on how to behave fairly; Dworkin mentions the application of the utilitarian ethics of Jeremy Bentham here. In this case a particular conception of fairness is used to instruct people. Dworkin emphasizes that this is a difference "not just in the *detail* of the instructions given but in the *kind* of instructions given." In the case of invoking the concept of fairness, the instructor does not attach particular value to their own views, i.e., they do not deem their views to be superior. When specifying a particular conception of fairness, this sends out the signal that the instructor believes that their views have a higher standing. Dworkin argues that "when I appeal to fairness I pose a moral issue; when I lay down my conception of fairness I try to answer it."

Criley (2007) proposes that "a concept F simply is a cluster of norms providing standards for the correct employment of the corresponding linguistic term *F*." This raises the question which norms belong to or are part of a concept? For an essentially contested concept, this question can of course never be answered conclusively; otherwise, the concept would not be essentially contested. On the contrary, it is crucial for the understanding of essentially contested concepts that they always involve a normative dimension which cannot be laid down into fixed principles and guidelines. According to Criley, a conception is also a cluster of norms. However, in contrast to a concept the provided norms are much more concrete, resolving some of the vagueness or conflicts found in concept clusters, even up to the point of stating explicit rules. Criley (2007) goes on to discuss the relationship between concepts and conceptions: "Notice, however, that the point of having a distinction between concepts and conceptions becomes much clearer once we turn our attention to the possibility of distinct

rival candidate conceptions of a concept. If we focus exclusively on those concepts that have a single, uncontroversial, determinate conception that is implicit in any thinker who is competent with respect to that concept, then it becomes hard to see the importance of a distinction between concepts and conceptions."

We find this observation particularly useful when trying to understand the concept of meaningful research. As we have already seen in the last section, the concept of meaningful research is multi-faceted and highly complex. In their work, researchers are striving for something that cannot be defined conclusively, as meaningful research is an essentially contested concept. In essence, researchers follow different conceptions of meaningful research, depending on their research area and their personal experience. We would even expect a certain degree of rivalry between groups using different conceptions. Also, individual researchers might switch from one conception to another during their careers. Following the interpretation of Dworkin, encouraging researchers to do meaningful research by invoking the abstract concept hands the responsibility on how to achieve this to the individual researchers. If a certain conception of meaningful is laid down, though, to describe what high-quality research looks like, then this particular conception is enforced as a standard, pushing researchers into a certain direction.

### 4.3 Multidimensionality of Meaningful Research

In a series of studies, meaningful, high-quality, or excellent research is characterized as a multidimensional concept. In particular, various empirical studies (Aksnes and Rip, 2009; Bazeley, 2010; Hug et al., 2013; Mårtensson et al., 2016) have revealed that "researchers' conceptions of research quality [include] a multitude of notions [which] span from correctness, rigor, clarity, productivity, recognition, novelty, beauty, significance, autonomy, difficulty, and relevance to ethical/sustainable research" (Aksnes et al., 2019). Which of these dimensions of quality or meaningfulness is predominant in a specific assessment of research quality depends of the context of the assessment. Moreover, characterizing the concept of meaningful research as multidimensional implies that it has no abstract meaning detached of the specific dimensions. Rather, it should be viewed as a "boundary object that [...] offers some ground for constructive discussion via a shared framework" (Hellström, 2011).

This characterizing of meaningful research as multidimensional nicely fits within the conceptual framework developed in the previous sections: since meaningful research is essentially contested, researchers in practice adhere to different conceptions of meaningful research, which, in turn, emphasize and deemphasize different dimensions of research quality. The concept of meaningful research hence becomes meaningless if we abstract from these dimensions. However, characterizing meaningful research as essentially contested also implies that it is more than an umbrella term for the various, often pairwise incompatible dimensions of research quality. Rather, it entails that researchers following a specific conception of meaningful research must in principle be willing to defend their conception in discussion with others. Or put differently: a researcher who states that they simply follow their conception of meaningful research and does not care about what the community says or thinks about it is no longer involved in the progressive cooperation of doing research.

