Contents lists available at ScienceDirect

Rhizosphere

journal homepage: www.elsevier.com/locate/rhisph

Book review

The essence of university research ranking and bibliometrics.

Abstract

A book review of “Bibliometrics and Research Evaluation – uses and abuses” by Yves Gingras, 2016, MIT Press, Cambridge Mass. This is a summary of what I thought were the core ideas presented in this book, occasionally padded with my own comments. The main three metrics used in the evaluation of research impact that are the focus of the essay are the H-index (or Hirsch index), journal impact factors, and global university rankings. I do not describe the book by chapter, or even in the same order concepts are discussed. Instead each section might be a condensate of ideas that occur in various places or throughout. The point was to create a concentrated version to deliver the essence of the thoughts more fully presented by Yves Gingras in his book.

Keywords

Bibliometrics; H-index; Scientometrics; University Ranking; Webometrics

1. Introduction

The popularization of the big three (the H-index, journal impact factors, and global university rankings) in evaluation and ranking of academic research output has its roots in the 1980s with the implementation of the “New Public Management” ideology. There has been a multiplication of indicators of excellence to cater to this new found wisdom, but what do these actually mean in the context of academic research in universities? This is the task set-out for this short book by Gingras (2016), a sociologist, historian, and scientific director at the Observatoire des Sciences et Technologies at the Université du Québec à Montreal.

The book begins with a concise but precise historical sketch of the development of scientometric and bibliometric measures, followed by three chapters called: What Bibliometrics Teaches us about the Dynamics of Science; The Proliferation of Research Evaluation; and The Evaluation of Research Evaluation. Zitt (2015) provided a chapter by chapter summary and commentary in his 5 pages book review of the French version that I will not repeat here.

2. Origins and criticism

The uses of bibliometric evaluation and ranking of research output by university and government administrations has proliferated over the last 15 years, but has been common for almost fifty years, and have their modern origins in the post-World War II period. It is driven by the very reasonable desire to go beyond anecdotes, to inform and enlighten choices and spending priorities. On the one hand researchers of a more competitive nature embraced these indicators, while more thoughtful ones also criticized their misuse and inadequacy for the task. This book argues that in fact due to their shortcomings, and supported by numerous studies, the commonly-used indicators have no scientific validity and rarely measure what they set-out to do (p VIII).

This perspective is fully in agreement with the countless and, as the author remarks, now often redundant blogs, letters to editors, and papers that restate these shortcomings and the statistical failures of the common indicators for evaluation and for ranking the performance of universities, departments, and individuals. There is an ill-defined relation between a concept to be measured and its indicators, and also of scale at which the indicator is valid. Some metrics are meant for coarse evaluation of aggregated data but instead are frequently applied to individual level evaluation. Others require a several-years window to be measured but are calculated on very short time scales. As Gingras and most people recognize, the critics pointing to limits and counter-productive consequences are numerous, while those that dug deeply enough statistically to evaluate “their epistemological foundation” are few (see Beauvallet, 2009). In his summary and commentary on the book, Zitt (2015) remarks accurately that in this book Gingras offers a “committed vision, albeit one without nuances”. He buttresses the commentary by providing a well selected list of more than 50 important papers on the topic that would save any reader new to this issue a lot of time. Two additional useful references are the book of indicators compiled by Todeschini and Baccini (2016) that describes about 150 bibliometric indicators alphabetically, and the review of 108 author-level metrics by Wildgaard et al. (2014).

Gingras is in fact quite fair in pointing to the “lack of serious methodological reflection that has led to the anarchic use of bibliometric indicators in research evaluation” (p VIII) on the one hand, while pointing to their adoption by academic researchers on the other hand, to defend university administrations which cannot be solely blamed for the “bibliometric fever”. He extends the remarks by asking why do universities thoughtlessly

http://dx.doi.org/10.1016/j.rhisph.2017.07.003

Available online 20 July 2017
2452-2198

adjust their priorities and direction based on admittedly flawed and erroneous rankings and indicators? He then describes ways of using indicators that match the metrics to be measured.

2.1. Recap on the historical origins

The term scientometrics was coined by Vassily V. Nakimov in a book *Naukometriya* published in 1969, which covered all aspects of scientific output analysis (at the time) across disciplines. The term bibliometric was used by Alan Pritchard also in 1969, limiting it to a subset of scientometric, the statistical analysis of publications in various contexts. The term technometric is used for the analysis of patents, and the newer term webometrics refers to the analysis of publications available online. But the origins of thinking about patterns in academic research productivity are much older. In a classic paper Lotka (1926) described the distribution of scientific productivity and used the term bibliometric to do so, and derived a power law (aX^{-2}) to describe the pattern. But before him, psychologists were already applying this type of analysis to the evolution of trends in publication, as were later other disciplines, albeit without using the term.

