

# Counting research $\Rightarrow$ directing research. The hazard of using simple metrics to evaluate scientific contributions. An EU experience.

*Kai A. Olsen*

MOLDE UNIVERSITY COLLEGE AND DEPARTMENT OF INFORMATICS, UNIVERSITY OF BERGEN


*Alessio Malizia*

HUMAN CENTERED DESIGN INSTITUTE, BRUNEL UNIVERSITY LONDON

Journal of Electronic Publishing

Volume 20, Issue 1, 2017

DOI: <http://dx.doi.org/10.3998/3336451.0020.102> [<http://dx.doi.org/10.3998/3336451.0020.102>]

 [<http://creativecommons.org/licenses/by/3.0/>]

*This paper was refereed by the Journal of Electronic Publishing's peer reviewers.*

## Abstract

*In many EU countries there is a requirement to count research, i.e., to measure and prove its value. These numbers, often produced automatically based on the impact of journals, are used to rank universities, to determine fund distribution, to evaluate research proposals, and to determine the scientific merit of each researcher. While the real value of research may be difficult to measure, one avoids this problem by counting papers and citations in well-known journals. That is, the measured impact of a paper (and the scientific contribution) is defined to be equal to the impact of the journal that publishes it. The journal impact (and its scientific value) is then based on the references to papers in this journal. This ignores the fact that there may be huge differences between papers in the same journal; that there are significant discrepancies between impact values of different scientific areas; that research results may be offered outside the journals; and that citations may not be a good index for value. Since research is a collaborative activity, it may also be difficult to measure the contributions of each individual scientist. However, the real danger is not that the contributions may be counted wrongly, but that the measuring systems will also have a strong influence on the way we perform research.*

**Keywords:** Counting research, h-index, journal publications, JCR, ranking

## Introduction

For centuries, the public image of a university professor has been someone in a dusty office with piles of books and papers. While he appears to be working to increase our knowledge of the world, we do not know exactly what he is doing or if it will indeed turn out to be useful. In many cases even his boss or peers may not have a precise idea of what he is working on. One possible justification for such a free system is that research cannot be controlled—we hope that some researchers will produce valuable results, some will end up as Nobel laureates, and some research results may even lead to a better world.

This image is now changing. Taxpayers want to know exactly what they are getting back for the large amount of money that is put into universities. Politicians demand that educational institutions like hospitals, schools, railroads, and garbage collection services, show that they offer value for the money they have been given. This is not easy for universities. While teaching can be evaluated by the number of students enrolled, exams taken, and degrees awarded, measuring research is a more intricate exercise. However, many national research organizations and universities are now introducing automatic metrics to measure research, and hopefully, also research quality (Paul, 2008).

On an individual level, researchers have always been evaluated by their peers to be hired, promoted, and awarded grants and positions. The disadvantage of reading and evaluating research contributions is that it is time-consuming and expensive. Sometimes it can also be influenced by personal likes or dislikes. In any case, in-depth peer evaluation is not feasible when the accumulated results of all the researchers at an institution are to be periodically evaluated, perhaps as a basis for funding or for ranking a university. To evaluate research on this scale, especially for evaluations that may be performed on a yearly basis, automatic methods seem to be the answer.

Ideally one would like to measure the impact of research from a socio-economic perspective, i.e., to measure what impact the research has on society (Penfield et al, 2014). One could, for example, measure the products that were developed based on the research, their economic impact, the number of jobs created by the project, etc. However, the causality between a piece of research and such effects may not be easy to determine, especially since it may be many years before the effects materialize.

Since publications are the most common form of research results, the majority of national evaluation systems are designed to count publications as a measure of impact. All the available data can then be put into a formula that produces one number for each researcher and one for the university as a whole. This could be as simple as counting results, publications, patents, and products. However, since a publication is seldom a product in itself, these numbers should express the *value* of the research. This is not easy to formalize, since value may have many aspects. Further, since each result is a part of a process, it may not even be possible to pin the value to a certain research effort.

Thus, in order to count research one needs more formalized methods where there is data available. Typically, the *impact* of each research publication is measured as the number of citations to the journal where the research results are presented. That is, one tries to measure the impact the publication had on the research community. Ideally a measuring system should be in the background. But there is a risk, at least where the results are important for individuals or organizations that they convert into systems that *direct* research. That is, researchers and institutions perform a form of “reverse engineering” where one analyzes the counting methods and then optimizes the results by producing the research contributions that offers the maximum score with the least effort.

We shall use Norway as an example. The country has established a list of journals that are high quality [<http://www.nsd.uib.no/>]—the level 1 journals. Publications in these journals are given one publication point. Of the complete set of journals, the 20% best are raised to level 2, where publications are given three points. Similarly, we are rewarded with 5 or 8 points for books. The system is also used in Sweden and will be introduced in South Africa. Basic research funding is based on the points collected by each institution. In addition, there is a clear pressure from the Ministry of Education that colleges and universities should generate more points, in order to get a stronger international reputation and increase the institution’s visibility.