### 4.4 A Conception of Meaningful Research

As a basis for discussion, let us briefly give a, necessarily inconclusive, description of our conception of meaningful research. For a start, we would like researchers to work on questions the answers to which would help solve major problems faced by society. Also, we would like researchers to attempt to solve the really challenging problems and not get sidetracked by minor details that have no or very low impact. We call such research *practically useful*. Identifying meaningful with practically useful research would be too narrow, though. It would rule out a lot of fundamental research that helps us gain a deeper understanding of a field. Someone reading about fundamental research of this kind should be intellectually stimulated or conceptually enriched by it. We call this research *stimulating*. Ideally, we would like research to be stimulating also for readers from a different research community, otherwise we could end up with closed research communities who develop very esoteric questions that thrill their members but are irrelevant or even unintelligible to the rest of the world. So, from our point of view, a promising starting point for characterizing meaningful research could be to require that it should be stimulating or practically useful.

Note that this description leaves out important dimensions such as ethics (according to our characterization, the Manhattan Project or a project on drone warfare would be classified as meaningful), as well as methodology (we assume that researchers adhere to sound and scientific methods). Moreover, recall again that our characterization of meaningful research as practically useful or stimulating should not be misread as an attempt at defining meaningful research. Rather, it should be understood as a starting point for a discussion in which stakeholders should engage continuously.

After discussing meaningful research from a theoretical point of view, we now turn to practical aspects of evaluating research, namely the use of metrics, which we view as proxy indicators, and the issues associated with them.

## 5 Quantitative Metrics as Proxy Indicators

Metrics are quantitative measures used for evaluating work artifacts such as academic publications. We focus on publication- and citation-based metrics, which are very common, see Aagaard et al. (2015). We argue that these metrics are *proxy indicators* – indirect measures – which need to be applied and interpreted carefully, and propose a *feedback loop* approach.

If we are right in characterizing meaningful research as essentially contested, then there cannot be a metric that captures it. A set of reasonably good proxy indicators is the best we can get. Moreover, proxy indicators target a particular conception instead of measuring the general concept. Trying to explicitly formulate the conception that forms the basis of an indicator adds context and helps in understanding what we are actually measuring. This also brings hidden conflicts between different conceptions into the open, and can be used as a starting point for discussions.

Publication- and citation-based metrics can be weighted based on the type of publication (journal article, book chapter, monograph, etc.), a quality rating of the publication venue (which can be decided by committee or taken from a trusted source such as the ISI Web of Science by Thomson Reuters), and the number of authors. Different weighting schemes are in use, and a lot has been written about how to implement them fairly, for instance by Aagaard et al. (2015) and Piro et al. (2013). It is plausible for publication and citation-based metrics to be positively correlated with other metrics targeting research quality, and indeed

this correlation has been shown to exist, e.g. by Jarwal et al. (2009). However, that study also showed that while there is a correlation, the bibliometric values they studied did not account for the full variance of the quality ratings assigned by independent peers. Michels and Schmoch (2014) have shown that citation counts are generally lower for journals published in non-English languages. In certain areas, such as linguistics and cultural studies, it makes perfect sense to publish in another language, but this does not say anything about the quality of a publication. Additionally, as Nygaard and Bellanova (2017) point out, publication metrics crucially depend on the criteria for deciding what counts as a scientific publication and its value. For example, among computer science researchers conference papers, especially those published in top-tier conferences, are considered to be on par with journal papers, whereas designers and artists are often evaluated by their portfolios and not their publications. Numerous other studies have complained that bibliometrics do not capture research quality fully or fairly (see Aksnes et al. (2019) as well as Grimson (2014) and papers cited there), and moreover, that bibliometric processes influence the object of their measurements. Section 2 and Michels and Schmoch (2014) discuss that effect.