One of the earliest intellectual transformers of bibliometrics that contributed in numerous ways to modern thinking about analysis of research productivity was Eugene Garfield. He proposed a formal citation index in the mid-fifties, based on Sheperd's Citations, which listed all citations to legal decisions in the USA. He founded the Institute for Scientific Information (ISI) in 1959 which published the first Science Citation Index (SCI) in 1963. These occurred in the context of managing the ever-growing amount of scientific publications, both in increasing number of journals and of papers published, with no direct connection to research evaluation. In a paper published in 1951, Derek de Solla Price showed a doubling about every 15 years of the number of both journals and papers published between 1700 and 1950. He then showed this could be modelled as a power law similar to Lotka's power law of productivity of researchers in 1965. He called for a *science of science* and published the "Scientific Foundations of Science Policy" (de Solla Price, 1965). After Lotka's paper, and at least since the twenties and thirties, libraries had been analysing citations to manage their collections, to make space for new items based on usage and contemporary relevance. It was well-known already that in some disciplines papers were mostly cited for a couple of years, while in others they remained relevant for two or three decades. However, despite differences in the period the overall pattern was similar across disciplines, with papers being cited most after a lag due to publication time constraints, then would fade as they were cited less. Both the peak citations and absolute numbers over time were relevant to understand breadth of impact across disciplines.

In parallel to these developments, in the sixties the Organization for Economic Cooperation and Development (OECD), a creation of the US Marshall Plan after WWII and based in Paris, and the newly created National Science Foundation (NSF) in the US were pondering national science policies and how to evaluate the impact of scientific research. The OECD commissioned a series of reports on national science policies focused on economic inputs and patents. The NSF had received a mandate from the US Congress to address the evaluation of impact based on a set of indicators. It commissioned Francis Narin and his company Computer Horizons for a study. The lengthy report from Computer Horizons provided the basis for what became the foundations of "evaluative" bibliometrics. In 1972 the NSF published *Science Indicators* and it became the *Science and Engineering Indicators* in 1987. In the eighties and nineties SCI matured into its modern form and adapted to the new web-based reality. It is during this period that its index was applied to the output of individuals, as well as coarser evaluations of universities and countries for which it was designed.

Gingras calls this development and subsequent prolific misapplication of indicators "*wild bibliometrics*" in reference to Sigmund Freud's reference to "*wild psychoanalysis*". The data in these databases were intended for coarser analysis and not for the output of individuals. He describes the desire by administrations to down-skill and systematize research output evaluation. What used to be the domain of experts within a discipline was gradually handed over to improvising bureaucrats masquerading as experts in the evaluation of an individual's research output. One unfortunate counter-productive effect was that the same few indicators were imposed on all disciplines. The evaluation indicators must be adapted to the culture of the discipline, and not the other way round. And although it is simpler to de-skill the process and move it out of the hands of the experts, into the hands of bureaucrats looking at a few simple numbers, those indicators cannot replace human decision-making by those in the discipline. This is a critical and pernicious error. The disregard for the scientific validity of indicators and tools for the object to be measured or evaluated is unconscionable. Wrong indices or measures were (and are) applied and imposed uncritically because they are accessible.

Another development in 2005 was the publication of the H-index by Hirsch (2005) that attempted to provide an individual measure of research output quality. Bibliometric *evaluation* of individuals gradually became an *evaluation and ranking* of individuals, as well as academic departments, research centres, universities and countries. Gingras goes to some length reminding us and non-statisticians that evaluation is not the same as ranking. Using one to develop the other is prone with methodological errors that can render the ranking false, invalid and ultimately useless. These issues are tackled in chapters two and three.

There was one significant development at the beginning of the 21st century that contributed to the explosion of *wild bibliometrics*. This was the appearance of Scopus, Google Scholar and the transformation of SCI into the modern Web of Science. ISI published the SCI from 1963 to 2004 without competition, and in 1993 it was integrated into the Thomson Reuters Web of Science platform. In 2004, Elsevier one of the world's oldest publishers of scientific journals released Scopus, a modern more versatile platform to compete with the Web of Science. Google Scholar went online also in 2004, having accumulated a large enough database to provide a less structured and less defined platform to access publications and metrics about the publications. (Other popular platforms such as Mendeley (now assimilated with Elsevier) or ResearchGate are not discussed). These platforms continue to improve tools and provide metrics to analyse the research output of everything from individuals, to journals, disciplines, universities, and countries. Web analytics became accessible but alongside it misused by self-declared experts.