The most difficult and time consuming way to achieve this is to publish papers in high ranked international journals; the easy way to obtain a comparable number of points is to reengineer the system. Doing this we see that many options for gaming appear. These are also used. Among the many research centers in Norway, the Center for Rural Research is ranked near the top based on its high amount of publication points. While they had a few papers in international journals many years ago, the majority of their research is not published internationally. Over the past few years they have concentrated mainly on publishing books in Norwegian. These efforts offer far better pay-back in points than international publications in high ranked journals. Similarly, over the last five years, half of the publication points within the general medical area are from publications in one journal, the *Journal of the Norwegian Medical Association* (published mainly in Norwegian). While these publications do not address the goal of an international orientation, they clearly offer an easy way to obtain more points.

In a new version of the counting system the ministry wants to give some incentive for international collaboration; instead of dividing the points between authors as the current system does, additional points are given if the publication is a joint international effort. However, this new system can also be reengineered. Haugen and Sandnes (2016) show that a paper will receive more points if it includes more than one author, especially authors from abroad.

## “Counting” becomes important

The ivory tower is crumbling. Universities are asked to produce value for money. With ideas from New Public Management, funding in many countries depends on the number of students enrolled, exams taken, and degrees awarded. In addition, there are requirements for publishing. For the latter, many researchers are provided with a list of journals that “count,” and also sometimes a scoring system in which publications in a smaller set of journals offer more points than others.

The authors of this paper work under different “counting” systems. The Norwegian author works under the system described above. For the UK author, the Research Excellence Framework (REF 2014) is important, both for himself and the institution where he works. Ideally this is a peer review system, where a panel of experts produces an overall quality profile for each submission. However, due to a combination of the workload involved and the resources set aside for performing the review, one ends up counting publications in good journals. For example, the 2014 REF [<http://www.ref.ac.uk/>] included approximately:

- 2,000 submissions
- 50,000 academic staff
- 200,000 research outputs
- 7,000 impact case studies

Thirty percent of the submissions were judged to be world-leading (4 stars), 46% internationally excellent (3 stars), 20% recognized internationally (2 stars) and 3% recognized nationally (1-star). For a research university to be able to compete efficiently for research funds a 4-star level rating is required, stressing the importance of publication in world-leading journals and conference proceedings.

We see these numbers are becoming important in evaluating researchers and research institutions in the EU and elsewhere, and they are often used directly to determine promotion, tenure or research grants. For example, a researcher may be told that her h-index, a measure of references, is too low for her to be a principal investigator for an EU project. Evaluation and promotion may be directly based on the h-index; the actual research contributions may not be required. As we see, countries and universities are making lists of journals that count. Many researchers receive a clear directive from management: publish—in these journals!

That is, the automatic counting systems offer clear rewards both for the individuals and the institutions that manage to get their numbers right.

## Measuring impact

So how do we measure impact? The most common systems define impact as an appearance in a journal that receives more than a certain threshold of citations. Numbers from Thomson JCR are commonly used; these figures are based on references from more than 9,000 journals. JCR counts the references from all publications in a given year to articles published the preceding two years in a particular journal in a particular year. This provides a journal impact value, which can be applied to all papers published in the journal.

Journal impact value can help researchers express the value of the publications, both for themselves and for their research institutions. The individual scores are often expressed as the h-index, the largest  $n$  for which the researcher has published  $n$  articles, each with at least  $n$  citations. Note that we have removed the actual publications from the process. The papers, their abstracts, and titles are not necessary in order to perform the evaluation. Everything can be computed based on the number of references or the name of the journals in which the papers are published. That is, we have defined that all publications in a journal have the same value, based on the number of citations that the papers in this journal receives.

Does this method work? In a paper from 2009, Adler et al. offered an excellent review of many of the drawbacks of using citation data. These included the following:

- There may be vast differences between the impact factor for a journal and the citation rate for a given paper in that journal.
- Impact factors vary from year to year.
- Impact factors cannot be used to compare different journals.
- The two-year scope for measuring citations may be too short for many fields. This is certainly the case today, when bibliographic databases and excellent search systems make older papers much more accessible than they used to be.
- The idea of measuring impact by counting references may be initially flawed, especially considering that references range from negative (criticizing a paper) to rhetorical (for example, referencing the paper in a research context) to an acknowledgment (expressing an intellectual depth to the cited paper).
- The h-index cannot be used to directly compare researchers (for example, a researcher with 10 papers that have 10 citations each will have the same h-index as another with 10 papers with 100 citations each).