The considerations above underline that metrics should be chosen and applied carefully. But what does this mean? First, we need to specify what we want to measure. It is very hard to define a metric if the goals we want to achieve are vague and unclear. For instance, in their report about the development of indicators for research excellence by the European Commission, Ferretti et al. (2018) highlight that, when asked to define research excellence, many stakeholders were not able to give an answer. This is not an ideal starting point for choosing or defining metrics. We also have to be aware that often metrics are applied by administrators or scientists rather than bibliometricists. This is sometimes called *citizen bibliometrics* (Leydesdorff et al., 2016). According to Hammarfelt and Rushforth (2017), a considerable number of citizen bibliometricists are aware of the shortcomings, and apply metrics thoughtfully. However, we believe that they could get better operational support, e.g. in the form of training, frameworks, and documented best practices. We agree with Wang and Schneider (2020), who, in the context of interdisciplinary measures, argue that "the operationalization of interdisciplinary measures in scientometric studies is relatively chaotic."

Our recommendation is to evaluate quantitative metrics periodically to ensure they are well chosen to meet requirements, and that they are being applied correctly. This is the feedback loop idea formulated by O'Neil (2016), who argues that we need to evaluate an algorithmic evaluation process itself from time to time to check if it is (still) fit for purpose by comparing metric outcomes with other assessment techniques. This allows us to notice when metric results are wrong and to take corrective action. The idea of a feedback loop also appears in the context of concepts and conceptions: Criley (2007) calls this *reflective equilibrium*, which is "a method for inducing or restoring coherence between general principles and particular judgments through a process of *mutual adjustment* of the conflicting principles and particular judgments." Proposing a feedback loop raises the question: how do we evaluate proxy indicators against a notion of quality? This implies we need an alternative way of assessing how well given research work matches our conception of research quality. This is difficult; some authors use independent peer review (e.g. Jarwal et al. (2009)) but of course peer review itself is a fraught metric (see Brezis and Birukou (2020), Krummel et al. (2019)). We investigate test of time awards in Section 6; in some fields, retractions may also provide a useful signal. Rafols et al. (2012) speak out for indicators that do not reduce the quality of research to a single number, but provide contrasting perspectives. They call this *opening up* the decision-making process and argue that these indicators should be embedded into an assessment or policy context, so they can be used to interpret the data and

not as a substitute for judgment. We believe that such indicators could also be helpful in implementing a feedback loop.

Moreover, we advocate more personal responsibility for researchers, creating space for them to apply their own intellectual judgment. Luhmann (2017) states that the two most important concepts making complex social systems feasible are trust and power. It follows that the alternative to putting faith in researchers is to have an authority deciding unilaterally and controlling the researchers' actions. However, Pollitt (1993) points out that treating staff as "work units to be incentivized and measured" instead of as "people to be encouraged and developed" will lead to demoralized and demotivated employees by taking away their intrinsic motivation (cf. also Shore and Wright, 2000). In their review on the evolution of performance management, Pulakos et al. (2019) come to the conclusion that "formal performance management processes disengage employees, cost millions, and have no impact on performance." In the context of taxpayer honesty, Kucher and Götte (1998) have shown that observing tax laws and regulations is not just a matter of how strictly taxpayers are controlled. When taxpayers have trust in a government and in return are trusted by the government, taxpayers feel more obliged to follow the rules. Additionally, when people have the possibility to participate in a decision process and have some control over the outcomes, they are more willing to accept these decisions and outcomes. We believe that leveraging intrinsic motivation and allowing researchers to act more independently is the way to go, since demotivating staff will hardly result in meaningful research. Dance (2017) reports that a number of academics have already taken matters into their own hands by not joining traditional academia at all or leaving and working as independent researchers.

## 6 An Analysis of Quantitative Publication Metrics

We want to round off our work by taking a closer look at some concrete numbers, i.e., analyze citation numbers and investigate alternative evaluation methods. As two authors belong to the computer science research community, we decided to use the data set provided by the dblp computer science bibliography originally created by Ley (2009), which tracks all major computer science publications. At the time of writing, dblp indexed over 4.7 million publications. For our analysis, we used a version with citation numbers created for ArnetMiner by Tang et al. (2008) containing more than 3 million publications.[2] In a first step, we did some exploratory data analysis to get a better feel for the data and to check its consistency. Figure 1 shows the average and maximum citation counts for papers arranged by year since their publication. The older papers have a higher average citation count, but there is some bias here. First, they have been around longer, which means they had more opportunities to be cited and, second, there are fewer older papers, so outliers have more influence. The different number of papers is also illustrated in the graph for average citation counts: it is very smooth on the left-hand side and gets more and more ragged towards the right, as the number of papers in the denominator used for computing the average decreases with the years. Additionally, dblp did not collect old papers as systematically as newer ones, so it is likely that only influential old papers made it into the collection. While many papers reach their maximum citation count in the first ten years after publication[3], there is a considerable number of papers that do not. Most evaluation schemes, such as REF and VQR, only look at

