Vergara Fernández, Melissa

Working Paper

Towards measuring journal impact - Properly

CHOPE Working Paper, No. 2020-13

Provided in Cooperation with:
Center for the History of Political Economy at Duke University

Suggested Citation: Vergara Fernández, Melissa (2020) : Towards measuring journal impact - Properly, CHOPE Working Paper, No. 2020-13, Duke University, Center for the History of Political Economy (CHOPE), Durham, NC

This Version is available at:
https://hdl.handle.net/10419/228866

Towards Measuring Journal Impact– Properly

Melissa Vergara Fernández

CHOPE Working Paper No. 2020-13
December 2020

Towards Measuring Journal Impact—Properly.

In this paper I evaluate the Journal Impact Factor using a theory of measurement. I argue that JIF does not stand up to close scrutiny. To measure a concept adequately, our theory of measurement requires correspondence between three steps: the characterisation of the concept, its representation, and the procedures followed to carry out the measurement. Characterisation involves defining the concept: identifying its boundaries, which fixes the features that belong to it. Representation involves defining a metrical system that appropriately represents the concept. The procedures are the rules formulated for applying the metrical system to the tokens. These three steps do not line up together neatly for JIF. There are at least two problems. First, the procedures to measure JIF do not reflect an unequivocal characterisation. Second, the representation strategy of JIF is inappropriate and not justified, given the kind of concept it tries to capture: one without strict boundaries. The bottom line is not that the JIF ought to be eschewed. Sufficient reasons related to how JIF distorts scientists' incentives have been provided to this end. But path-dependence is a tricky issue—the longevity of qwerty keyboards demonstrates it. The bottom line is that, given that JIF is unlikely to vanish, we better start giving it some proper scientific basis.

> when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind…
>
> —Lord Kelvin

> If you cannot measure, measure anyhow.
>
> —Frank Knight

I. Introduction

It is difficult to quarrel with the motivation behind measuring things, phenomena. Who would want to know less, rather than more? Quarrels with measurement arise with respect to the ways in which phenomena are measured, how the information collected is used, or both. The metric of "journal impact" is no different. In the modern world, in which our heartbeat is measured with precision by a wristwatch or Google Trends has the potential to predict unemployment spells (Choi & Varian, 2012), there's no reason not to also want to measure the impact of our scholarship. Intuitively, it makes sense to have knowledge of how ideas within and across scientific communities travel. As discussed below, this was the idea that gave rise in the first place to Journal Impact Factor (JIF), the most widely use metric of journal impact. Besides, tax payers' monies

around the world fund a great deal of our scholarship. Arguably, academia should be accountable for how it uses those monies and a way to do that is by measuring its impact. Or trying to. And that's part of the quarrel. There is criticism across disciplines that JIF is a poor measure, for instance because average citations grossly misrepresent the actual citation record of sometimes the majority of the articles published by a journal (e.g. Editorial (2005); Leydesdorff et al. (2016)). The other part of the quarrel is about how the information collected is used. On this front, criticism has been levied against JIF in that it generates perverse incentives for the academic community (e.g. Brembs, Button, & Munafò (2013); Moustafa (2015); Perez, Bar-Ilan, Cohen, & Schreiber (2019)).

In this paper I shall suggest that the way in which "journal impact" is currently measured has no scientific basis. Simply put, if pitched against a theory of measurement, this theory tells us that JIF doesn't measure adequately what it says. But my purpose is not to merely provide additional grist to the critics' mill. Rather, my purpose is two-fold. First, I want bring attention to the following. We collectively subject our research—that is, our experiments, models, evidence, etc.—to certain standards of scientificity. Yet, since we uncritically continue to use JIF, we aren't following suit with the measurement of our scholarship. "How come the double standards?" is the question the paper intends to animate. Second, and more importantly, I want to suggest that understanding the drawbacks of JIF with respect to a theory of measurement may help us think of ways in which this metric can be modified and improved.

The paper is organised as follows. First, I will briefly describe what JIF is and how it is used (section II). Then I will introduce the theory of measurement (section III) against which I will pitch "journal impact" (section IV). I will then suggest how this exercise offers lessons that might help us to modify and improve JIF (section V). A short conclusion follows.