Onodera and Yoshikane (2015) study how citation rates are influenced also by the type of article (e.g., short report) and the country and language in which the journal is published. Sanderson (2008) finds that h-indexes for a set of academics within Library and Information Science and Information Retrieval are dependent on which database, Web of Science or Google Scholar, was used for the calculation. Wainer et al. (2013) show how, in terms of publications, productivity and impact differ even within subareas of Computer Science; that is, the total productivity of researchers in Computer Architecture, Communications and Networking, Distributed Computing, and Image Processing and Computer Vision is significantly higher than those for researchers in Management Information Systems and Operational Research and Optimization. This implies that comparisons of impact factors for journals and h-index for researchers may be invalid even within the same scientific field.

Alan Dix (2015) found a similar effect when analyzing data from the 2014 REF, showing that some areas of Computer Science have many more 4-star papers than what one could predict from the overall amount of citations. For example, looking at the top 1% of cited papers, the probability that these are ranked as 4-star can differ by up to ten times for different areas. From his study it becomes apparent that the formal and theoretical areas are the winners and the applied areas the losers; that is, the bias is in favor of the old universities that tend to emphasize theoretical work, and against newer universities that may concentrate on applied areas.

## Variation from country to country

While the EU is trying to establish a common research area in Europe with institutions like the European Research Council and programs for funding that include all member countries (such as Horizon 2020 [<https://ec.europa.eu/programmes/horizon2020/>]), the evaluation system for individual researchers varies from country to

country. One of the authors of the present paper gained his PhD in Italy, held an academic position in Spain for six years, and then moved to another position in the UK. Interestingly enough, a publication that got a full score in Italy could have a much lower value in Spain. Moving from Spain to the UK was especially difficult, as the UK institution employs a different list of journals and conferences that count in comparison to Spain. By luck, he had publications in some of these journals. After living through these different evaluation systems, being one person with many different research profiles can feel somewhat schizophrenic.

The other author of this paper is evaluated under the Norwegian system. When coauthoring papers, we see that there are also discrepancies between the British and Norwegian systems. One of us may receive full points for a publication whereas the other may receive zero.

Differences from some selected journals and conferences are presented in Table 1. The selection is made among the journals that we would consider as relevant to our research. For each journal and country, the table indicates the score expected for a paper published in these journals.

**Table 1. Value of journal and conference publications in different European countries.**

Journal name	Area	UK	Norway	Italy	Spain
ACM CHI Conference	Human-Computer Interaction	3	0	3	1–2
ACM Communications	Computer Science	0	3	3	3
ACM Computing Survey	Computer Science	1	3	3	3
IEEE Computer	Computer Science	0	3	2	3
ACM Symposium on Principles of Programming Languages (POPL)	Formal methods	3	0	3	0–1
European Journal of Information Systems	Information Systems	3	3	3	2
IEEE Transactions on Software Engineering	Software Engineering	3	3	3	3
International Conference on Information Systems (ICIS)	Information Systems	3	0	2	0–1
JASIST	Information Science	2–3	3	0	0
Pattern Recognition	Artificial Intelligence	3	3	3	3
PloS Computational Biology	Biomedical and healthcare informatics	3	3	3	3
Scandinavian Journal of Information Systems	Information Systems	0	3	0	0
The Computer Journal	Theoretical Computer Science	3	1	3	2

While there is much in common between the various systems, there are also major differences. A quick comparison of the UK and Norwegian systems reveals sizeable discrepancies. While the UK REF system asks for originality, significance, and rigor, the Norwegian system will consider most journals with a high JCR ranking top notch, without regard to the depth or novelty of the papers.

The main difference between these two systems is that the Norwegian author has a formalized list with the name of each publication and its ISBN number and will therefore know in advance just how many points a publication will be awarded, while the UK author will be subject to a less mechanical and thus less quantifiable a priori evaluation. The REF will offer general criteria (rigor, novelty, and impact) used to provide a classification in 1, 2, 3, or 4 star for a paper; this will likely induce the researcher to aim at top journals and conferences to reduce the probability of getting a lower number of stars.

Norway has a list of 2,000 top journals, and Spain, which also relies heavily on the JCR values, has a similar number of journals listed. The Italian system is mostly based on the abstract and citation database Scopus. This should provide data similar to JCR and the Web of Science citation index. In practice, researchers in these three countries will have a long list of good journals to choose from.

As we have seen, the UK system requires papers to have a degree of originality to be recognized. Thus the *ACM Computing Survey* (Table 1), highly recognized in other countries, might be risky for a UK author since it will be difficult to prove the originality of a survey. In fact, UK authors inclined to make a good impression in the next REF should perhaps keep clear of survey papers at all. Similarly, *IEEE Computer* and *ACM Communications* are widely read journals, but some of the papers accepted here may not include enough “rigor” as expected for a top publication in the REF. However, for a Norwegian or Spanish author, the JCR of these journals will ensure a top score, while an Italian author would only get an average score for a paper in *IEEE Computer*. One reason why Sanderson (2008) got different h-index results for different databases was that Google Scholar included conference papers, while Web of Science did not.