[2] DBLP-Citation-network V10 is at `https://aminer.org/citation`, the Jupyter notebooks used for the analysis can be found at `https://github.com/kpaschen/spark/tree/master/dblp/jupyter`

[3] Please note that the average and maximum citation counts use different scales. Clearly, the maximum count, which reaches ten for 80 years since publication, is always greater than the average count.

papers published in the last five to seven years, though, which means that they may overlook important publications. What we do not show in a graph here is the (unsurprising) fact that the citations are not uniformly distributed: a small number of papers get the most citations.

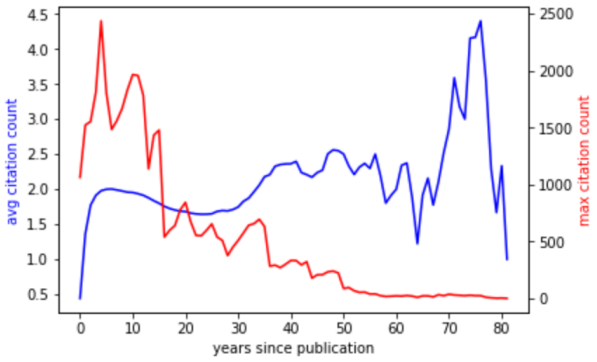


**Fig. 1** Average versus maximum citations

For the next step, we asked ourselves how we could analyze the suitability of citations for determining the quality of publications and look at an alternative. The validation of metrics performed in Jarwal et al. (2009) uses independent peer reviews for this purpose. This is useful but labor intensive; Tennant (2018) estimates that more than 2.5 million English language research papers are published annually and that the rate is still increasing. Considering that usually two to three reviewers are needed for each review and that the number above reflects the number of accepted publications, i.e., the number of submissions going through the process is even higher, this puts an increasing burden on reviewers (sometimes referred to as *reviewer fatigue*; see Breuning et al. (2015); Fox et al. (2017)). Not only is there a lack of time: Tennant (2018) goes on to state that there is no proper incentivization for reviewers to do a good job. Moreover, even if a peer review is done to a high standard, some research needs time to be appreciated; a peer review performed now may come to different conclusions than one performed five or ten years later.

A number of publication venues have introduced *test of time awards*, which retrospectively identify important and influential work done ten or even twenty years ago. We found this an interesting approach by the scientific community to acknowledge quality in hindsight.[4] For our study, we selected the Test of Time Award of the International Conference on Very Large Data Bases (VLDB), which is a prestigious publication venue in the area of database research. Figure 2 shows the citation counts for the award winners compared to the average citation count of all VLDB papers.

While award winners tend to be cited more often than the average VLDB paper, there are also award winners whose citation numbers are below the average. This shows that citation numbers are not the only criterion used by the committee for selecting the winners, otherwise the papers with the lower citation counts would not have received the award. This is also reflected by the number of times that the award winners were cited after receiving the award; for VLDB the award is usually given ten to twelve years after publication and many papers still have a considerable number of citations after this period.

[4]Note that we are still using publications as a proxy for scientific accomplishment here.

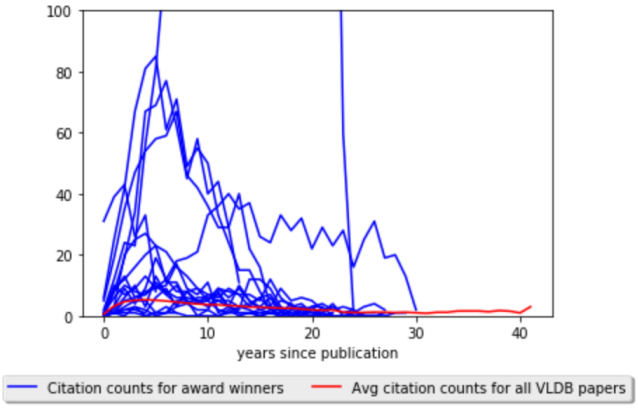


**Fig. 2** VLDB Test of Time Award winners versus all VLDB papers. One paper has so many citations that including its peak would make the bottom third of the graph unreadable.

