Is it any good?
Measuring scientific merit

A talk presented at the meeting of the German Physical Society (DPG)
Regensburg, 26 March 2007

Philip Ball
Nature
4-6 Crinan St
London N1 9XW
UK

Abstract:
How do we know if a paper is any good? How do you evaluate someone's scientific output? Which is more praiseworthy: a solid piece of work in an established field, or a stimulating new hypothesis which may or may not be right? Who are the 'best' scientists? The more deeply we probe into the question of quality in scientific research, the more contentious it becomes. That's why it is useful to have objective measures of quality, so that assessment – an important aspect of any human enterprise – does not become a lottery of personal opinion. The measures commonly in use are generally based on citation analysis. But it is widely acknowledged that merely counting up the number of papers in *Nature*, *Science* and *Physical Review Letters* is not the best way of quantifying quality. What alternatives exist, and how good are they? Can there ever be a one-size-fits-all metric of the merit of one's scientific output? In this talk I shall look at some of those that have been proposed, discuss what they tell us about the state of science (and of physics in particular), and ask where citation analysis – and the role of scientific publishing in general – seems to be heading.

–

The explosion of interest in the past several years in citation analyses, research assessment and the metrics that might be used for that, all hinge on a question that is perhaps not often voiced as explicitly as it should be: what is good science? I suspect most researchers would like to feel that, even if they cannot define it precisely, they know good work when they see it. But if we look at the various criteria that are typically used to evaluate research and allocate funding, the potential for disparity and contradiction is rather striking. For example, we might identify the following 'indicators' of 'good research":

- research that attracts funding, whether from the state or privately
- research that is highly cited
- research that other researchers agree is good (and those two are *not* necessarily the same)
- research that has the potential to lead to new technologies and economic growth
- research that leads to patents
- research that opens up new areas and poses new questions
- research that wins Nobel prizes

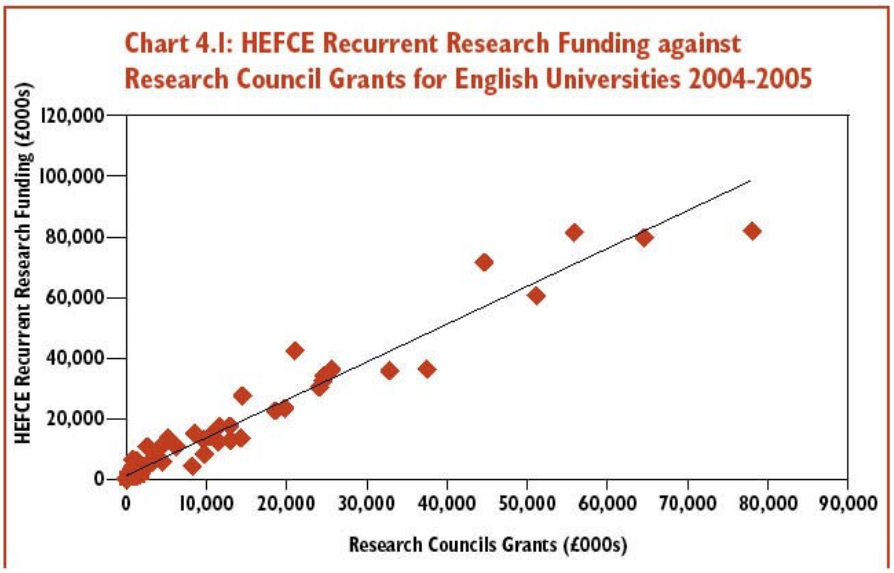
This is an incomplete list, but some of the potential answers that it omits are ones that are probably rarely considered, or easily overlooked – for example,

- research that inspires students or excites public interest
- research that takes risks
- research that can be applied in other fields
- research that benefits humankind