*JASIST*, the *Journal of the Association for Information Science and Technology*, offers an interesting case. In 2014, it changed its name—“American Society” was changed to “Association.” The old *JASIST* was registered as “ceased” while a “new” journal was registered. The UK and Norwegian systems managed to move the impact factor from the old to the new, while the systems that are based on JCR and Web of Science did not manage to capture its history. Thus its impact factor is zero.

Publication in conference proceedings might also offer different scores depending on the country. For example, the ACM CHI conference is considered to have a large impact worldwide by all researchers within Human-Computer Interaction. This is supported by its low acceptance rate, which is generally considered a good measure of quality although sensibly questioned by Freyne et al. (2010). The conference is a candidate for the highest score in Italy and the UK, but a publication here only offers an average or low score in Spain. In the Norwegian system most of the conference publications are given zero points, including ACM CHI.

This lack of recognizing conferences will be a drawback for Computer and Information Science authors because conferences are more important here than in any other field (Freyne et al., 2010; Patterson et al., 1999; Fortnow, 2009). An alternative approach is to try to diminish the reliance on conferences. Moshe Y. Vardi, the editor in chief of *Communications of the ACM* (Vardi 2015) says that:

The underlying issue is that while computing research has been widely successful in developing fundamental results and insights, having a deep impact on life and society, and influencing almost all scholarly fields, its publication culture has developed certain anomalies that are not conducive to the future success of the field. A major anomaly is the reliance of the fields on conferences as the chief vehicle for scholarly publications.

Thus, if we agree with Vardi, the Norwegian system goes in the right direction.

If we look at all countries, journals seem the best bet. However, even here there are discrepancies. A paper in *The Computer Journal* will only be given one point in Norway while earning higher points in all other countries. Similarly, the *European Journal of Information Systems*, considered a high impact journal in many countries, will only score average in Spain. When co-authoring papers, authors will end up checking for those discrepancies between national evaluation systems. To avoid the situation where one author may get full points for a publication and another zero, one may end up choosing journals based on the point systems.

Note that Table 1 only recognizes the scores offered for different journals. There are other discrepancies as well. Some countries, such as Norway, would normally give points based on the journal only. However, in other countries, such as the UK, one would also determine the type of the paper. Thus an opinion paper, such as the one you are reading, may get a score of zero, even if published in a high-ranked journal.

These differences do not seem to be a part of any national research strategy. It seems to be influenced more by the choices that are made when formulating the evaluation schemes. Or perhaps these differences are just an indication of the haphazard nature of the whole evaluation system, which could be a symptom of using methods that are not fully understood. In any case, it is not clear why the European countries, with their centrally planned research funding, have so many different ways of evaluating researchers. The EU uses large funds to encourage cooperation and movement of researchers between member countries. In this respect these different national evaluation systems are counterproductive. It may be costly to try and move from one national evaluation system to another.

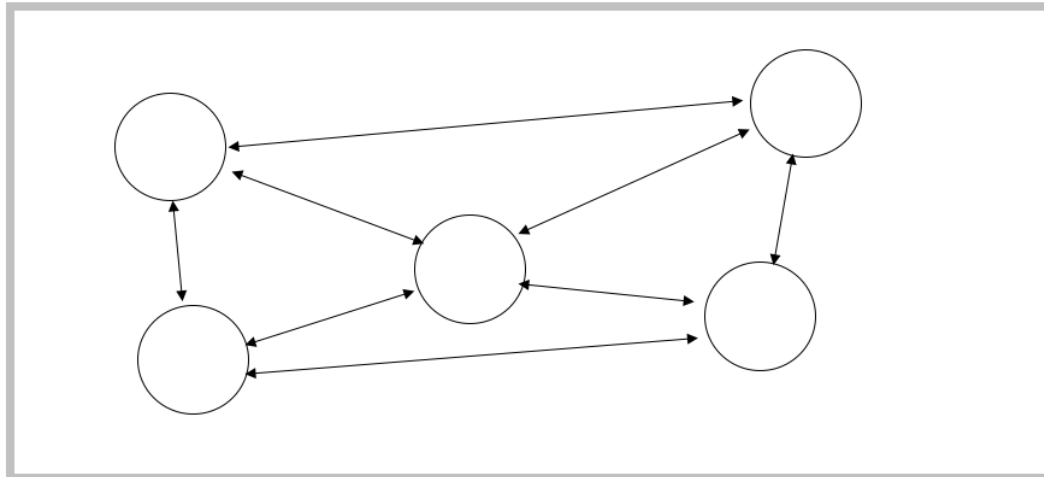
Reviewers for international journals, research grants, or academic positions may come from one country (or publication culture) and may not appreciate the profiles from researchers coming from another: there may be too many conference papers, too many opinion papers, not the necessary rigor, qualitative instead of quantitative methods, etc. These national discrepancies, in culture or in research measuring systems, may become apparent when researchers cross national borders, for example, when applying for an EU grant or submitting to an international journal.