While not conclusive, we believe this data supports the hypothesis that citation counts are not a complete measure of research quality on their own. We think that other indicators, such as test of time awards, should be investigated and that this would be a promising area for future research. However, since these awards provide recognition for research with a time delay, there are many contexts where they cannot be applied directly. They can be used for calibrating other proxy indicators, though.

## 7 Related Initiatives

We are by far not the only ones criticizing the effects of quantitative metrics on research and calling for improvements in the area of research evaluation. For instance, there are the San Francisco Declaration on Research Assessment (DORA) (DORA, 2012), the Leiden Manifesto for research metrics (Hicks et al., 2015), and the Metric Tide report (Wilsdon et al., 2015).

While we agree with the sentiments expressed in DORA, we want to focus on the operationalization of research metrics and we believe this to align with arguments made in the Leiden Manifesto and the Metric Tide report. The authors of the Leiden Manifesto call for open and transparent high-quality processes in the context of decision-making, part of which are the regular scrutinization of indicators and their improvement if they are found lacking. This is in line with the feedback loop we are proposing. Other important points are that quantitative assessment should go hand-in-hand with qualitative evaluation and that it should consist of a suite of indicators to reflect different aspects of research. Similar points were made in the Metric Tide report and its notion of *responsible metrics*. The report noted, that "there is potential for the scientometrics community to play a more strategic role in informing how quantitative indicators are used across the research system and by policymakers." We think that a useful next step toward operationalizing bibliographic metrics is to propose and evaluate feedback loop mechanisms. These mechanisms should not only consist of quantitative evaluation methods, but also include qualitative ones.

## 8 Conclusions and Outlook

Applying purely metric-based indicators for evaluating research has the potential to lead to a dystopian research environment in the style of a "Black Mirror" episode. There is a hidden danger in just accepting certain indicators, such as citation numbers in Google Scholar or Scopus, because they are convenient. Many researchers are left with the impression that important decisions are taken over their heads and that they have no say in what is happening. This situation is at risk of devolving into a low-trust environment, which cannot be in anyone's interest. In our view, over-reliance on indicators also hands too much power to the institutions managing these indicators and makes it difficult to introduce changes, locking in a particular conception of research. Stilgoe (2014) expressed this provocatively: "[research] excellence tells us nothing about how important the science is and everything about who decides."

Given the diversity of academic disciplines, it is difficult to come up with a set of universally applicable methods for assessing research. Nevertheless, the question of what we want research to achieve has to be asked and discussed explicitly. Over the course of history, the roles that universities played have already changed before: for example, in the nineteenth century they were transformed from educational institutions to organizations that also pursued research (Willetts, 2017). It seems that at the moment universities are undergoing another transformation, but it is far from clear where they are heading (Swain, 2011).

Moreover, the discussion we are asking for is not a one-off event: because research is an essentially contested concept and the environment is constantly changing, this needs to be an ongoing public debate. As we expect quantitative metrics to stay with us as an evaluation tool for some time to come, we at least need more transparency in how they are created and which policies drive them.

## Acknowledgements

We would like to thank an anonymous reviewer for very helpful comments and many pointers to relevant literature.

## References

Aagaard, K., C. Bloch, and J. W. Schneider (2015). Impacts of performance-based research funding systems: the case of the Norwegian publication indicator. *Research Evaluation 24*, 106–117.

Aksnes, D. W., L. Langfeldt, and P. Wouters (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open 9*(1), 1–17.

Aksnes, D. W. and A. Rip (2009). Researchers' perceptions of citations. *Research Policy 38*(6), 895–905.

