European Review, Vol. 15, No. 3, 275–282 © 2007 Academia Europæa
doi:10.1017/S1062798707000300 Printed in the United Kingdom

Time for a rethink of research proposal evaluation?

DENIS WEAIRE

School of Physics, Trinity College Dublin, Ireland. E-mail: dweaire@tcd.ie

The processes of research proposals and evaluation have become ever more elaborate, but the system remains primitive. This paper argues for a fresh look at existing practices, with a view to developing less onerous and more efficient systems.

A waste of time?

Getting and spending we lay waste our days said a poet: it applies well to today's research scientists. Trapped in a maze of prescriptive procedures, they toil into the night to satisfy the ever-increasing and repetitive demands for proposals, reports and accounts. All this and taking the kids to the crèche, suing for divorce, and the other *petites commissions* of the modern lifestyle.

The evaluators who determine the fate of these scientists are not very comfortable either. Gone is any informality and the old-fashioned quality of trust. Judgement based on gut feeling and the wisdom of experience must be set aside, or veneered with explicit rationalisation addressed to specific criteria. Just like the proposer, they must slavishly fill every single box.

Is it not therefore time to ask whether this system is worth the pain and irritation that it engenders, whether it achieves its own aims, and what its side-effects may be? Does it really deserve the A+ that its widespread adoption might suggest?

Goals

In an age of thriving management consultants (some of them refugees from research), we might recall the question that they predictably and expensively ask their clients. *What are you trying to achieve?* They have taught us that there should be no perspiration without aspiration.

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press

By general consent we seek to advance the course of science (or more generally, knowledge). We want to promote excellence in its creation and development.

Part of our trouble comes when we add that it must 'serve the needs of society'. Ministers in some countries move quickly to this particular PowerPoint slide, which is likely also to contain the words 'national' and 'wealth creation'. It is hardly a new theme: an Irish physicist[1] said in 1894:

> It is but a short time since the pursuit of experimental research was regarded merely as a matter of individual curiosity; but owing to the high commercial value and important bearings of many of the recent discoveries in the fields of science, the public mind has now become awakened to the conviction that knowledge is wealth […]

Other societal motives also intrude, bearing on education, outreach, quality of life, the environment, gender, the cohesion of Europe, and so on.

Elaboratism

Elaboratism (a word coined by the author, without copyright) is the tendency to elaborate procedures and processes that could arguably be kept simple. There is an epidemic of elaboratism. In the case of evaluation, every officially endorsed aspect of scientific virtue is to be identified, codified and laid down in Times Roman 14pt.

This cast of mind comes naturally to lawyers and software designers (for whom a skilled touch of elaboratism can greatly enhance their income), and indeed some scientists. *Lord, tell me what I am to do, and I shall do it …*

It would be reckless to condemn the tendency in, let us say, the construction sector. We do not tell builders 'Go build something and we'll see how it stands up later'. Within the nominal research budget there will always be a portion that really pertains to development, surveys, or to engineering on a large scale, and it should be set aside in the present argument. We presume that something substantial is left for science.

The long tail problem

What remains is 'pure' and 'applicable' science, where intuition and insight rule, where serendipity sometimes blesses the prepared mind with the gift of a sudden discovery, worth ten thousand days of fastidious labour in the lab. Once recognised, these serendipitous events are celebrated with prestige and prizes. The trouble is that they are rare events, statistical outliers unrelated to the mean. We are confronted by a long tail of low probability.

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press

This phenomenon is not confined to research. In the development of technical start-up companies, it has to be recognised that most will fail, and that a few outstanding successes will be of overriding importance. The problem is: how can we spot them in advance?

It is much easier to measure the past than the future, but even that can be misleading, if simply extrapolated. There is a story about Shuji Nakamura, the inventor of the long-sought blue diode laser, told by Gerhard Fasol. Nakamura went to the founder and chairman of his company, Nobuo Ogawa, with his dream of creating the laser. He offered little by way of a track record commensurate with this grand challenge, and did not even have a PhD. The young man was given millions of dollars, considerable freedom, and some years to find the blue laser. He did so.

How on earth could you make such an extraordinarily bold decision, Fasol (himself a young man, seeking enlightenment) asked the chairman. 'How did you choose your wife?' was the reply.

Government agencies can hardly be expected to do this. There is no free hunch: something more analytical seems to be required. Why not a weighted average of measures of every conceivable aspect of merit?

Just add the numbers

Well identified desiderata having been identified and assigned their due proportions in the software, the evaluator (or a reviewer) must place in each corresponding box a numerical mark. These are then added, and the results are ranked. Second thoughts based on consultation are sometimes admitted. In the age of Google and Mathematica this is prehistoric mathematics, applied to the most sophisticated intellectual purposes. Why should a kindergarten formula, the linear sum of a few numbers, be fit for this purpose?

Of course there is positive merit in simplicity, but not in naivety. We must consider the consequences. In particular the object of scrutiny will react to it.