II. JIF: What it is and how it's used

The JIF is a component of the Journal Citation Reports (JCR). These are published by Clarivate Analytics every summer. The reports offer detailed quantitative information and analysis of every journal Clarivate has indexed[1] into its Web of Science selection of journals. JCR are a tool that, given the interests of the user, can be used for the purposes mentioned above. Clarivate warns that the JIF, which is the most important of the indicators and the basis of much of the analysis, should not be used on its own. Clarivate thus suggests that the reports as a whole are a tool that aids judgement.

JIF is a measure of the frequency with which the average article of a journal is cited by other journals in a specific year. Generally, it is calculated by dividing the number of citations to journal x in year t to items published in $t - 1$ and $t - 2$, into the total number of articles published by x in $t - 1$ and $t - 2$[2]. The idea is that the more citations a journal gets on average, the more impact it has. But the number of citations is normalised with respect to the number of citable items a journal produces. Otherwise, journals that publish little or are new—and thus have fewer citable items—couldn't possibly be compared with those that publish a lot or are old.

This metric is used in at least four ways (Clarivate, 2019). First, it helps librarians to select or remove journals from their collections. Second, it aids publishers and editors to determine a journal's impact in the marketplace and to, in turn, set its publishing strategy. Third, it aids researchers to identify the journals that best fit their interests and to decide where to have their work published. Finally, it aids research managers to track bibliometric and citation patterns that may be important for grants decisions, among others.

[1] Arguably, how this selection is made is also a contentious issue. I will not deal with it in this paper.
[2] I say "generally" because the two-year window is the most common and the one known as JIF. The 5-year JIF is also calculated and published, but it is, naturally, less current. It is known as the 5-year impact factor. Whether certain fields like history should use less current JIFs is a point some commentators make.

III. A theory of measurement

Measurement is a privileged source of knowledge. Although philosophers have found it difficult to give an unequivocal definition of measurement, many agree that it is an activity that involves interaction with a concrete system with the aim of representing some of its features in abstract terms, such as in classes, numbers, or vectors (Tal, 2015). In this section I'll discuss a theory of measurement introduced by Cartwright et al. (2017) and Cartwright & Runhardt (2014) that is helpful to understand the pitfalls of JIF. It is concerned mainly with measurement in the social sciences. According to this theory, three requirements have to be fulfilled in order to measure adequately. That is, that the abstract terms correspond systematically with the concrete system of interest. First, the concept of interest has to be characterised. This means that the criteria that determine the boundaries of the concept are set. Second, the way in which the concept is represented has to be defined. So given the characterisation, a corresponding representation has to be chosen. Finally, a set of procedures has to be established to make sure that the tokens that are picked out given the concept, are really the ones intended to be pick out. Let me discuss each of the requirements in turn. First, characterisation. Some concepts pick out qualitative or quantitative properties that individuals or populations have like sex or age. This is easy. Others, however, sort things into categories that are based on criteria that are blurry and thus more difficult to characterise. Unlike the age of trees or even the weight of the W boson, some concepts related to social phenomena exist only because they are of interest to us. If we didn't care about unemployment or civil war or journal impact, such phenomena would be nowhere to be found[3].

[3] I use the controversial case of the W boson to illustrate that the distinction is not merely a matter of observability. The weight of the W boson, if only indirectly observable, is measured under the assumption that it exists. It would continue to exist even if we hadn't discovered it or measured it. We wouldn't be able to make the same claim about "civil war" if we didn't care about it as a phenomenon. To be sure, citizens of a country could still be in conflict and kill each other. But if we

It is not nature that neatly distinguishes them, but it is our interest in them that does. And this often isn't neat. There isn't anything in nature that tells us when civil war is taking place. Or when an alcoholic has become one. Or when climate change set in.

The first requirement of our theory is therefore that we set the boundaries of our concept. That we define the criteria that will allow us to decide whether a token—e.g. a country or an individual—fall under our criteria. Take the example discussed by Cartwright & Runhardt (2014) "civil war". Four aspects have been commonly taken to define it: presence of internal fighting, active government involvement, appreciable amount of force applied by the involved parties, and a certain number of deaths as a result of the conflict. If we were to say that, for instance, in addition to these four, involved parties are mostly women, our concept of civil war would be rather different. In turn, we would assign other tokens—probably none—to our newly created concept of "civil war".

The second requirement is that the abstract characteristics of the formal representation of the concept reflect and are warranted by the characterisation of the concept we intend to represent. Take "civil war" again. Provided that we have defined it as presence or absence of the four aforementioned features, we can only represent it as a binary variable. We can only tell *whether* Colombia or the Netherlands are in civil war. To track the severity of Colombia's civil war before and after the peace accord between the government and the largest guerrilla group FARC, a different representation—and characterisation—would be needed.