Armstrong, J. (2012). A question universities need to answer: Why do we research? `https://theconversation.com/a-question-universities-need-to-answer-why-do-we-research-6230`. [Online; accessed September 2018].

Austin, R. D. (1996). *Measuring and Managing Performance in Organizations*. New York: Dorset House Publishing.

Bazeley, P. (2010). Conceptualising research performance. *Studies in Higher Education 35*(8), 889–903.

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature News 489*(7415), 179.

Breuning, M., J. Backstrom, J. Brannon, B. I. Gross, and M. Widmeier (2015). Reviewer fatigue? why scholars decline to review their peers' work. *Political Science and Politics 48*(4), 595–600.

Brezis, E. S. and A. Birukou (2020). Arbitrariness in the peer review process. *Scientometrics 123*, 393–411.

Broad, W. J. (1981). The publishing game: Getting more for less. *Science 211*(4487), 1137–1139.

Collier, D., F. Daniel Hidalgo, and A. Olivia Maciuceanu (2006). Essentially contested concepts: Debates and applications. *Journal of Political Ideologies 11*(3), 211–246.

Criley, M. E. (2007). *Contested Concepts and Competing Conceptions*. Ph. D. thesis, University of Pittsburgh.

Dance, A. (2017). Flexible working: Solo scientist. *Nature 543*, 747–749.

DORA (2012). San Francisco declaration on research assessment. `https://sfdora.org/`. [Online; accessed September 2018].

Dworkin, R. M. (1972). The jurisprudence of Richard Nixon. *The New York Review of Books 18*, 27–35.

Dworkin, R. M. (1978). *Taking Rights Seriously: New Impression with a Reply to Critics*. Oxford: Duckworth.

Ferretti, F., Â. G. Pereira, D. Vértesy, and S. Hardeman (2018). Research excellence indicators: time to reimagine the 'making of'? *Science and Public Policy 45*(5), 1–11.

Fox, C. W., A. Y. K. Albert, and T. H. Vines (2017). Recruitment of reviewers is becoming harder at some journals: a test of the influence of reviewer fatigue at six journals in ecology and evolution. *Research Integrity and Peer Review 2*(3).

Gallie, W. B. (1955). Essentially contested concepts. *Proceedings of the Aristotelian Society 56*, 167–198.

Gewin, V. (2012). Research: Uncovering misconduct. *Nature 485*, 137–139.

Grimson, J. (2014). Measuring research impact: not everything that can be counted counts, and not everything that counts can be counted. In W. Blockmans, L. Engwall, and D. Weaire (Eds.), *Bibliometrics. Use and Abuse in the Review of Research Performance*, Volume 87 of *Wenner-Gren International Series*, pp. 29–41. Portland Press.

Hammarfelt, B. and A. D. Rushforth (2017). Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation. *Research Evaluation 26*(3), 169–180.

Hellström, T. (2011). Homing in on excellence: Dimensions of appraisal in center of excellence program evaluations. *Evaluation 17*(2), 117–131.

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature 520*, 429–431.

Hug, S. E., M. Ochsner, and H.-D. Daniel (2013). Criteria for assessing research quality in the humanities: a Delphi study among scholars of English literature, German literature and art history. *Research Evaluation 22*(5), 369–383.

Hvistendahl, M. (2013). China's publication bazaar. *Science 342*(6162), 1035–1039.

Ioannidis, J. P. A., R. Klavans, and K. W. Boyack (2018). Thousands of scientists publish a paper every five days. *Nature 561*, 167–169.

Jarwal, S. D., A. M. Brion, and M. L. King (2009). Measuring research quality using the journal impact factor, citations and 'ranked journals': blunt instruments or inspired metrics? *Journal of Higher Education Policy and Management 31*, 289–300.

Krummel, M., C. Blish, M. Kuhns, K. Cadwell, A. Oberst, A. Goldrath, K. M. Ansel, H. Chi, R. O'Connell, E. J. Wherry, and M. Pepper (2019). Universal principled review: A community-driven method to improve peer review. *Cell 179*, 1441–1445.