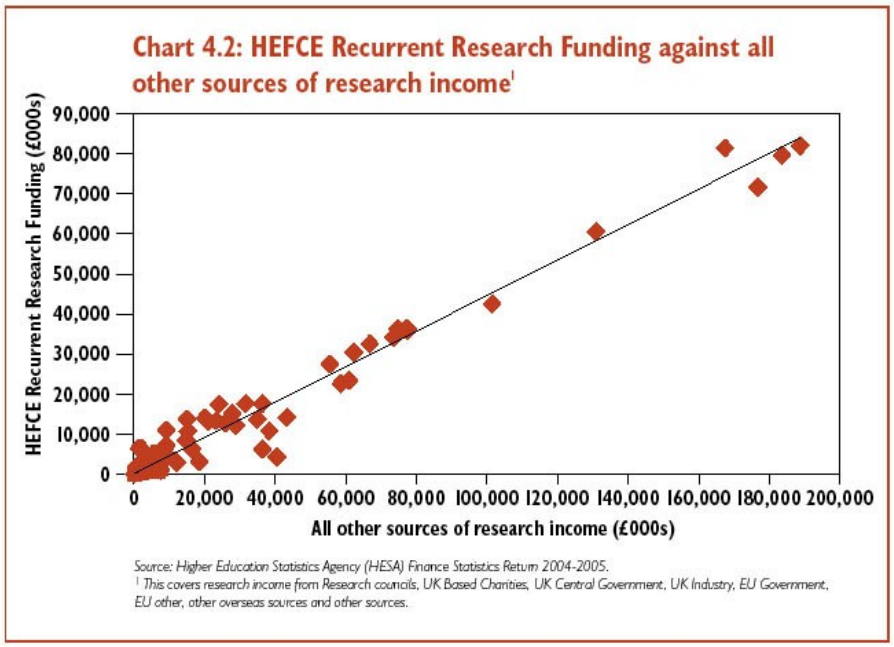
Of course, I don't mean to say that, for example, research that takes risks is necessarily good for that reason. But I'd suggest that some of these are factors that could weigh in favour of a project, if it is productive.

My initial point, then, is that while the discussion about assessing quality in scientific research typically becomes focused rather quickly on the relative merits of quantitative metrics and expert peer review, we need to acknowledge first of all that there are many different ways of defining what 'quality' means in the first place, before we worry about how best to measure it.

Let me give you one example of why these issues are vital to the way research is supported. I've taken it unashamedly from the country I know best, but at the same time I think this is one of the most striking examples of where thinking on research metrics may be heading. Last year the UK government announced an intention to do away with the current system of assessing the quality of university research departments, called the Research Assessment Exercise, which is highly dependent on peer review. They say that this system is administratively very expensive and is rather opaque, taking place behind closed doors. Instead, the government wants to introduce a system based almost entirely on metrics: "The Government's firm presumption is that after the 2008 RAE, the system for assessing research quality and allocating QR funding will be mainly metrics-based." (*Science and Innovation Investment Framework 2004-2014: Next Steps*, March 2006.) And what will that metric be? The government proposes that it be the level of funding that each department attracts, whether from applications for grants to the UK research councils, or from private funds coming from industry or elsewhere. In other words, the measure of quality is not what your peers think or how many *Nature* papers you have, but simply how much money you can attract. Now, this might sound pretty venal, but the government has what looks like a good argument. It points out that all the hidden, time-consuming and costly process of the current Research Assessment Exercise seems to do is reproduce the decisions made by the research councils when they come to award grants: the two outcomes are very highly correlated:



Moreover, that correlation holds up for all other sources of departmental income too:



Here, then, is a clear suggestion that quality assessment be based on a market-economy approach: the market decides what is good.