Gingras provides us with another short history lesson regarding the analysis of relations between disciplines and what became webometrics. In the sixties, Michael Kessler described bibliographic coupling by describing patterns based on papers that shared common citations. In the seventies, Henry Small described results from co-citation analysis, based on any two papers that cited each other. These analysis provided maps of connectivity and networks between disciplines, at the level of countries to individual researchers. These networks and their terminology are not unlike food web and ecosystem studies in ecology. Early results demonstrated research interconnectedness was more circular than linear. That is, all disciplines are more or less interconnected – as suggested by Piaget (1967), as opposed to all his predecessors from Francis Bacon to Auguste Comte who thought it should be linear (several examples are provided in chapter two). This type of analysis led to a number of now commonly recognised observations, as well as to several myth-busting observations.

2.2. Patterns in citations

The second chapter shows bibliometric has much broader scope than research evaluation, and that when used well it is a powerful tool. However its recent application to research evaluation is unwise (p IX). It can be summarized into a few examples of broad scale patterns, and of falsification of myths. First the main trends:

- Productivity has not grown significantly. There is an exponential increase in the number of researchers, published papers, and a tendency to cite more. There is a wide variation between disciplines that tend to cite more (Health and Biomedical Sciences) or much less (Humanities).
- Self-citations remain low at about 8% of total citations. These are necessary and unavoidable for a researchers to recap on previous results and contextualise the manuscript. Citations by a paper to others in the same journal constitute about 20% of total citations. This is again unavoidable since papers in the same journal are more likely to be relevant to that discipline. Similarly to previous observations by Lotka and de Solla Price, the distribution of citations over time follow a power law function (aX^{-y}) called the *Pareto distribution*.
- About 20% of researchers are responsible for about 80% of grants and citations, but about 20% are producing about 60% of the published papers. There is what is called the *Matthew effect*, whereby better known and already well-cited authors tend to be cited more frequently, reinforcing their prominence in citations. The prestige of the home university or academic department is another significant factor in selecting papers for citation.
- Total publications (and patents) from a country are strongly linked (correlated) to its GDP, and its willingness to invest in research and development.
- Citation practise and cultural behaviour within disciplines identifies four main clusters of citation behaviour. These are the Health and Biomedical sciences, Natural Sciences and Engineering, Humanities, and Economics.
- Data analysis identifies behavioural and cultural differences at the level of countries, disciplines, and sub-disciplines, or even type of university, so that cross-comparisons are fraught with interpretation errors unless corrected. The differences are in preference for books, book chapters, papers or conference proceedings; for single or multiple authors; for the assignment of the order of authors, who is first and who is last, or if it's an alphabetical list, and who is the corresponding author.

The chapter also tackles often recited falsehoods. For example,

- the myth that papers are cited for less than five years is false. In fact the median is increasing towards older papers, and in most disciplines it is much longer, although it can be short even 1–2 years in the biomedical sciences;
- the myth that most articles are never cited is false. This is in part due to the inappropriate narrow window used by some platforms and indicators. In fact using a 5–10 year window shows clearly it is not so. Very few papers are never cited.

Gingras makes an argument against the criticism that papers that turn out to be false, or even fraudulent are highly cited. One example used is that of the cold-fusion research, but he does not discuss the anti-vaccine topic. He argues correctly that results that turn out to be wrong are part of the natural progress of any subject and they need to be discussed. Fraudulent papers will be caught out and the authors eventually penalised within their discipline. Neither are the problem some make them out to be to cheat the system. I am not fully in agreement with this; yes the academic discussion is useful, but the bureaucratic implementation of indicators by those outside of the discipline would lead to false compensation, falsely-awarded recognition or even merit, and research grants. In fact, a few pages later, he provides a good example of why the issue is not so clear-cut when bureaucrats and politicians are involved. Gingras provides an example of how simple-minded bureaucratic approach leads to false conclusions. Rather than pick on recent fraudulent cases, he uses the famous example of Trofim Lysenko in the 1940s and fifties who was promoting a discredited theory of inheritance to advance agriculture, with his government's backing. This set-back Russian agriculture by decades. Yet he was at the time the most cited in his discipline because he was criticised for promoting an erroneous idea if not fraudulent research. By bureaucratic evaluation of indicators alone, he could have been awarded a Nobel Prize, and most institutions today would have cherished and encouraged his work just for the citations and impact against all reason.

In his review, Zitt (2015) brings a clarification by correcting Gingras on one point. That is regarding the perceived notion that the size of a discipline affects its impact (pages 33 and 66). Correctly stated, what matters and should be tracked is “the ratio between citing and citable literature” in the discipline (Zitt and Cointet, 2013).

Another topic broached but not well articulated is that of avant-garde forward thinkers in their discipline, or so-called *sleeping beauties*. Some papers are premature and remain undiscovered for years or decades until the subject is ready to assimilate and understand the results as a consequence of developments in that discipline. These papers and individuals working on them are precotious and insightful relative to their discipline. Yet any indicator or evaluation would fail to recognise this and would implement counter-productive restrictions on their work if it fails the test within the narrow window for calculating indicators or merit (Beauvallet, 2009). This implementation of indicators promotes a herd mentality of doing what is popular rather than innovation and discovery. This has certainly been the trend in government funding of research. This issue is addressed again with regards to the H-index and journal impact factors.

3. H-index, journal impact factors, university rankings

We are reminded that in various ways, at least for the past 350 years of experimentation, research by academics was evaluated by peers, their societies, and by universities. The author traces the peer review process to 1665 in the Philosophical Transactions of the Royal Society – London. We can remark in passing that the purpose of Society journals was not to reject manuscripts, but to set standards and to guide the improvement of manuscripts to that standard. This very important point is lost in journals that market themselves for profit, at the expense of the traditional Society journal that serve its membership and set standards for the discipline, instead of seeking only papers that might be highly cited for a publisher's profit. Gingras gives examples of low rejection rates as editors and reviewers tried to work with authors to improve manuscripts. The author steers clear of the contemporary debate on scientific publication, the role of universities, professional societies, and journal publishers. Indeed this is a distinct nest of vipers.

3.1. Grant proposal reviews

The peer review of grants is much more recent, as research tended to be funded by philanthropists, organizations, and industrialists, either provided directly to researchers or indirectly through universities. The first modern research university was probably the Free University of Berlin created in 1810. In 1901, France established grants for research. It was followed in 1916 by establishing the National Research Council in Canada. The United States NSF program was established in 1950. However, research into the awarding of grants indicates they are awarded almost at random as they are affected by the composition of the committee, among other problems (Cole et al., 1981). Gingras tends to avoid the topic and does not elaborate very far. For he could have used an example from his own Canada. In the mid 2000s the National Science and Engineering Research Council (NSERC) changed its mechanism for reviewing and awarding grants, against all recommendation. Bureaucrats were elated to show graphs as evidence of the success of the new method. It indicated that some previously well-funded researchers were receiving less or none, while others were suddenly receiving substantially more without a discernable change in their productivity or research direction. Perversely, one that had just received a top award from the agency for their research accomplishment was discontinued the next round of funding. Statistically, the system had been randomised. At a time when national governments are disinvesting from university research except in China and a few small countries, this is a contemporary issue. But even in China, as in almost every country, the *wild bibliometric* fever directs academia away from innovative research into applied research and popular research, at the expense of discovery. The subsidising of industry research is appealing in some respect but does not contribute to research, only to development and commercialization without maintaining the pipeline of diverse ideas that would generate discovery.

With far more researchers in the system than available research funding, and in this almost random process of awarding grants, many have given up and do not bother. When the success rate is too low and the amount provided too little many academics quite reasonably find better ways to spend their time – as with Sisyphus assigned forever to push his rock up a hill only to slip as he approaches the top. Many researchers continue to try to push their rock up the hill, but who gets there is typically random. After several failures, many are resigned to directing their energy, skill set, and intellect to other tasks, while some try to benefit from the *Mathew effect* above. The author targets the granting agencies by referring to their methods as baseless pseudo-quantifications that satisfies bureaucrats. He does not state it is at the expense of research careers and interrupted research programs that are lost. What a waste it is to spend as much money and years invested in forming a research scientists as it costs to form medical specialists or surgeons, only to have them idle without grants and instruments to do research.