While we could argue that diversity in research and diversity between nations may be an advantage, there is a need for a more international perspective. As research funds are moved from individual countries to the large EU grants, the systems for evaluating proposals need some degree of standardization. That is, the standardization work that the EU performs in many areas, from EAN numbers to road security, should perhaps also encompass research evaluation.

## Impact—only as citations

Since impact is related to citations, any form of impact that does not directly lead to citations is of no value—that is, it does not “count.” For example, papers that are read by non-academics and the general public will not usually result in citations and will therefore not have a countable impact. Papers that are downloaded often, are commented on in social media, or go viral on the Internet do not count. Papers published in journals with a low JCR value or in magazines will not be credited.

Researchers that have achieved a breakthrough may not always have the time to present rigorously scientific papers and submit them to a top journal. Gordon Moore presented the ideas behind his “law” fifty years ago. The prediction has had a widespread impact on the development of computer technology (Brock, 2006). However, as it was presented in a magazine, and more as a “viewpoint” than as a research paper, it would have been given a zero score in most systems today. Tim Berners-Lee made the program code for the WWW freely available in 1991 by posting it on a newsgroup. At that moment, no one was able to predict the impact that his invention was going to have. By sharing he offered the tool to technicians, researchers, and ordinary people who saw new possibilities of applying his ideas. The result, as we know, was dramatic changes to society, but it did not “count.”



[/j/jep/images/3336451.0020.102-00000001.png]

Figure 1.

*References in a sandbox.*

Even papers or journals with high JCR ratios may have little real impact. For example, we may find closed research communities, visualized as a sandbox in Figure 1, with communities of researchers that all reference each other’s papers within a small set of journals. While these journals and papers may obtain high impact ratios, the effect on society at large may be very small. This is not only a theoretical argument. When references can offer economic gain for institutions and promotion for individuals, authors will do everything they can to be referenced. This may be achieved, honestly, by ensuring that early versions of manuscripts are sent to all interested parties or, perhaps not as honestly, by having agreements to cross-reference with other institutions.

## Best bet—top journals

In this world of numbers, researchers who expect to move between countries, or face future evaluation schemes in which the list of journals that “count” may not yet be clear, will have to strive for publication in top journals. A subset of these top journals offers credit in all countries, which we can confidently expect will be part of any upcoming evaluation. However, this may lead us to a situation in which these journals will receive even higher numbers of submissions. The resulting low acceptance rate will then cause good papers to be rejected. This may result in a situation where being accepted may feel like winning the lottery. We see a trend in this direction today for certain high-ranked journals. This is not a sustainable system; many resources are wasted when perhaps only one paper out of a number of good papers is selected.

In order to select the good papers, careful reviewing will be necessary, but this may not be so easy to achieve. Universities that have lost state funding are trying to increase their attractiveness towards potential tuition-paying students. One way of achieving this is to offer higher service levels, such as weekly meetings between students and professors, e-mail consultations, and more lab hours. At the same time professors are expected to use more time for fundraising and, as we have seen, face stricter requirements for publication. A side effect of this pressure may be that there is less time for voluntary jobs such as reviewing. In other words, the publication system may bite itself in the tail.

A system in which only journals with high JCR values are counted will be static. Few researchers will be willing to devote their time and effort to publish in journals that do not count. This quest for publication in high-ranking journals will make it very difficult to establish journals in new areas of study that may open doors for new forms of presentations or present research results for new audiences.

## Can better tools be the answer?

New emerging tools using Web technologies (e.g. Research Gate, Google Scholar, academia.edu) might offer a more comprehensive suite of measurements that is less subjected to unwanted side effects. We have been offered many different methods for tracking, disseminating, and reviewing research. The tools that take ideas from social networks, such as Research Gate, are especially interesting.

Van Noorden (2010), in a paper in *Nature*, called this trend “a Cambrian explosion of metrics.” Bollen et al. (2009), who carried out a study on these multiple new impact measures says that “Scientific impact is a multi-dimensional construct that cannot be adequately measured by one single factor.” Indeed, Bollen’s claim is corroborated by Wouters and Costas (2012) who studied 16 different tools, including Google Scholar Citations, Microsoft Academic Search, and CiteUlike. They concluded that an effort is required to understand the dynamics and potential use of such new Web based tools in order to build new useful metrics for the scholarly community. However, when introducing new metrics, there is always the risk that a data-driven approach will replace an evidence-driven approach (Grimson, 2014). Indeed, a large number of indicators do not necessarily offer a better or more accurate estimate of research quality.

However, using more metrics and perhaps also changing metrics, may at least make the “reverse engineering” more difficult. This may diminish the controlling effect of the counting systems.