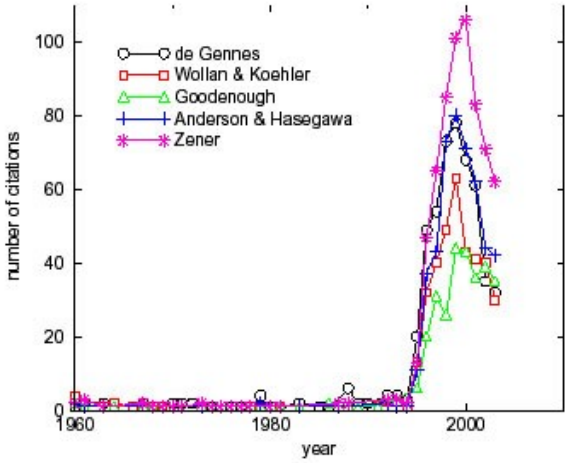
Now, this may seem anathema to some, but it is not hard to make the argument that this is not actually so different from the measure of quality long favoured by the research community and often used to evaluate questions not just of funding but of tenure and reputation – for what are citation figures if not a kind of market-based quality metric? The whole point of citation counting is that, once the paper is out there, it really doesn't matter if eminent Professor X thinks the work is a load of rubbish, because Professor X is just one voice among many, and he can't determine what the entire field thinks.

Well, let's just have a little look at what this market-based metric of citation statistics tells us. Sidney Redner at Los Alamos has analysed all the citation statistics from the *Physical Review* journals since it was begun in 1893, looking at all the cross-citations within *Phys. Rev.* to pick out which papers – and which scientists – were by this measure the most influential.

Publication	# cites	Av. Age	Title	Author(s)
PR **140**, A1133 (1965)	3227	26.7	Self-Consistent Equations Including Exchange and Correlation Effects	W. Kohn, L. J. Sham
PR **136**, B864 (1964)	2460	28.7	Inhomogeneous Electron Gas	P. Hohenberg, W. Kohn
PRB **23**, 5048 (1981)	2079	14.4	Self-Interaction Correction to Density-Functional Approximations for Many-Electron Systems	J. P. Perdew, A. Zunger
PRL **45**, 566 (1980)	1781	15.4	Ground State of the Electron Gas by a Stochastic Method	D. M. Ceperley, B. J. Alder
PR **108**, 1175 (1957)	1364	20.2	Theory of Superconductivity	J. Bardeen, L. N. Cooper, J. R. Schrieffer
PRL **19**, 1264 (1967)	1306	15.5	A Model of Leptons	S. Weinberg
PRB **12**, 3060 (1975)	1259	18.4	Linear Methods in Band Theory	O. K. Andersen
PR **124**, 1866 (1961)	1178	28.0	Effects of Configuration Interaction on Intensities and Phase Shifts	U. Fano
RMP **57**, 287 (1985)	1055	9.2	Disordered Electronic Systems	P. A. Lee, T. V. Ramakrishnan
RMP **54**, 437 (1982)	1045	10.8	Electronic Properties of Two-Dimensional Systems	T. Ando, A. B. Fowler, F. Stern
PRB **13**, 5188 (1976)	1023	20.8	Special Points for Brillouin-Zone Integrations	H. J. Monkhorst, J. D. Pack

Here's the ten most highly cited papers, and while there's lots of interesting things to say about them, you'll see that they are by no means all 'grand ideas' papers of the ilk of Bardeen, Cooper and Schrieffer's theory of superconductivity. Several, including the number one, are concerned with methods, in this case particularly with the methods that underpin density-functional theory, one of the most valuable tools of condensed-matter theory. This needn't surprise us. Writing 30 years ago, Eugene Garfield, who introduced the Citation Classics scheme in *Current Contents*, pointed out that his list of classics was dominated by methodology papers. Walter Schneider of the National Cancer Institute was quoted then as saying what seems of course to be obvious: "methods are the backbone of all scientific research". But of course this raises an important point, because methods papers aren't exactly the kind of thing that sets news networks buzzing. To put it bluntly, the can seem, and indeed often are, downright boring to anyone who doesn't need to use that method. Important work isn't necessarily eye-catching – and citation statistics, although much maligned, do at least have the virtue of pointing out the value to the community of work like this that might otherwise fall off the radar.