3.2. H-index

There have been numerous papers specifically on this index and its demerits that Gingras summarises (and also Zitt in his 2015 review). These include criticisms that this measure of quality is not independent of the quantity or an author's productivity counted in the number of papers. Its value only increases with time, like a temperature thermometer that can only read up and never down. Another criticism is that the H-index can identify the very best (the leading front tail of the distribution) as the paper showed, but it is much poorer at ranking or evaluating the middle of the distribution, which is most people. The general conclusion from statistical studies is that the H-index "cannot be considered an appropriate indicator of a scientist's overall scientific impact" (van Leuven, 2008; Waltman and van Eck, 2011). Its numerical application has obvious perverse effects that any thinking person will identify in minutes. He suggests that had the paper been published in its bibliometric discipline it might have received peer review by knowledgeable experts. For Gingras this is another example of *wild bibliometrics* but not as Hirsch claimed "a more democratic assessment of people's research". As it stands, the paper ignored bibliometric past-knowledge, best practice, known conceptual pitfalls, do's and don'ts, and tried to reinvent the wheel. It ignores defining the conditions for its validity, its statistical distribution, and relevant significant figures. Gingras further remarks that it was already known that using the number of publications and the number of citations to obtain the median citation per paper, compared within the discipline, is the much better well-defined quality indicator.

3.3. Journal impact factors

As the Web of Science became more accessible and more popular in the nineties, there was a misconstrued notion that a paper in a high impact journal is a good paper. In fact it is inappropriate as a measure of the quality of its papers. In reality any journal will have a statistical distribution of papers that are highly cited, poorly cited, very good or with errors, or even fraudulent. Since citation rates follow the Lotka power law, most are not cited many times. He provides an example with the journal Nature in 2004, stating 89% of all citations came from 25% of the papers. So forcing authors to submit to high impact factor journals simply increases their rejection rate, and delays time to publication.

In addition, the one-size-fits-all narrow window for calculating a journal's impact factor is designed to favour Health Sciences and Biomedical Sciences where the citation cycle or period is very short, in contrast to all other disciplines where a paper is evaluated and cited after several years, and where the research-to-publication delays are much longer. As Gingras points out, if a ten year time frame is considered the difference between the impact of Biomedical Sciences and Social Sciences reverses. Therefore the window for the calculation needs adjusting to accommodate different disciplines, if they are to be compared, or a much longer time frame is required. Correcting this calculation error today is simpler and more trivial than it would have been in the seventies or eighties.

The chapter provides several examples of abuses of the impact factor, and of journals, and journal cartels that have tried to manipulate their value. These journals are excluded from the metrics published when their manipulation is discovered.

Gingras also discusses the more extreme misguided ways some agencies produced ranked lists of journals provided by bureaucrats to encourage publishing into high impact factor journals and discouraging publications in low impact factor journals. He gives the example of an important, difficult, very small discipline that might have one journal. Inevitably it will be a low impact factor journal, but it doesn't reflect at all the quality of the papers within it, or on the importance of the discipline. He also refers to universities that along the mindless use of the impact factor of journals, reward their researchers monetarily with substantial sums if they publish in higher impact factor journals. As he points out, research shows playing to numbers and indicators improves the performance against the indicators, but it does not change the overall quality of the research output. That is because a lower quality paper that sneaks into a high impact factor journal will still be a less cited paper. To my recollection, and as Gingras shows, many governments have succeeded in changing performance against indicators without improving performance or quality, whether it is in the provision of health, education, or research. This is another failure of the New Public Management ideology, of using indicators as objectives with the naïve and misconstrued belief that it will provide rationalization, efficiency, and excellence in using resources. It reminds me of the failed communist government indicators and objectives in their five year plans and production targets to be met. The objectives were often only met on paper, if at all.

3.4. University rankings

The three better known global rankings of universities are those from the Times Higher Education, the QS World University rankings, and since 2003, the Shanghai Ranking of the top one thousand universities, published annually. The chapter takes issue with the rankings based on methodologically dubious practises to reach the rankings, and through the book often refers to black-box calculations by various platforms and organizations. He shows that it is improper to rank universities between countries as they have very different structures and objectives. The rankings are also open to abuse, and he gives an example of one country that has its universities pay highly-cited authors from other countries to add a second affiliation address in their papers to this country, in order to boost its indicators. He uses examples to show how universities with no reputation appear one year at a very high standing, but then at a low standing the subsequent year, and how it is not possible for a university to be excellent one year but poor the next, or to improve by several hundred places from one year to the next when there has been no particular change. The pace of change and the things that make a reputable university are decades in the making and very rarely are they shattered or dramatically improved in one or two years. For measures that change very slowly, fluctuating annual rankings should be a caution that it is probably not a scientific endeavour. There is no reason why the exercise is not carried out every ten or twenty years, and more rigorously.