Several kinds of representations are used in science. The most common are the nominal, which assign different numbers (or letters) to the tokens that fall under the concept; ordinal, which rank tokens that fall under the concept; interval, which orders tokens on a scale with equal intervals; and ratio, which order tokens on a scale with equal ratios and a true-zero point.

---

didn't care about this fact, the concept (and phenomenon) of "civil war" wouldn't exist. See Searle (1995) for a discussion of how social, or 'institutional' facts, as he calls them, come into existence.

For those concepts that have blurry boundaries, there are three common strategies for representation (Cartwright & Runhardt, 2014). The best is as tables of indicators. None of the features by which we want to define a concept can be singled out as essential. We must therefore consider them all, if we are to characterise our concept as comprehensibly as possible. Civil war would be much coarser—let alone uninformative—if we were to characterise it—and thus represent it—solely as "presence of internal fighting", as per above. The Netherlands would have a civil war too, by dint of the Dutch gangs in the southern province of Limburg, who compete for the illicit drugs market and often shoot at each other. The downside of tables of indicators is that it's difficult to compare across time or across the tokens that fall under the concept. An alternative is to strip much of the concept and pinpoint a precisely definable feature. Something like this has been done with the concept of "race", in different medical contexts (Efstathiou, 2012). Epidemiologists care about race in terms of regional heritage that may be associated with risks of diseases. Geneticists care about it in terms of genetic polymorphisms. The last alternative is representation as an index; a compromise between the other two. Examples are the Human Development Index or the Consumer Price Index. These keep the variety of features but weigh them according to some rule to obtain a number between zero and one. Therefore, it allows comparison across tokens and across time.

The final requirement is that the procedures followed to carry out the measurement correspond to the way the concept has been characterised and represented. The procedures are the methods used to find out which tokens belong in the categories. To stick to the "civil war" example, the procedures here would refer to say, how the number of casualties will be counted—will only people in military uniforms count?

Cartwright & Runhardt (2014) suggest that coming to correct procedures often involves reconsidering characterisation and representation. For instance, we might have characterised a concept in a particular way, but then we discover that, in practice, the procedures necessary to do

justice to this concept are cumbersome or too expensive to carry out. Another possibility is that, since social sciences often rely on the statistical data compiled by statistical offices, it is not always possible to determine the procedures on the basis of how the concept has been characterised. It is thus crucial to make sure that the three parts of the measuring process line up neatly. In that way we make sure that there is correspondence between the empirical system and the abstract terms by which we represent it.

IV. An assessment of JIF

Now that I have introduced the theory of measurement, let me pitch "journal impact" against it. I will argue that there are two problems; one related to its characterisation and the other to its representation.

Characterisation

The first requirement of our theory of measurement is that our concept of interest be properly characterised. This involves finding out what the boundaries of "journal impact" are. There are two related problems here. First, there is not an unequivocal characterisation of the concept. The boundaries aren't properly defined. "Journal impact" is understood and therefore used in different ways. I can think of three in which we could try to determine how these boundaries can be defined. First, to look at whether, and if so how, Clarivate has done this. This is JIF, as we know it. Second, to attempt to reconstruct its characterisation by looking at how JIF is used. Third, to dig the history of the concept and find out what its original purpose was. These are all different possibilities for characterising the concept. The second problem is that none of these three ways is satisfactory. I shall discuss each of them in turn.

Officially, Clarivate defines JIF according to the operation required to calculate it introduced above This is at least what can be inferred from website, including their training guides and documents like "Journal Citation Reports: A Primer on the JCR and Journal Impact Factor" (King, 2017) This

way of defining a concept is known as operationalism. Percy Bridgman, its most famous proponent, stated that "a concept is synonymous with the corresponding set of operations" (Bridgman (1927) quoted in Tal (2015)) arguing that nothing more was needed for definition. Such a strategy is the most extreme version of a pragmatic perspective towards measurement: there are no facts of the matter about which operations *truly* measure a specific quantity (Tal, 2015). An implication of this perspective is that there can't be more than one operation for one concept. Length measured by using a ruler or by timing electromagnetic pulses are, strictly speaking, two different concepts.