Redner's analysis also highlighted the fact that some papers become highly cited only long after they were first published. He identified several 'revived classics' that were barely visible until suddenly, decades after they appeared, new research made them significant. Here, for example, is a graph of the citations of five papers published between 1951 and 1960 on the electronic structures of perovskite manganites – not the most sexy of subjects, you might have thought, until colossal magnetoresistance was discovered in such materials in the 1990s:



Of course, it's not clear how *any* research metric might anticipate developments of this sort; but this perhaps serves as an indication of the need for diversity in research, so that the available resources are not dictated by current fashion. It goes without saying that these revived classics achieved that status not just because their subject later came into vogue, but because the work they reported was sound and insightful in the first place – *bad* papers on perovskite manganites would never have been revived.

This is fascinating, but of course it is a very incomplete story. Most obviously, it uses *Phys. Rev.* as a proxy for the physics literature overall – not a bad proxy, one would guess, but limited and no doubt filled with biases, most obviously towards the English language in general and the USA in particular. But there is also the factor that, to publish in *Phys. Rev.* journals, you first have to pass peer review. So this isn't a true 'market mechanism', but only one filtered by expert opinion. That's true in general of the scientific literature, of course, although preprint servers are changing it, and so might new publication models such as that being pioneered by the life-science journal *PLoS One*, which has undertaken to publish (online) every paper submitted, subject to a simple check of the methodology:
"Each submission will be assessed by a member of the *PLoS ONE* Editorial Board before publication. This pre-publication peer review will concentrate on technical rather than subjective concerns and may involve discussion with other members of the Editorial Board and/or the solicitation of formal reports from independent referees. If published, papers will be made available for community-based open peer review involving online annotation, discussion, and rating."
(from the *PLoS One* web site www.plosone.org.)

In some ways, this expert pre-filtering is precisely what is acknowledged by journal impact factors. In effect, these can be used to give greater weight in citation statistics to those papers that have been deemed to clear the highest hurdles, being published in journals that the 'market' has again identified as the most significant – meaning the most highly cited. The impact-factor weighting of citation is of course very often boiled down to a rather simple, even crude, formula which insists that, if you want tenure, you'd better have a few papers in *Nature, Science* or *Phys. Rev. Letters* under your belt.

Now, I have an instinctive tendency to deplore this sort of reductionist thinking, because

as an ex-editor at *Nature* I am all too aware of how much desperation it can instill in authors to push their paper through the peer-review process. But I deplore it mostly because I know that getting a paper in *Nature* or *Science* is a very limited and even misleading measure of anything. Firstly, those journals of course would like to assert that they are publishing the top papers in all of science, but the fact is that they are simply publishing what they hope is the best of what they receive. There are disciplines that do not tend to publish in those venues, largely because their peers do not read them. This means, contrary to what is often believed, that it can be easier rather than harder to get a paper into *Science* or *Nature* in a field in which the journal rarely publishes – not just because the journals can be prepared to lower the bar in order to enter a new field, but also because they have less of a sense of what is truly significant in such areas and because the quality of the reviewing can, frankly, be more indifferent. Beyond this, however, it should be obvious that *Nature*, *Science,* and not some extent *Phys. Rev. Lett.*, cater for particular kinds of papers, and of courser there is plenty of first-class work that would never really be suitable for or adaptable to such a publication outlet.

Citation statistics used to be dominated to the point of monopoly by the Science Citation Index, compiled by the American firm Thomson Scientific, formerly the Institute for Scientific Information (ISI). This is now an integrated online resource called the Web of Science (http://scientific.thomson.com/products/wos/). But it has competition from other web-based resources, such as Google Scholar and Scopus. Since these resources don't all search an identical literature data base (they may include preprint servers such as ArXiv, for instance, or book chapters and dissertations), they don't give identical results, and can sometimes differ significantly – researchers in some fields may have up to twice as many citations indicated by Google Scholar than by the SCI.

Quite aside from all of this, how trustworthy are citation metrics anyway as an indicator of quality? This is a subtle and difficult issue that is now becoming better understood thanks in particular to recent work on the growth and form of networks. It has become clear that the growth of a network such as that formed by papers linked via citations isn't a simple meritocracy. Whereas we're now used to the notion of web-based networks that are continually revised, updated and therefore in some sense optimized through constant use, citation networks are pretty much set in stone – no one can, even if they wanted to, alter a reference in one of their old papers because a better one has come to light. What tends to happen in a situation like this is the well-known Matthew effect, otherwise known as 'the rich get richer': the more citations you get, the better your chances of getting more. This means that the disparities between rates of citation of different papers don't necessarily follow in proportion to their differences in quality. In particular, standard texts or 'classic' papers are often cited just because others have done so, not because the researcher has gone and read them and made a considered decision that they really are the most appropriate.