One case in point is the Research Assessment Exercise initiated in the UK under Margaret Thatcher, to measure the quality of every university department. The exercise occurs every few years only, and it was renamed the Research Excellence Framework in 2008. Ultimately it is a qualitative judgement based on site visits and reading texts provided by each department. This national scale exercise is similar to the university level program prioritization evaluations and ranking promoted in North America by private companies and consultants. In its last reiteration in the UK there were about 1500 evaluators in 67 research areas, at a cost of £250 million pounds sterling. It is not surprising many ask if there is any real value added to this exercise, and if the same amount could not have been better used to fund the universities instead. Given the misguided, scientifically invalid, and erroneous indicators used to evaluate these departments, the cynicism abounds. Furthermore, since it is evident that specialists within their discipline understand better the quality of their research than bureaucrats, I would argue the whole exercise could be carried out at a fraction of the cost by placing the department heads of each discipline in a room for half a day to come-up with a much more educated ranking. In fact, they could do it online.

Gingras argues that within a discipline there is always an unofficial or implicit understanding of who is better, and which department in what university has better scholars. Any organization or agency can come-up with an “official” ranking. Its worth really ought to be in its merit as scientifically sound, in how it is done, and what is measured. When these rankings are made public, then it becomes advertising. So I can ask, why would a government want to discredit some of its own universities, or disciplines, based on statistically invalid and scientifically unsound indicators and rankings?

There is a misperception that measuring individual output, departments, and universities with even erroneous metrics and indicators is better than not doing it at all. Gingras proposes there is a *Social Law* that using *any number is better than not having any*. However, in their encyclopedic review Todeschini and Baccini (2016) provide plenty to work with and improve upon, as do Wildgard et al. (2014). He thinks the simplified tools or indicators are preferred for bureaucrats to understand and use, to dumb down the evaluation process – a form of de-skilling the process to make it easier to carry out. Although it would be laughable if these evaluations and rankings were conducted by politicians, it should be unacceptable when they are conducted by universities who ought to know better. Sadly, university administrators use this demonstrably failed approach at the expense of those whose studies show the principles and tools used are inadequate and invalid. Recently, government reports in Canada in 2012 (Council of Canadian Academies, 2012) and in the UK in 2015 (HEFCE, 2015) have also reached the same conclusion and there are more citations to reports and studies in the book. Universities could instead rely on their own specialists to advise them on how to do it correctly. (I think Gingras suggests this is more about power and control rather than institutional stupidity).

3.5. Internationalization

There is an implemented perception, with funding and merit implications, that research of local significance is less important and grandiose than research of international scale – that local is less, global is more. Unfortunately there are local issues that need to be researched and they are not less important, or not important; this research needs to be done if global scale or international research is to be given any interpretation. The reason for this perception is that local research is cited mostly at the local level, while research of an international scope gets more citations. Gingras provides an example, of an economist that might publish on the economy of Canada or the USA. Since much more people study the American economy, those papers and the journals they are published in would be much more cited, with higher journal impact factors. To obey agency, government, and university administrators towards higher indicator values, it is clear what should not be studied and what should be. Is this really a clever way to distribute funding to universities, researchers and departments? In other words, local journals and local languages matter and have value that needs to be accounted for in any indicator. This had huge research funding implications in the choice of subjects that are accosted by new professors in academia.

Long before any topic within a discipline is popular, it has a period of maturing in the hands of very few that establish the ground work. The methods and concepts need to be developed; they have to be explained to others in the discipline, they have to find connections with other disciplines. That means there will be fewer citations and so the research will not be accepted by high impact factor journals. This then becomes an excuse not to fund the research because of these two indicators. How is any novel research to occur, to drive ideas and discovery that eventually reaches the development and commercialization pipeline?

Gingras reminds us of the *Zhdanov Doctrine* at the beginning of the cold war in 1946/47. Andrei Zhdanov proposed his famous doctrine in response to the *Truman Doctrine*, and described the world in two camps, that led by the *imperialists* (USA and its allies) and that led by the anti-imperialist & anti-fascist *new-democracies* (Russia and its allies). The strict application of the doctrine meant to force everything to conform to the government's perspective and priorities. He draws a parallel between that with the application and consequences of *wild bibliometrics,* and the enforcement of scientific and cultural activities to conform to misconstrued government priorities based on invalid indicators. The third chapter ends with a reminder that rankings of the official kind are neither inevitable nor productive. Australian researchers succeeded in beating back their government's attempt at imposing a British style *Zhdanov Doctrine* on what journals and disciplines mattered, to some extent.

4. Correcting the evaluation of research

The proliferation of new indicators was never accompanied by rigorous evaluation of their validity, or determining if any indicator actually measures the intended metric well. In other words, they are never explicitly submitted to testing their validity before they are used. Three criteria are proposed for evaluating institutions:

- It should be evaluated based on its mission and mandated objectives;
- It should be evaluated based on stated goals to accomplish, and these should be stated fairly precisely to develop appropriate metrics;
- Then, indicators should be defined to measure if those goals were achieved in the time frame covered.