Cartwright et al. (2017) suggest that, rather than always being a pragmatic perspective, operational definitions are often used when understanding of the concept and knowledge of alternative features that might capture the concept are deficient. In addition, and more importantly, operationalisation makes knowledge accumulation difficult. This means that there's no basis to establish conceptual and (potential) causal relations with our concept of interest. An example illustrates this point. Take "civil war" again. We are able to say that 'youth bulge' is a possible cause of civil war (Heinsohn, 2003) because we can associate unemployed and dissatisfied young men with civil unrest. . However, if we were to define civil war operationally as say, "the number of military deaths", such a link would be more difficult to make. In particular, because there is no other set of procedures that warrants that we're measuring what we intend; any other procedure simply points to a different concept.

In the case of JIF, because it's difficult to justify that different procedures measure the same quantity—journal impact—we have little guidance with respect to what might determine it. Is it quality, popularity, signalling? If we can't tell what determines journal impact, the metric loses its normative force: librarians, researchers, publishers have no other reason to care about it except for its own sake. At least in principle though, we're supposed to care about journal impact because it signals something we value.

Let me now turn to the uses of JIF. Can we reconstruct a more comprehensive characterisation based on its uses? Recall that above I mentioned four ways in which the JIF is used. Arguably, the idea underlying these different uses is that high impact somehow tracks prestige and, in turn, prestige tracks quality. If librarians want to subscribe to high-impact factor journals, publishers want to publish high-impact journals, researchers want to have their work published in high-impact journals, and research managers want to fund research that has prospects of being published in high-impact journals, high-impact journals must be of high quality[4]. Granted, perhaps publishers are not necessarily interested in high quality academic work insofar as their journals make money. And perhaps something similar is true, though to a lesser extent, for the researcher—it's impossible to deny the perverse incentives of the publish-or-perish academic landscape. Still, if collectively we endorse these uses, it must be because collectively we have come to understand JIF as a measure of quality. At least tacitly it has been characterised as such[5].

Alas, the problem here is that there are good reasons to challenge that tacit characterisation. Let me discuss two. First, JIF gives prevalence to the short-term publication record; what Leydesdorff, Bornmann, Comins, & Milojević (2016) call the research front. They argue that a distinction should be made between this short-term research front and the long-term processes. The research front tends to involve transitory knowledge claims. These are claims whereby the researchers inform one another about progress. They reflect involvement in current discourses. By contrast, in the long-term, knowledge claims become codified into large bodies of knowledge. The problem here is not that quality is only to be found in the long-term, established bodies of knowledge. Rather, that we have no a priori reason to presume that research quality is *only* associated with short-term

[4] Naturally, "quality" is not without its problems. Do we mean by "quality" rigorousness in research, break-through ideas, accessible writing, all of them? More could be added.
[5] although in writing Clarivate makes explicit that the JIF is just one metric and that it should be used with discretion, in one of their informational videos about JIF, the voice over of the video says "Journal Impact Factor scores help you compare journals to assess the relative quality of different publications"(Web of Science Training, 2017).

high average citation frequency, as journal impact does If anything, long-term processes are, prima facie, better indicators of quality. They have passed the test of time.

Now the second challenge. Worse than that there is no association between quality and impact factor, Brembs et al. (2013) provide evidence that impact factor can be negatively correlated with quality. They discuss two features associated with quality: reliability and methodological soundness. In terms of reliability, there are two well-known phenomena in publication patterns that lead to the negative correlation in some fields. One is publication bias, the phenomenon that you're more likely to get novel and surprising results published than if you try to replicate a known one. The other is the decline effect, which is that published effect sizes tend to decline with time. So, the first time a causal relation is established, the effects published are large and they tend to decline in later attempts to replicate the effect. These two phenomena, together with the fact that initial publications occur in high-impact journals, suggest that the effects published in high-impact journals are overestimated. If this is so, they are likely to be less reliable. With respect to methodological soundness, (Brembs et al. (2013) cite several studies, and one of their own, of journals in different fields in the medical sciences that fail to find statistically significant correlations between high impact factors and levels of evidence or adherence to statistical guidelines.

The implication of these challenges for what concerns us here is that JIF can't be characterised as being the quality of scientific research. As such, JIF doesn't seem to always assign high JIF values to high-quality journals and low JIF values to low-quality journals. It doesn't do this systematically. It is as if my wine thermometer would be able to tell me that my *Pouilly-Fuissé* has the right drinking temperature only sometimes and I couldn't tell when. It would be a useless instrument.