These days, the current generation of researchers is determined and skilled in the composition of proposals. Universities run courses on how to do it, how not to fatally miss a trick. Consultants stand in the wings, ready to do it for us, at a price. They have boilerplate texts in their laptops, sufficient to meet every case with a blend of the most effective superlatives, tested in the field. Hesitancy, modesty and sometimes honesty may be suppressed to give an impression of incisive science at the leading edge, serving all of our needs.

So at least superficially, the standard of proposals is consistently very high. Most evaluation is rather superficial. It becomes a struggle to discriminate, a search for the negative, for evident lacunae in some category that will drag the proposal down a little. In these circumstances, even the categories that carry lesser

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press

weightings can be crucial. Rather than belabour the point, one might consider the case of Einstein (or some other great scientist). In all but the category of sheer creativity, his profile was patchy. No chance!

It is as if God were to assign a weighting percentage to each of the Ten Commandments, with a score of 90% being obligatory for the status of angel. Many maverick martyrs would have to be ejected from heaven. Might not there be some better way of arriving at a conclusion than all this earnest toil? Could it not include as a dominant factor some deeper, more sensitive measure of present standing and past performance than the reviewer's quick scan of the standard buzzwords? This logic has, in particular, led the UK into the morass of 'research rankings', making matters considerably worse, at least in terms of the imposed burdens.

It is nevertheless an attractive notion that established excellence might accumulate credit and credibility towards future support, and the UK has been pondering how it might be done better and with less collateral damage to morale. This brings us quickly to the tricky topic of citations analysis.

Citations

Citation records, when they first appeared, were a tool for actual research, rather than the meta-research that concerns us here. They have rapidly become a prime tool of performance assessment. Sometime in the 1990s, as the originator of the citations business said,[2] 'the dam broke'.

Scientists rather enjoy this game (although citation envy must soon emerge as a recognised emotional disorder). In the US, the ranking of individuals is a national sport in every walk of life. In research, citations seem to provide an objective basis for it.[3,4]

Today, there are many different ways of compiling a citations analysis from the Web: they give different results. Impact factors and Web publishing without referees further confuse the issues.

But again, simple arithmetic is naïve. One often-cited paper, containing a useful but banal early definition or finding, might become a traditional landmark, widely used to signal where we are at. On the other hand, hundreds of trivial or routine papers, with a few citations each, might also gather a spurious measure of citational merit, when totalled.

In 2005, Jorge Hirsch questioned the use of a crude total.[5] His 'H-Number' has caused quite a stir, and some serious interest. A letter to Nature has pointed out that Sir Harold Jeffreys formulated a similar algorithm to monitor his cycling prowess, confirming that truly original ideas always arise in Cambridge.

What is the largest number N, such that I have published N papers with at least N citations? The answer is my H-number. Even the H-number is a crude, blunt tool, but it may point in the right direction. We can expect further refinements,

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press

such as higher-order indices (citations of citations?). The sceptical might regard this kind of thing as further elaboratism, but it is not of the kind that imposes much burden on the individual. (You can get your H-index by typing your name in a website, provided that it is unique in the scholarly world, and just clicking once. Brace yourself.)

Some sort of enhanced 'H-profile' could efficiently provide a 3D colour snapshot of an entire career, not easily altered or enhanced by the proposer, or veiled by the inflated 'leading-edge' rhetoric of relevance. It is an uncomfortable fact that a large proportion, perhaps even 90%, of scientific papers have *never* been cited, at least in terms of the present methods of citation recognition. They are 'born to blush unseen'.

Winner takes all

Our determined researcher, having met all criteria, may be one of the 20% that are to eventually receive blessings and contracts, from among the anxious cohort. What of the others? There are no silver or bronze medals in these games.

It may well be that parallel funding processes, pursued by other agencies, arrive at a similar selection, so that funding is tightly focused on particular individuals or groups. This concentration will be further enhanced by the publications that ensue, which go into the hopper for next year's applications, and so on. So nothing succeeds like success, and nothing fails like failure, in the game of grantsmanship.

Governments increasingly designate strategic priority areas of research, reducing the range of opportunities for many at the very outset. All of them set much the same priorities as their targets, independent of the peculiarities of their own economies, so local fashion cannot be evaded by flight.

Fair enough, our social Darwinists might declare, let the weak and the irrelevant go to the wall. But the proposers are mostly tenured full-time university staff, and will remain so. Without research funding what are they to do after lunch?

Frank Gannon has identified the problem in an EMBO editorial.[6] He praised the 'foot soldiers' of science, and spoke of their worth and their present neglect. It was the genius of the Duke of Wellington to recognise that his foot soldiers needed to be properly clothed and fed, if they were to fight – a novel idea among his brother officers. In an ideal system, one might look for a balance between the focused funding and a much broader, shallower system of research support. In a sense we have that already, in that lecturers' salaries are often included in such statistics as the OECD might gather to estimate national research budgets. From a lofty height, it all makes sense. But the discretionary funds that universities might apply to the costs of research by these unfunded lecturers have been squeezed dry in most countries, so there can be no balance. The foot soldiers have no boots!

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press

Peer review

Peer review is proudly paraded as the greatest single merit of the system, as it is in research publication. But consider the differences in practice. In publication there is often a dialogue between reviewer and author. This is generally ruled out in proposal evaluation. If authors do not eventually prevail against the negative judgements of a journal's reviewers, they are not without recourse. Instead of the *Journal of Bombastic Breakthroughs*, they can send their treasures to the *Annals of Mediocrity*. They will still be 'out there'. It may not even matter. Suppose Crick and Watson had not published in *Nature*?

Peer review of proposals has much more serious consequences, yet it is often as casual as the Emperor's languid thumb-down in the arena. This is likely to be so, because it takes a lot of time to do this job right, and it is at best only minimally remunerated.

'Sociologist fools physics judges'

This is a headline from an article in *Nature*, reporting on the work (some would say mischief) of a social scientist, Harry Collins. Collins, who has a layman's interest in gravitational waves, arranged for five questions on the subject to be set by an expert and then answered by himself and another expert.[7] Nine further researchers in the field were, on balance, unsuccessful in the game of Spot the Physicist, on the basis of these answers.

One must wonder about this troubling conclusion: would it have been so provocatively and prominently reported if the result had been less embarrassing? But it is certainly food for thought for those who conduct evaluation by peer review.

As it happens, the same issue of *Nature* (6 July 2006) also reported the first jail sentence given to a US researcher for 'making false statements to secure a $542,000 research grant'. Fabricated data and spurious claims are nothing new to science, but are they on a rising trend? If so, we need to be on our guard. The extraordinary ethical tradition of science is a priceless inheritance, but it is under siege.

To dwell on the odd case of outright fraud, however fascinating,[8] would be to miss the point. There is a wide spectrum of lesser and more typical disingenuousness of which these are the extreme examples, to which a few individuals are driven in the limit of high panic or perfidity.

Disingenuous discourse, which some philosophers are now calling *bullshit*[9] as a technical term, is common in the commercial world, to which we apply a different standard. What fool would ever *believe* a detergent advertisement? Yet a research paper, even emanating from the same detergent company, would be

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press

taken at face value, even before its validation by independent laboratories. Such validation is essential in principle but often difficult or impossible. Rigorous insistence on traditional ethical standards is a marvellously efficient convention, quite apart from any moral content.

What can we do?

This gloomy assessment should not conclude without some hope of improvement, but it would be hasty to list specific measures to be recommended at this stage. Rather, the problem should be recognised, and debate and analysis should follow. And in the meantime, we can say that whatever the trends, most evaluation is still conscientiously pursued within the limitations of the systems that are handed down.

The Academia Europaea organised a symposium in 2006 in Pavia, on the wider topic of Quality Assessment in Higher Education, and is considering the launching of a specific initiative on research evaluation.

The Pavia proceedings are about to be published.[10] They should provide a good starting point in developing an answer to the question we pose. How are we to optimise the output of research, and give our independent researchers the support they deserve, without guiding their weary footsteps each year down long paths of perceived righteousness towards an all-or-nothing outcome? It is a particularly daunting prospect for the young scientist: our system seems geared towards those in mid or late career. Must new ideas be conceded to a group or institute leader, in order to pursue them? The new ERC has already recognised this in biasing its funding towards younger researchers.

Acknowledgement

Thanks are due to Gerhard Fasol for advice, and to the Compagnia di San Paolo, who sponsored the Pavia meeting, which brought many of these thoughts to the surface in the author's mind.

References

1. T. Preston (1894) *Theory of Heat* (London: Macmillan).
2. J. M. Perkel (2005) The future of citation analysis, *The Scientist*, **19**, 24.
3. E. Garfield (2006)The history and meaning of the journal impact factor, *JAMA*, **295**, 90–93.
4. L. I. Meho (2007) The rise and rise of citation analysis *Physics World*, **29**, 32–36.
5. J. E. Hirsch (2005) An index to quantify an individual's scientific research output, *Nature*, **436**, 90.

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press

6. F. Gannon (2002) The foot soldiers of science. *EMBO Reports*, **3**, 1003.
7. H. Collins, R. Evans, R. Ribeiro and M. Hall (2006) Experiment with interactional expertise. *Stud. Hist. Phil. Sci.*, **37**(4), 656–674.
8. R Park (2000) *Voodoo Science: the Road from Foolishness to Fraud* (Oxford: Oxford University Press).
9. J. Holt (2005) Say anything, *New Yorker*, 22 August, p. 69.
10. A. Cavalli (2007) *Quality Assessment in Institutions of Higher Education in Europe: Problems, Practices and Solutions* (London: Portland Press).

About the Author

Denis Weaire is Erasmus Smith Professor of Natural and Experimental Philosophy at Trinity College Dublin, and a Vice-President of Academia Europaea. This article expresses his personal views, based on extensive experience of research evaluation.

https://doi.org/10.1017/S1062798707000300 Published online by Cambridge University Press