Typically a set of indicators are applied and universities are asked to adjust to the indicators, rather than the correct way around. Not all universities can have the same mandate, purpose, and objectives. We are cautioned against using too few indicators to measure the multifaceted role of a university (economic, societal, cultural, environmental, education, research, contribution to progress, etc.). It would tend to create the *lamp-post syndrome* of looking only under the lamp-post for lost keys, which is only looking at certain aspects of a complex institution. In national and global rankings, reducing a variety of indicators, even if they are statistically valid and rigorous, into a single number or rank is easily discredited. The multidimensional space of multiple indicators, reduced to a single value without dimensions or one dimension, such as a rank, loses most of the information in the multiple axes. The aggregation can only go so far before one is mixing apples and oranges, and metrics that cannot be aggregated.

Another issue in developing metrics is to determine which numbers can be obtained. For example post-graduation data are scarce. So evaluations can only rely on metrics that can be reliably obtained and tracked. I have personal experience in ranking departments in my university, and the extent to which numbers become inaccurate or vague as they are aggregated, and how important it is to understand how a university tracks numbers in its database. Most numbers that can be accurately determined by a department are fuzzified or falsified by the institution in its attempt to disaggregate, and in back-calculations of the numbers. This is not a trivial university-level issue because many numbers are now obtained from metrics provided by online platforms such as Google Scholar, Scopus, Web of Science, and others. None, for example, will calculate the same H-index for the same person, largely because of differences in their data base and how their numbers are obtained. Even though it takes minutes to calculate your own correctly, the web platform numbers are those accepted by institutions. It raises the relevant question, I might ask, what are the error bars and significant figures associated with each metric, if they are to be used to rank people?

The various platforms have different databases and operationally search and find different objects. There are private databases that can be contracted but they do not necessarily disclose what their database is or how their calculations are done. According to Gingras, Web of Science has about 12,000 journals (18,000 including conference abstracts and other items), Scopus about 22,000 journals (about 29,000 including all other items). Therefore not all journals are in these databases and there is a strong bias against certain languages or local publications. There is a discussion included that although books are not in common databases, citations to them are, so the difference this would make to any metric is very small. However to represent a discipline proportionally and fairly there must be an even representation of books across languages, especially in the Social Sciences and Humanities.

The "Altmetric Manifesto" of 2010 is criticised as containing clichés and criticisms but being short on providing sound statistically valid alternatives, while referring to poorly defined "others". The immediacy factor is curious because on the one hand, it recognises that it takes years to fully evaluate the impact of research, while on the other hand promoting the instantaneity of twitter, blogs and other social media. As Gingras points out, the half-life of tweets is measured in minutes, that of blogs in days, and that of citations in years. So only the latter scale is the correct one.

It is also easy to manipulate un-refereed or un-scrutinised instant online information. We are given the example of a webometric experiment by Cyril Labbé at the Joseph-Fourier University in Grenoble who posted 100 short articles online on a web site, each cross-referencing the others, using a fictional author name. The fictional author received instantly an H-index of 94 in Google Scholar, which is most difficult to accomplish by a real person.

Gingras refers us to the sixteen Berlin Principles on Rankings of Higher Education Institutions. Noting that "most measures used are not fit-for-purpose since they do not compare dynamic behaviour of the indicator with the behaviour of the concept to be measured" (p 71), he suggests four criteria that must be fulfilled for an indicator to be valid for the object to be measured. 1- *Adequacy* for the object; 2- *Sensitivity* to the intrinsic inertia; 3- *Homogeneity* of the dimensions of the indicator; and 4- The indicator must be a *monotonically increasing* function of the concept it measures. As we have seen so far, it is indeed false to claim that only the quality of the data matters (its quality is usually poor and can be unverifiable), and that indicators can be selected based on convenience. Both need a great deal of care is selecting and matching. Each criteria is explained with examples.

One example for *adequacy* is that of enumerating Nobel prizes that can be associated with an institution. Often the work recompensed was conducted decades earlier and is no longer a valid indicator of the institution today; or the research was completed at another more suited institution. An example of appropriate *sensitivity* is given using a thermometer that reads 20 °C, 12 °C a few minutes later, and 30 °C a few minutes later, when in fact the room is held at constant temperature. One quickly recognises these as instrument errors. Similarly, when an indicator fluctuates more than it is explainable for one or many institutions when there was no known change, one must question the tool or the indicator. An example used to demonstrate *heterogeneity* is combining different kinds of indicators together into one measure, such as combining "academic reputation" and "number of international students". When the composite indicator changes, there are too many variables and unknowns in each to properly evaluate what it means. Instead, each metric should be kept separate and presented on a spider-web graph. Lastly, the *monotonic* behaviour of the function is demonstrated using the example of the number of foreign professors and students at an institution. If this was a good indicator, then having 90% foreign professors or students is better than having 20% of each. However, anyone would quickly remark that would actually be problematic, and represent a colonial institution, or one that could not find qualified local students, or find suitable local professors. Unless there is a well-established reason and well-defined optimum value these indicators are not useful.

How do these criteria fit for the H-index and global university rankings? One obvious criticism is reliance on heterogeneous indicators in both the H-index and the Shanghai ranking of universities. The various indicators used by the latter do not meet most of the criteria outlined, even before they are aggregated to a single number. However, Gingras notes that more recent ones such as the CWTS Leiden ranking and U Multirank are much better because well-defined indicators are used, and they do not combine heterogeneous indicators with arbitrary weightings.

Lastly, one cannot leave the topic without a reflection on university branding and marketing administrators who have shamelessly used these indicators. It befalls the university community, and not just its administrators, to contemplate the damage done, to rectify it, and to push back on obviously erroneous indicators and tools, when better ones can be used. After all, it is for everyone in the system to determine what kind of a democracy we want, what kind of administration we accept, and when to call-out on the falsehoods and the mismeasure of universities and individuals. Notwithstanding, as Gingras also remarks, that it regularly becomes an opportunistic excuse to further defunding the system.

Acknowledgement

I thank my colleague Larry Stewart (History, University of Saskatchewan) for his valuable comments and corrections to the document.

References

Beauvallet, M., 2009. Les stratégies absurdes. Comment faire pire en croyant faire mieux. Seuil, Paris.
Cole, S., Cole, J.R., Simon, G.A., 1981. Chance and consensus in peer review. Science 214 (4523), 881–886.
Council of Canadian Academies, 2012. Informing Research Choices, XIII.
De Solla price, D., 1965. The scientific foundations of science policy. Nature 206 (4981), 233–238.
Gingras, Y., 2016. Bibliometrics and Research Evaluation – uses and abuses. MIT Press, Cambridge Massachusetts. (Translated from: Les Dérives de l'Évaluation de la Recherche: du bon usage de la bibliométrie (The Excesses of Research Evaluation: The Proper Use of Bibliometrics). Yves Gingras. Paris: Raisons d'Agir Editions, 2014. 122 pp).
HEFCE (Higher Education Funding Council for England), 2015. The Metric Tide, VIII-X.
Hirsch, J.E. An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences 102 (46): 16569-16572.
Lotka, A.J., 1926. The frequency distribution of scientific productivity. J. Wash. Acad. Sci. 16 (12), 317–324.
Piaget, J., 1967. Le système et la classification des sciences. In: Piaget, J. (Ed.), Logique et Connaissance Scientifique. Gallimard, Paris, pp. 1151–1224.
Todeschini, R., Baccini, A., 2016. Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research. Wiley, Weinheim, Germany.
Van Leuven, T.N., 2008. Testing the validity of the Hirsch-index for research assessment purposes. Res. Eval. 17 (2), 157–160.
Waltman, L., van Eck, N.J. 2011. The inconsistency of the H-index. ArXiv:1108.3901v1.
Wildgaard, L., Schneider, J.W., Birger Larsen, B., 2014. A review of the characteristics of 108 author-level bibliometric indicators. Scientometrics 101, 125–158. http://dx.doi.org/10.1007/s11192-014-1423-3.
Zitt, M., 2015. Book review (in English) of: les Dérives de l'Évaluation de la Recherche: du bon usage de la bibliométrie. J. Assoc. Inform. Sci. Technol. 66 (10), 2171–2176. http://dx.doi.org/10.1002/asi.23519.
Zitt, M. and Cointet, J., 2013. Citation impacts revisited: How novel impact measures reflect interdisciplinarity and structural change at the local and global level. In: Hinze S., Lottman A., (Eds.), Proceedings of the 18th International Conference on Science and Technology Indicator (STI), Berlin, Germany, September 4–6, 2013, pp. 466–476.

Sina M. Adl
College of Agriculture and Bioresources, University of Saskatchewan, Saskatoon, SK, Canada
E-mail address: sina.adl@usask.ca