Let us now turn to the history of JIF. There are two episodes that are important to highlight here. Both reveal that JIF was first used to solve a practical problem. In 1927, Gross & Gross (1927) proposed the idea to rank journals according to the number of citations to them with the purpose

of guiding librarians in small colleges to select scientific journals. They focussed on chemistry journals. Gross & Gross (1927) were addressing a challenge that, at the time, was arising for small colleges in the United States. Colleges were no longer seen as sufficient to prepare students for working life but rather, they had to prepare them for specialised graduate programmes at the university. Simultaneously, they had to impart cultural education. In addition, they hired faculty with PhDs, and libraries had to offer them access to periodicals that contributed to their research. The problem Gross & Gross (1927) were trying to solve was thus: What journals should a college library have that could both prepare students for advanced graduate work and serve the research needs of the faculty, while accommodating to their financial restrictions?

This first conception of a journal impact factor very much resembles the use that I mentioned above for librarians. Gross & Gross (1927) thought of it as "a standard of some kind by which to measure the desirability of purchasing a particular journal" (p. 386). So, in a way, it could be argued that, when it comes to librarians, their use of JIF has to do more with 'librarian desirability' than with 'quality', as I suggested. The problem with this interpretation is at least two-fold. First, the way Gross & Gross thought about it, is that each individual field and library, the "local needs", should determine what journals to purchase. This means that every party would have to do their own characterisation—determine the criteria by which they will find a journal purchase desirable. Currently, though the Journal Citation Reports offer a number of indicators and the possibility to analyse the data according to different criteria, the JIF is a one size fits all approach. Second, considering that publishers nowadays offer journal subscriptions bundled, it is legitimate to ask whether librarians are really able to assemble their libraries as they see fit[6].

The second episode involves Eugene Garfield, the creator of the Science Citation Index (SCI). He used a similar principle as Gross & Gross (1927) to select some of the source journals to be

[6] In 1998, Elsevier, one of the biggest publishers, introduced "The Big Deal", a plan for the internet whereby university libraries would pay a flat fee for access to bundles of journals. This system is still in place. See Buranyi (2017).

included in the first SCI in 1961 (Garfield, 2006). The SCI was conceived by Garfield as a means to lend credibility to scientific citations (Garfield, 1955). His main motivation was his concern with scientific work being cited to support a particular claim, but without the reader being able to tell what the previous history of the cited work was. The cited authority could have been previously criticised or shown to be incorrect, and a reader would not be aware of this. He got a tip to use the 'citator' system, used in law already since 1873, and published as the *Shepard's Citations* (Garfield, 1963).

Since precedent is crucial in the law, Shepard's Citations Inc. has published a listing of court cases in the United States, giving a record of the publications that have referred to a case, of other court decisions that have affected them, and any other information with respect to a case that may be relevant for a lawyer. A lawyer looking to find authoritative cases useful for their case can consult the *Shepard's Citations* for all the subsequent cases that have cited a case of their interest. The lawyer is then able to tell whether the cases they are interested in are still good as authority—e.g. not overruled, or reaffirmed (Adair, 1955).

Garfield intended to provide the same possibility to readers of scientific research by using a similar system. A crucial difference between a citation index for law and one for science was—and continues to be—the volume: when Garfield first wrote about the SCI in 1955, the order of magnitude between publications in science and the law was, according to Garfield, from fifty to a hundred times greater in science. A selection of the journals to be included in the science index had to be made. The Journal Impact Factor was created for this. Rather than a measurement tool, it was a tool that brought some kind of objectivity to the selection process[7]. The real gains, the proper contribution, were supposed to be offered by the SCI. I'll come back to this point below.

[7] See Porter (1995) for a discussion of how quantification is a "technology of distance"—from judgement, subjectivity and bias.

What can we conclude about these episodes as a way to characterise JIF? JIF was conceived as a selection tool. Both Gross & Gross and Garfield devised a ranking that was helpful to select journals either for library collections or the SCI. They were not meant to reveal any intrinsic features of the journals they were ranking. Besides the number of citations to them, there wasn't anything else to say about the journals selected.

In sum, when it comes to characterisation, neither of the three possibilities I considered for defining the boundaries of "journal impact" are helpful. Clarivate's definition is operational. These boundaries are too restrictive and hamper knowledge accumulation. The current uses can be said to define the boundaries around quality. But if so, there are other metrics that suggest it is a poor indicator of it. And it can't show otherwise because of its operational characterisation—it doesn't have the internal machinery to prove that it is the other metrics which are faulty. As for the history, there was no attempt at defining boundaries. JIF was meant as a tool to select, not to measure anything.