Kucher, M. and L. Götte (1998). Trust me – an empirical analysis of taxpayer honesty. *Finanzarchiv 55*(3), 429–444.

Ley, M. (2009). DBLP: Some lessons learned. *Proc. VLDB Endow. 2*(2), 1493–1500.

Leydesdorff, L., P. Wouters, and L. Bornmann (2016). Professional and citizen bibliometrics: complementarities and ambivalences in the development and use of indicators—a state-of-the-art report. *Scientometrics 109*, 2129–2150.

Luhmann, N. (2017). *Trust and Power*. Cambridge: Polity.

Mårtensson, P., U. Fors, S.-B. Wallin, U. Zander, and G. H. Nilsson (2016). Evaluating research: A multidisciplinary approach to assessing research practice and quality. *Research Policy 45*(3), 593–603.

Michels, C. and U. Schmoch (2014). Impact of bibliometric studies on the publication behaviour of authors. *Scientometrics 98*, 369–385.

Noorden, R. V. (2011). Science publishing: The trouble with retractions. *Nature 478*, 26–28.

Nygaard, L. P. and R. Bellanova (2017). Lost in quantification: Scholars and the politics of bibliometrics. In M. J. Curry and T. Lillis (Eds.), *Global Academic Publishing: Policies, Perspectives and Pedagogies*, pp. 23–36. Bristol: Multilingual Matters.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Parnas, D. L. (2007). Stop the numbers game. *Commun. ACM 50*(11), 19–21.

Piro, F. N., D. W. Aksnes, and K. Rørstad (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology 64*, 307–320.

Pollitt, C. (1993). *Managerialism and the Public Services: Cuts or Cultural Change in the 1990s?* Oxford: Blackwell.

Pulakos, E. D., R. Mueller-Hanson, and S. Arad (2019). The evolution of performance management: Searching for value. *Annual Review of Organizational Psychology and Organizational Behavior 6*(1), 249–271.

Rafols, I., T. Ciarli, P. van Zwanenberg, and A. Stirling (2012). Towards indicators for 'opening up' science and technology policy. In *The Internet, Policy & Politics Conference 2012*, Oxford, UK.

Royal Society (2017). Research culture embedding inclusive excellence. `https://royalsociety.org/~/media/policy/Publications/2018/research-culture-workshop-report.pdf`. [Online; accessed September 2018].

Shore, C. and S. Wright (2000). *Coercive accountability: the rise of audit culture in higher education*, pp. 57–89. London: Routledge.

Singh, G. (2014). Recognition and the image of mastery as themes in Black Mirror (channel 4, 2011-present): an eco-jungian approach to 'always on' culture. *Int. Journal of Jungian Studies 6*, 120–132.

Steen, R. G., A. Casadevall, and F. C. Fang (2013). Why has the number of scientific retractions increased? *PLOS ONE 8*(7), e68397:1–9.

Stilgoe, J. (2014). Against excellence. `https://www.theguardian.com/science/political-science/2014/dec/19/against-excellence`. [Online; accessed August 2019].

Swain, H. (2011). What are universities for? `https://www.theguardian.com/education/2011/oct/10/higher-education-purpose`. [Online; accessed September 2018].

Tang, J., J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proc. of the 14th ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD'08)*, Las Vegas, Nevada, pp. 990–998.

Tennant, J. P. (2018). The state of the art in peer review. *FEMS Microbiology Letters 365*(19).

Troullinou, P., M. d'Aquin, and I. Tiddi (2018). Re-coding black mirror chairs' welcome & organization. In *Companion of the The Web Conference WWW'18*, Lyon, France, pp. 1527–1528.

Wang, Q. and J. W. Schneider (2020). Consistency and validity of interdisciplinary measures. *Quantitative Science Studies 1*(1), 239–263.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics 62*(1), 117–131.

Willetts, D. (2017). *A University Education*. New York: Oxford University Press.

Wilsdon, J., L. Allen, E. Belfiore, P. Campbell, S. H. Stephen Curry, R. Jones, R. Kain, S. Kerridge, M. Thelwall, I. V. Jane Tinkler, P. Wouters, J. Hill, and B. Johnson (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE.