Research Policy

Problems with peer review and alternatives

RICHARD SMITH

Many of the medical researchers I met were far from happy with the peer review systems used by many bodies funding research, but they didn't think that there was any workable alternative. There are, however, alternatives, and they are being used by some organisations in Britain and overseas. The main alternative is to use objective quantifiable measures of research performance, and much work has gone into developing and validating these measures.[1-4] The other alternatives, which have been much less explored and studied, are to distribute money in different ways, which include giving a little to every potential researcher, which is rather what has happened in Britain with research funds distributed through the University Grants Committee; trusting a "strong manager" to direct research, as happens often in defence and industrial research[5]; offering prizes for solutions to difficult problems[6]; and distributing funds based on a formula that measures past performance and productivity.[5 7]

Here I examine the criticisms of peer review, describe some of the alternative objective quantifiable measures that have been developed, and consider how the alternatives might be married together with peer review to create a better system.

Problems with peer review

LACK OF CONSENSUS AND RANDOMNESS

Lock has summarised the many criticisms made of the peer review system.[8] He concentrated mostly on peer review as used by journals, but the systems used by bodies funding research are similar: most of the deficiencies of the systems are likely to be shared. Firstly, referees often do not agree on the value of a paper or research proposal, and when they do agree it is most likely to be on what is bad. Cole and others took 150 proposals submitted to the National Science Foundation, half of which had been funded and half of which had not, and sent each proposal to about another 12 reviewers.[9] They found considerable variation in how the reviewers rated the proposals, and in about a quarter of the cases the reviewers would have reversed the decision on funding. Cole and others thus concluded that "the fate of a particular application is roughly half determined by the characteristics of the proposal and the principal investigator, and about half by apparently random elements which might be characterised as 'the luck of the reviewer draw.'" Neither Lock nor Cole and others are unduly dismayed by how important chance is in the system, arguing that although it may be hard on individual scientists it does not matter much to science as a whole.[8 9] The proliferation in Britain of different sources of funds for research seems to support this conclusion (if you are turned down by one source you can go to another), but randomness in the system may matter more if there is also consistent bias, as there may be[8 10]; thus some researchers and some sorts of research, including perhaps that which is most innovatory, face bias as well as randomness.

BIAS

Within the peer review system there may be bias against particular individuals, against certain subjects, towards well known researchers and against the poorly known, against certain research methods (particularly the unorthodox, which may include the brilliant and revolutionary), and against more peripheral institutions and towards the elite. Certainly some medical researchers are convinced that the MRC is biased towards Oxbridge and London and against provincial universities.

Institutional bias was illustrated by a notorious study of peer review as used by journals. The authors, Peters and Ceci, took 13 articles published in influential psychology journals by researchers from prestigious institutions, changed them, and then resubmitted them to the journals in which they had originally been published.[10] They changed the names of the authors and the institutions from which they came to fictitious ones and then changed the titles, the abstracts, and the opening paragraphs of the introductions. One paper had to be excluded from the study, three papers were recognised as resubmissions, and one was accepted, but the nine others were rejected. And they were rejected not because they were unoriginal but rather because they were poor papers. Peters and Ceci thought that their paper illustrated bias against the provincial in favour of the central elite. Lock has summarised the many criticisms of this study, not the least being its small size.[8] But many people have heard of this study and are convinced that the authors were right to use what have been called "unethical" methods to illustrate a bias that is hard to illustrate in any other way.

Cole and others in their studies did not find much evidence of bias.[9 11] In their first study of 1200 proposals they found that reviewers from prestigious institutions were if anything less rather than more likely to review favourably an application from another prestigious institution; nor did Cole and others find that professional age (length of career) had any strong influence. The funding decision was, however, moderately or weakly correlated with prestige rank of academic appointment, academic rank, geographic location, and place of PhD training. Overall, indeed, they did not find a high correlation between previous scientific performance and the grant application being successful. This was especially surprising as past performance is one of the evaluation criteria.

BIAS AGAINST THE INNOVATORY

The most worrying bias is that against the truly innovatory. Lock quotes the case of the unknown J J Waterson who should have been given the priority subsequently given to Joule, Clausius, and Clerk

British Medical Journal, London WC1H 9JR
RICHARD SMITH, BSC, MB, assistant editor

Maxwell except that his work was unpublished because a peer reviewer wrote that "the paper is nothing but nonsense."[8] Small took 73 papers that had received high citations (more than 10 in a year) and managed to get 50 usable referees' opinions from the authors: he graded the opinions from "publish unchanged" to "do not publish" and found that the most highly cited papers generally received the lowest evaluation by the assessors.[12] Small describes, too, the case of a standard reference (cited 751 times in eight years), of which the referee had called for drastic revisions unless unjustified conclusions were to get into published reports. The editor published the paper unchanged, but how many research directors have the courage to overrule grants committees?