Representation

When I presented the requirement of representation above, I said that concepts that have blurry boundaries are represented in either of three ways, the table of indicators being the best. The reason for this, again, is that some concepts are best characterised by a set of features, none of which is essential. Like with members of a family, they may share several traits, but not all have skin problems or crooked noses.

I'll submit that "journal impact" is one of those blurry concepts. There are two reasons for this. First, there are many features I can think of that could fall under the concept. It could be impact in terms of the reach outside academia, or impact in terms of break-through ideas, or impact in terms of reaching different fields, or impact in terms of generating the most benefit for society, or impact in terms of having the most influence on politicians and public policy, or impact in terms

of being the most read, or impact in terms of being the most read in non-English speaking countries. Surely there are many more. These are all features that could well characterise as journal impact while none would we regard as essential. In science we probably value them all.

Second, as Cartwright et al. (2017) suggest, the distinction between blurry and pinpoint—precise—concepts is not sharp. In fact, some concepts such as "temperature" have evolved from being fuzzy to pinpoint[8]. So "journal impact" doesn't have to stay fuzzy. At the moment though, considering that it was rather a selection tool and not intended as *the* metric of our scholarship, it is fuzzy.

Hence the problem with its representation. As a fuzzy concept, it would best be represented as a table of indicators, including the features we, in academia, care about that are related to the impact of our scholarship, such as the ones I mentioned above. Should that strategy fail, perhaps because we do care about being able to make comparisons, we might opt for selecting a single naked feature. Or an index, if we want to compromise. But JIF is a convenience trick that stuck. This means that its current representation, the single naked feature of the ratio between citations and recent citable items published, is not the outcome of a conscious choice. More precisely, it isn't the outcome of a neat line up of the three requirements. It has no justification.

To be sure, I am not trying to suggest that we can only start measuring once we have perfectly discerned how it is to best characterise and represent our concepts and established the procedures accordingly. This is clearly not consistent with the history of science. It took nearly two centuries of experiments to settle on the freezing and boiling points of water as the fixed points in thermometry—in 1701 Isaac Newton proposed the melting point of snow and blood heat as candidates (Chang, 2007, Chapter 1). Usually going back and forth between the requirements is necessary to reach a satisfactory measuring strategy. And this is precisely the point. Finding out a

[8] For a fascinating history of the measurement of temperature, see Chang (2007).

measuring strategy that is adequate for our purposes is the result of our desire to improve our standards. Chang (2007) has described a similar idea as epistemic iteration. This is "a process in which successive stages of knowledge, each building on the preceding one, are created in order to enhance the achievement of certain epistemic goals" (2007, p. 45). This is how temperature was invented.

V. What is to be done?

There are at least two lessons we can draw from this exercise to help us modify and improve the way in which we measure "journal impact". The first is obvious. It is that for adequately measuring "journal impact", we first need to decide how to characterise it, represent it and establish procedures that do justice to these choices. If these three requirements do not line up neatly, we can't be sure that our empirical systems and the abstract entities by which we represent them correspond with each other. Reasons offered above suggest they do not correspond neatly. Something important to acknowledge here is that the effort to engage in this process of epistemic iteration must come from science. It will not come from Clarivate or Springer: their ethos is not that of science.

The second lesson we can draw comes from the exploration of the history of JIF. Above I said that the interest of Garfield was in the SCI. His interest with it was to allow scientists to keep scientific ideas in check. At the same time, he recognised that besides patrolling the cogency of scientific ideas, the SCI allowed scientists to know how their ideas travelled (Garfield, 1955). They could observe who else had profited from them and to discover other contexts in which those ideas had been put to use. This would, in turn, enhance communication between scientists and between fields. This is important because it makes us reconsider the purposes for which we measure journal impact.

While I haven't made it explicit, that purposes matter is an idea that has been lurking around. We measure for a particular purpose. We don't measure simply because we're interested in an indiscriminate accumulation of knowledge. Otherwise we would be counting our hairs and the weight of our clipped nails. And recording it. We measure because, in general, there is some higher purpose for the knowledge we garner. The measurement strategy, the fulfilment of the three requirements, has to be fit for this purpose. This implies that before we even start thinking about how to characterise "journal impact", represent it, and define the procedures to measure it, we first have to determine for what purpose.