That became clear in an analysis done by Simkin and Roychowdhury at UCLA in 2003, who used errors in citations as a measure of when a reference had merely been copied over from another paper: if you repeat the identical error, it's presumably because you never bothered to find the original paper. In this way, Simkin and Roychowdhury obtained a proxy for how often cited papers were actually read. They found that this

happened only in 20% of cases, and they estimated that fully 70-90% of scientific citations are copied from the lists of references used in other papers (M.V. Simkin and V.P. Roychowdhury, 'Read before you cite!' , *Complex Systems* **14**, 269 (2003) [cond-mat/0212043]).

Of course, if you haven't read the original paper, then by citing it you're not endorsing its quality at all, but only its fame. Lokman Meho at Indiana University, who has studied citation analysis, refers to such cases by the rather nice term of 'ceremonial citation'. This allowed Simkin and Roychowdhury to estimate what the citation statistics would look like if blind or random copying of other citation lists, rather than any genuine assessment of quality, were determining them. They showed that if scientists were to pick three random papers, cite them, and copy a quarter of their references, this would reproduce the empirically observed citation distribution. They say that "Simple mathematical probability, not genius, can explain why some papers are cited a lot more than the others." (M. V. Simkin & V. P. Roychowdhury, *Annals of Improbable Research* **11**, 24 (2005) [cond-mat/0305150].) This doesn't of course prove that really good papers don't get highly cited – but it says that things wouldn't look any different if that were so.

With the increasing prevalence of online literature sources, one alternative to citation statistics that avoids these ambiguities is to measure download statistics. If a paper is downloaded, there is a much better probability that it is actually read (though if you're anything like me, that's still not a guarantee). What's more, downloads pick up papers that have been used in some way even if they are not cited in that context. And download figures are instantaneous, rather than having to be compiled with a time lag like the ISI figures. Studies have shown that download counts do seem to mirror citation counts and impact factors rather well.

All the same, it seems there is a good case for looking for better metrics of research quality than mere citation or download counting weighted by impact factors. In particular, one can argue that good researchers and good departments are those that have a consistent record of high-impact research, whereas by today's methods a recent flurry of *Nature* papers may be all it takes to secure a good score. That, as we've seen, is part of the thinking behind Jorge Hirsch's h-index, which has stimulated a spate of proposals for other metrics that might pull signs of genuine quality out of the literature data.

Here are some of them:

a-index: Proposed by Jin Bihui, editor of *Science Focus* (B. H. Jin, 'H-index: an evaluation indicator proposed by scientist', *Science Focus* **1(1)**, 8-9 (in Chinese); 2006). It is equal to the average number of citations received by works in the 'h-index' publications. This aims to correct for the fact that the h-index does not really take into account exactly how many times the respective articles have been cited, except in terms of a minimum threshold. The a-index has been shown to be potentially over-sensitive to one or a few extremely highly cited articles (see R. Rousseau, 'New developments related to the Hirsch index'**,** KHBO (Association K. U. Leuven), Industrial Sciences and Technology, www.eprints.rclis.org/archive/00006376/).

g-index: Proposed by Leo Egghe, Universiteit Hasselt, Belgium (L. Egghe, 'How to improve the h-index', *The Scientist*, **20(3),** 14; 2006; and L. Egghe, 'An improvement of the H-index: the G-index', *ISSI Newsletter*, **2(1)**, 8-9; 2006). It is equal to the highest number of papers that together received g^2 or more citations. For example, a g index of 10 means that you have 10 papers that have together been cited at least 100 times. A scientist who writes many articles which are all fairly well-received will have a high h-index; a scientist who writes a few exceptional articles, while her other articles are hardly noticed by the scientific community will have a relatively low h-index but a high g-index.