## The effects of counting

The idea is to count research, not direct research. As we have noted, there are indications that the evaluation systems are doing the latter. Most researchers publish to advance their career through tenure, higher positions, and research grants (Lyytinen et al., 2007). Rigorous formal systems for counting research may then lead to a focus on quantity rather than quality (Vardi 2015).

Butler (2004) shows that a performance-based publication system introduced in Australia, which counted publications independent of the source, had the effect that researchers optimized by seeking out sources that had high acceptance rates but often lower impact. As a consequence the national citation impact dropped.

The Norwegian performance-based publication system was introduced in 2004. To avoid the drawbacks of the Australian system, a two tier system with level 1 and 2 sources was introduced. Aagard (2015) found that there was no evidence that the Norwegian citation impact fell after the system was introduced; however, the new system did have an effect on citation quantity. Even if this performance indicator only accounted for the distribution of 2% of overall funding to universities and colleges, the number of publication points nearly doubled in the 8 years after the system was introduced (Aagard, 2015). While there may be other reasons for part of this increase, it is highly probable that the performance-based system has influenced researchers to publish more.

The Norwegian system was intended to work on an aggregate level. In describing the purpose of this system, it was stressed that the system should not be used to draw conclusions about individual publications, since high quality papers can sometimes be published in less reputable channels and vice versa. However, if the system is expected to have an impact on the aggregate level it must also have an impact on the individual level; in practice, the system could influence each researcher to publish more. Many institutions have understood this and offer monetary compensation to individual researchers for each publication point. A new variant of the counting algorithms in the system tries to focus more on institutions than on individuals; however, it is still possible to count on an individual level.

While the overall goal of the measuring systems can be to encourage more high quality research activity, we see that individual researchers will optimize their activities to get as many points as possible. Even if the Norwegian system manages to retain the impact of national research, it is still a system that offers incentives for quantity, along with all of the other counting systems.

## In the United States

The “publish or perish culture” is in many ways an import from the United States. However, the US system seems to be more informal compared to the ones being built in European nations. There is no official list. If you ask an American researcher he or she will point to the lists maintained by Microsoft Academic [<http://academic.research.microsoft.com/>] or Google Scholar [[https://scholar.google.com/citations?view\\_op=top\\_venues](https://scholar.google.com/citations?view_op=top_venues)]. The advantage of a more informal system is greater diversity and perhaps also a reduced effect of “directing.”

In “Incentivizing Quality and Impact in Computing Research”, Moshe Vardi (2015), editor in chief of *ACM Communications* highlights the effort of the CRA (Computing Research Association) in developing a new best practices memo for evaluating research (Friedman et al, 2015). This is directed toward hiring and promotions but can potentially have an impact on national evaluations. For example, the memo presents a recommendation for hiring, tenure, and promotion based on a very limited set of publications—between two and five. The idea is that “quality and impact need to be incentivized over quantity.” When only the most important publications are submitted it becomes possible for the hiring committees to actually read the papers.

Vardi calls for a statement signed by top scholars worldwide to foster the adoption of those practices. Interestingly, the memo calls for institutions to diminish the relevance given to research grants and focus more on the research the grants have helped produce; this could practically free researchers from the frenzy of selling for funding. Moreover, it criticizes the use of h-index and suggests that academia pay less attention to institutional rankings. In this respect, the CRA memo is a welcome effort to find better ways of evaluating research than just counting publications.

## Discussion and conclusion

The idea of evaluating such a complex area as research with seemingly objective numbers is alluring. When the results are converted into money, in the form of funding, promotion, rankings, etc., there will be a strong motivation to improve these numbers by performing “reverse engineering,” i.e., analyzing the counting systems and producing results that offer a maximum score. These incentives will be there for all parties, from individual researchers to university presidents. We also see that these ideas are adopted by PhD students. Instead of focusing on the important research questions many students, even from the very start of their research work, seem to concentrate on publication. The goal is no longer to increase knowledge, but to be able to produce a number of publications in top journals and conference proceedings.

Of the counting systems discussed here, all try to incorporate a measure of quality. As we have seen, the uncertainty around which journals count in the UK REF may cause researchers to target only a small set of “4-star” journals for potential publication. That is, the institutions want to be prepared for the next evaluation and the best bet is to go for the top journals. Since the evaluation, also for the UK REF, in practice will consist of “counting” there is a risk that papers published in other journals will not be included. Furthermore, as Dix’s (2015) analysis of the 2014 REF shows, the strong inter-area biases may well direct research toward the areas that are favored; this could mean that institutions may be tempted to adjust their hiring plans toward theoretical researchers that might have a higher probability of writing 4-stars outputs.

One advantage is that the number of papers submitted to the REF is limited. So while a researcher in Norway, Spain, Italy, or the United States is concentrating on getting as many publications as possible, the UK researcher will try to publish, hopefully 4 papers, for every REF in the small set of 4-star journals. With 6 years between REFs the researcher should produce about one paper every 1.5 years. Asking for a paper every 1.5 years is an improvement, but the number may still need to be lowered in order to let researchers concentrate on their research, not on their publications.