It seems to me that the purpose Garfield had with the SCI was a noble one. Perhaps there's a way to keep some of the richness of his original idea in the definition of the purpose for measuring the impact of our scholarship.

VI. Conclusions

In this paper I have argued that Journal Impact Factor, the currency for measuring the impact of our scholarship is flawed. There are at least two problems. First, there is not an unequivocal characterisation of the concept of "journal impact". An exploration of avenues for characterisation doesn't take us very far either. Second, the way JIF is represented doesn't correspond to the kind of concept it tries to capture. Neither is this form of representation justified. Instead, the history of how the concept came about suggests that its current use is the result of a convenience trick that stuck. The suggestion I make on the basis of this exercise is that, if we're interested in actually measuring the impact of our scholarship, first we have to consider the purposes for why we do it. Then we need to conceive of a measuring strategy fit for these purposes. Otherwise we're fooling ourselves.

References

Adair, W. C. (1955). Citation Indexes for Scientific Literature? *American Documentation*, *6*(1), 31–32.

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00291

Buranyi, S. (2017, June 27). Is the staggeringly profitable business of scientific publishing bad for science? *The Guardian*. Retrieved from https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science

Cartwright, N., Bradburn, N., & Fuller, J. (2017). A Theory of Measurement. In L. McClimans (Ed.), *Measurement in medicine: Philosophical essays on assessment and evaluation*. Retrieved from http://dro.dur.ac.uk/19766/

Cartwright, N., & Runhardt, R. (2014). Measurement. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of Social Science: A New Introduction*. Oxford University Press.

Chang, H. (2007). *Inventing Temperature: Measurement and Scientific Progress* (First Edition). Oxford ; New York: Oxford University Press.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*, 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Clarivate. (2019). Journal Citation Reports. Retrieved 26 March 2019, from Clarivate website: https://clarivate.com/products/journal-citation-reports/

Editorial. (2005). Not-so-deep impact. *Nature*, *435*, 1003.

Efstathiou, S. (2012). How Ordinary Race Concepts Get to Be Usable in Biomedical Science: An Account of Founded Race Concepts. *Philosophy of Science*, *79*(5), 701–713. https://doi.org/10.1086/667901

Garfield, E. (1955). *Citation Indexes for Science. 122*, 4.

Garfield, E. (1963). *Science Citation Index* (Vol. 1). Retrieved from http://garfield.library.upenn.edu/papers/80.pdf

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. *JAMA*, *295*(1), 90. https://doi.org/10.1001/jama.295.1.90

Gross, P. L. K., & Gross, E. M. (1927). College Libraries and Chemical Education. *Science*, *66*(1713), 385–389. https://doi.org/10.1126/science.66.1713.385

Heinsohn, G. (2003). *Söhne und Weltmacht: Terror im Aufstieg und Fall der Nationen*. Orell Füssli.

King, C. (2017, January 24). *Journal Citation Reports: A Primer on the JCR and Journal Impact Factor*. Retrieved from https://clarivate.com/blog/science-research-connect/journal-citation-reports-new-primer/

Leydesdorff, L., Bornmann, L., Comins, J. A., & Milojević, S. (2016). Citations: Indicators of Quality? The Impact Fallacy. *Frontiers in Research Metrics and Analytics*, *1*. https://doi.org/10.3389/frma.2016.00001

Moustafa, K. (2015). The Disaster of the Impact Factor. *Science and Engineering Ethics*, *21*(1), 139–142. https://doi.org/10.1007/s11948-014-9517-0

Perez, O., Bar-Ilan, J., Cohen, R., & Schreiber, N. (2019). The Network of Law Reviews: Citation Cartels, Scientific Communities, and Journal Rankings: The Network of Law Reviews. *The Modern Law Review*, *82*(2), 240–268. https://doi.org/10.1111/1468-2230.12405

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, N.J: Princeton University Press.

Searle, J. R. (1995). *The Construction of Social Reality*. Simon and Schuster.

Tal, E. (2015). Measurement in Science. *Stanford Encyclopaedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2017/entries/measurement-science/

Web of Science Training. (2017, October 13). Journal Citation Reports—Journal Impact Factor—YouTube. Retrieved 5 April 2019, from Youtube website: https://www.youtube.com/watch?v=VJc3PC697oc&list=PLyh-Yuqjd7yqRabcyeChfycIdoVXgxyFI