Creativity index: Proposed by José Soler of the Universidad Autonoma, Madrid (www.arxiv.org/abs/physics/0608006). Soler's thinking is that a paper that has lots of references but only a few citations will have a low level of creativity, while a paper with just a few references and lots of citations will have a very high creativity. The creativity index is thus calculated from the number of references, *n*, that a particular paper makes to previous papers as well as the number of citations, *m*, that it receives from papers written at a later date. According to this count, the most creative physicists, in descending order, are:

1. P. W. Anderson
2. A. J. Heeger
3. E. Witten
4. S. Weinberg
5. M. L. Cohen
6. M. Cardona
7. A. C. Gossard
8. P. G. deGennes
9. M. E. Fisher
10. G. Parisi

How well do these measures discriminate between individuals? That hasn't been thoroughly evaluated. But using Bayesian reasoning applied to the SPIRES data base for papers in high-energy physics, Sune Lehmann at the Technical University of Denmark and his coworkers have argued that "compared with the h-index, the mean number of citations per paper is a superior indicator of scientific quality in terms of both accuracy and precision" (S. Lehmann *et al.*, *Nature* **444**, 1003; 2006 – see also www.arxiv.org/abs/physics/0701311). They also argue that at least 50 papers are needed to draw conclusions about long-term performance with satisfactorily small statistical uncertainties. It's not clear that this amount of data will be available for young scientists applying for tenure.

It seems the best one can say is that each of these measures works better under different circumstances, and none can capture the true picture of a scientist's 'impact' in all cases.

There is also the potential problem with the introduction of a transparent metric is that it sets people thinking about how to fix the results. This needn't be a matter of conscious corruption or rigging, although that certainly can happen, for example by self-citation. No, instead researchers may simply alter their publication habits with the metric in mind. Several years ago the introduction of a metrics-based funding system in Australia was found to produce radical changes in the publication behaviour of that county's researchers, while the lead-up period to the UK's RAE in 1996 was marked by an increase in the production of journal articles at the expense of conference proceedings. One of the reasons Jorge Hirsch proposed his h-index is that, relative to simple counting of publications or citations, it is much harder to manipulate.

So where does all this leave us?

The usual conclusion in this field is a rather anodyne one: that research metrics have their place, but can be no absolute substitute for expert peer review. Even the UK government's proposal for an income-based metric for assessing departments has basically that implication: it simply says that there is no point in duplicating the efforts of the research councils in their peer-review-based allocation of grants. A conference on 'the use of metrics in research assessment' at Wolfson College in Oxford in 2004 concluded much the same, saying that while productivity can be measured quantitatively through the analysis of publication and citation metrics, quality must be judged by others through a process of peer review. The participants of that meeting signed the following summary statement:

Publication and citation metrics play a key role in research assessment, in particular by supporting, challenging and supplementing more direct forms of peer review. The contribution they can make varies across disciplines; they have a role to play in all areas (including the social sciences, arts and humanities) but more direct forms of peer review are also needed in all areas to ensure valid assessment s of quality. The possible effects on behaviour of using particular metrics as indicators of quality or productivity needs to be considered carefully in advance.

I don't know that I would suggest anything so very different, except to say that, as I implied at the beginning, we should be clear about *why* we need some subjectivity as well as some objectivity. It is not simply a matter of there being no ideal metric of research quality that does not suffer from some shortcoming or other. Rather, we must recognize that there is no single definition of what quality in research is. As Lehmann *et al.* point out, "Institutions have a misguided sense of the fairness of decisions reached by algorithm, and unable to measure what they want to maximize (quality), institutions will maximize what they can measure." (*Nature*, 2006).

In that sense, we face the same issues as those that confront a democratic society: we can argue all we like about which system is 'fairer' or 'more open' or 'more accountable', but in the end we lack any consensus for what democracy itself is. And one thing is for sure: good research, like good democracy, is not something that can be engineered, let alone imposed: all we can do is find ways of making our system healthy and fertile enough that it has the opportunity, in its many manifestations, to arise.