While the more informal system in the United States may be more difficult to “reengineer,” many of the systems used in the EU will, to a greater degree, direct the type of publications that are produced. The UK’s demand for “rigor” may steer researchers away from opinion papers, in the same way that “originality” may stop survey papers. All systems are open for “gaming.” Researchers learn to play the game—to publish the types of papers that are recognized by the counting systems and, in systems where quantity is important, divide publications into the “least publishable unit” (Vardi, 2015 [<http://www.apa.org/research/responsible/publication/>]). Not unexpectedly, some journals and publishing houses are also involved in “gaming” [<https://web.archive.org/web/20160606110024/https://scholarlyoa.com/2014/10/14/the-scientific-world-journal-will-lose-its-impact-factor-again/>] and use various methods to increase the impact factor of their products.

The question, then, is what is the correlation between the real value of research and the scores offered by the counting systems? We define real value as research that offers a better world, or—perhaps not so demanding—that has an impact on the world. As researchers we know that this implies hard work, patience, and perhaps also creativity. It is important to be critical, both of our own research and also that of others. This may not be compliant with trying to fulfill the requirements of the counting systems. That is, the current systems of measuring research may force researchers into a system that does not produce good research.

Terry Eagleton, a former Chair at Oxford University, writes about the “slow death of the university” and denounces the whole idea of counting research (Eagleton, 2015). He is afraid that “the institutions that produced Erasmus and John Milton, Einstein and Monty Python, capitulate to the hard-faced priorities of global capitalism,” and further that there is “plenty of reason for them to produce for production’s sake, churning out supremely pointless articles.” There may be a long way from the humanities division at Oxford to the research departments of lesser known universities and colleges, but the warnings that Eagleton offers seem to be regrettably accurate.

Nobel laureate Peter Higgs seems to agree with Eagleton. In an interview with the *Guardian* [<http://www.theguardian.com/science/2013/dec/06/peter-higgs-boson-academic-system>], he relates that he became an embarrassment to his department when they did research assessment exercises. When a message would go around the department saying “Please give a list of your recent publications,” Higgs would send back a statement: “None.” He also tells the *Guardian* that no university would employ him in today’s academic system because he would not be considered “productive” enough. He doubts that a breakthrough similar to the discovery of the Higgs boson could be achieved in today’s academic culture.



There is, of course, a middle way. By letting the counting systems be more in the background their tendency to control research may be diminished. For example, the direct connection between scores and funding, grants and positions should be removed and the actual research result should come in the foreground. While the UK REF has as the intention of measuring research outputs as “an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia,” it still looks at 200, 000 research outputs. The process of reviewing actual papers can only be achievable if there is a limit on the number of papers that can be submitted. Then the goal for each researcher would not be to produce yet another paper, but to produce *better* research and a *better* paper than what has already been published.

A similar policy can be used at the institutional level. Perhaps each institution should only submit their breakthrough research results for the reviews, concentrating on describing the effect of their research on society.

## Acknowledgments

The authors would like to thank Dr. Alan Serrano, Dr. Salvatore Sorace, and Prof. Juan Manuel Dodero for their insights respectively on British, Italian and Spanish national evaluation systems.

---

**Kai Olsen** is a professor in informatics at Molde College and at the University of Bergen, Norway and an adjunct professor at the School of Information Sciences, University of Pittsburgh. He has more than forty years of academic experience in the area of Information Technology and has published many books papers within this area. See for example Olsen, Kai A. (2012) *How Information Technology Is Conquering the World: Workplace, Private Life, and Society*, Scarecrow Press, December 2012, Lanham, Maryland, Toronto, Oxford, ISBN 978-0-8108-8720-6 (paperback) and 978-0-8108-8721-3 (e-book). He is also an enthusiastic skier and hiker, and has produced a large set of guidebooks for the northwest part of Norway ([www.turbok.no](http://www.turbok.no) [<http://www.turbok.no>]). Brittveien 2, N-6411 Molde, Norway. + 47 40287150. Kai.Olsen@himolde.no

**Alessio Malizia** is a Senior Lecturer at Brunel University London and a distinguished speaker of the ACM; he lives in London but is a “global soul” and has been living in Italy, Spain, and the United States. He is the son of a blacksmith, but thereafter all pretensions of manual skills end. Alessio began his career as a bearded computer scientist at Sapienza–University of Rome and then, after an industrial experience in IBM and Silicon Graphics, moved on with his career in research. He was a visiting researcher at the Xerox PARC, where he was appreciated for his skills in neural networks and as a peanut butter and chocolate biscuits eater. He worked as associate professor (and Spanish tapas aficionado) at the University Carlos III of Madrid. In 2012 he joined Brunel University London and the Human-Centred Design Institute. London, UB8 3PH, UK. +447450486664. Alessio.Malizia@brunel.ac.uk

## 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 (2): 106–17.
- Adler, R., J. Ewing, and P. Taylor. 2009. Citation Statics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). *Statistical Science* 24 (1): 1–14.
- Bollen, J., H. Van de Sompel, A. Hagberg, and R. Chute. (2009). A principal component analysis of 39 scientific impact measures. *PLOS ONE* 4 (6): e6022.
- Brock, D. C. (2006). *Understanding Moore's Law: Four Decades of Innovation*. Philadelphia, PA: Chemical Heritage Foundation.
- Butler, L. (2004). What Happens When Funding is Linked to Publication Counts?. In *Handbook of Quantitative Science and Technology*, eds. H. F. Moed, W. Glänzel, and U. Schmoch. , 340–89. Dordrecht, UK: Kluwer Academic Publishers.
- Dix, A. (2015). Citations and Sub-Area Bias in the UK Research Assessment Process [<http://ascw.know-center.tugraz.at/2015/05/26/dix-citations-and-sub-areas-bias-in-the-uk-research-assessment-process/>]. *Proceedings of Quantifying and Analysing Scholarly Communication on the Web*. Oxford, UK: ASCW15.
- Eagleton, T. (2015). The Slow Death of the University. <http://m.chronicle.com/article/The-Slow-Death-of-the/228991/> [<http://m.chronicle.com/article/The-Slow-Death-of-the/228991/>]
- Freyne, J., L. Coyle, B. Smyth, P. Cunningham. (2010). Relative Status of Journal and Conference Publications in Computer Science. *Communications of the ACM* 53 (11): 124–32.
- Friedman, B., and F. B. Schneider. (2015). *Incentivizing Quality and Impact: Evaluating Scholarship in Hiring, Tenure, and Promotion*. <http://cra.org/resources/bp-memos> [<http://cra.org/resources/bp-memos>]
- Fortnow, L. (2009). Viewpoint: Time for computer science to grow up. *Communications of the ACM* 52 (8): 33–35.

- Grimson, J. (2014). Measuring research impact: not everything that can be counted counts, and not everything that counts can be counted. In *Bibliometrics: Use and Abuse in the Review of Research Performance*, eds. W. Blockmans, L. Engwall, and D. Weaire, 29–41. Wenner Gren International Series, vol. 87. London: Portland Press Limited.
- Haugen, K.K., and F. E. Sandnes. (2016). The new Norwegian incentive system for publication: from bad to worse. *Scientometrics* 109:1299. doi:10.1007/s11192-016-2075-2
- Lyytinen, K., R. Baskerville, J. Iivari, and D. Te'eni. (2007). Why the old world cannot publish? Overcoming challenges in publishing high-impact IS research. *European Journal of Information Systems* 16 (4): 317–26.
- Onodera, N., and F. Yoshikane. (2015). Factors affecting citation rates of research articles. *JASIST* 66 (4): 739–64.
- Patterson, D., L. Snyder, and J. Ullman. (1999). Evaluating computer scientists and engineers for promotion and tenure. [http://cra.org/resources/bp-view/evaluating\\_computer\\_scientists\\_and\\_engineers\\_for\\_promotion\\_and\\_tenure/](http://cra.org/resources/bp-view/evaluating_computer_scientists_and_engineers_for_promotion_and_tenure/) [[http://cra.org/resources/bp-view/evaluating\\_computer\\_scientists\\_and\\_engineers\\_for\\_promotion\\_and\\_tenure/](http://cra.org/resources/bp-view/evaluating_computer_scientists_and_engineers_for_promotion_and_tenure/)]
- Paul, R. J. (2008). Measuring research quality: the United Kingdom government's research assessment exercise. *European Journal of Information Systems* 17 (4): 324–29.
- Penfield, T., M. J. Baker, R. Scoble, and M. C. Wykes. (2014). Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation* 23 (1).
- Sanderson, M. (2008). Revisiting h measured on UK LIS and IR academics. *JASIST* 59 (7): 1184–90.
- Van Noorden, R. (2010). Metrics: A profusion of measures. *Nature* 465:864–66.
- Vardi, M. Y. (2015). Incentivizing quality and impact in computing research. *Communications of the ACM* 58 (5): 5.
- Wainer, J., M. Eckmann, S. Goldenstein, and A. Rocha. (2013). How productivity and impact differ across computer science sub areas. *Communications of the ACM* 56 (8): 67–73.
- Wouters, P. and Costas, R. (2012). Users, narcissism and control—tracking the impact of scholarly publications in the 21st Century. Utrecht, NL: SURF Foundation. <http://research-acumen.eu/wp-content/uploads/Users-narcissism-and-control.pdf> [<http://research-acumen.eu/wp-content/uploads/Users-narcissism-and-control.pdf>]