That these poor opinions of excellent work should arise is not surprising, argues Roy, because there is no agreement on what is the best science.[13] Cole and others concluded from their study that: "Contrary to a widely held belief that science is characterised by wide agreement about what is good work, who is doing good work, and what are promising lines of inquiry, our research . . . indicates that concerning work currently in process there is substantial disagreement in all scientific fields."[9] Peers are better at agreeing on the importance of what happened 10 years ago than on the importance of what is happening now or what should be happening next.

SETTING THE WRONG PRIORITIES, NO DEFINITION OF A PEER, FRAUD

Roy, an American materials scientist, has broadened the attack on peer review, and one of his arguments is that the scientists who administer the peer review system follow (perhaps unconsciously) a code that rates "pure" (and often almost irrelevant) research highly and much more relevant research poorly.[13] His experience is that scientists prefer analysis research using the modern instruments over making real materials or systems, highly mathematical deductive work over experimentation and induction, and ephemeral theoretical computer simulations over painstaking measurements to provide the databases of modern science.

Roy frets too over the problem of who is a peer. Does, for instance, double Nobel laureate John Bardeen have a peer? Roy's next worry is that "the system . . . presupposes a level of objectivity, disinterestedness and honesty such as never obtained in any group." This concern, which is shared by Lock[8,14] and others,[15] extends beyond bias to fraud and plagiarism. Many examples of both have arisen from peer review and are now well described.[8,14,15]

COST AND DELAY

The fourth of Roy's arguments against peer review is perhaps the most substantial—its enormous costs in time and money. Roy quotes (without giving the data) an estimate that between a quarter and a half of the total intellectual time and energy of America's best scientists is spent on writing, visiting, discussing, reviewing, and serving on panels—that is, in the yoke of the peer review system.[13] The Nobel laureate Leo Szilard playfully suggested in 1961 that the day would arrive when 100% of the time of the scientific workforce would be spent in peer review.[16] Roy puts together the figures that an average grant from the National Science Foundation is $60 000 in some disciplines, that a full time academic costs $100 000 a year, and that two to four weeks are spent on preparing and following through a proposal to calculate that the Szilard point will be reached when the success rate of grant applications is one in 10—which it almost is for some subjects. Indeed, it may be worse: some requests for applications produce 30 applications and only one grant, meaning that the money spent on designing and reviewing the applications far exceeds the time spent doing the research. Turney has used figures supplied by the Natural Environment Research Council to estimate that research councils in Britain spend £4m a year processing about 6400 grant applications, well above the 1% of total turnover that the House of Lords select committee on science and technology suggested should be spent on research assessment.[17,18]

Hostile reaction to a study of peer review

Peters and Ceci conducted a study of peer review, which they think showed the bias against the provincial in favour of the central elite.[10] Afterwards they described the reaction to their study.[27]

"Our study seemed so straightforward and simple that we had often wondered why it had never been done before. We soon discovered a possible reason for the dearth of research on the peer review practices of one's own profession. Upon collection of the data we entered a period lasting approximately two years during which we experienced an intense and negative reaction from many powerful individuals in our profession for having conducted our study.

These personal attacks took their toll. For a couple of years we doubted the wisdom of our decision to do the research. Finally, after two unsuccessful attempts to publish our findings, replete with personally insulting, ad hominen reviews, we found a publisher and positive reviews. Soon press releases were telling a diverse audience our findings. Letters of support (over one thousand) came pouring in. Every one of them was complimentary. We realised for the first time in two years that the idea we had found so attractive so long ago was still an attractive idea to most people 'out there.'"

FINAL PROBLEMS

There are still further problems with peer review: the slowness of the system does not fit with the speed of scientific development—many of the best ideas cannot wait a year or more for funds; constant review takes a heavy psychological toll of research workers; and, Roy argues, the competitive nature of peer review conflicts with much of what is truly innovatory in science arising from collaboration.

Other ways of measuring performance

Dissatisfaction with peer review and the growing need to improve methods of distributing limited research funds have together stimulated the search for better ways of measuring performance. Most research has concentrated so far on finding, improving, and validating what are called "science indicators"—quantitative measures of the inputs and outputs of scientific research. The researchers who do this work (science policy analysts), many of them social scientists, have met sometimes with indifference and sometimes with hostility (see box) from the researchers whose work they are studying.

Many British medical researchers seem never to have heard of science indicators, and when they have they are worried that they will be used indiscriminately by science administrator/managers to stop the funding of particular projects. But this is to tilt at a false target, for science policy analysts have never suggested that the indicators should be used in this way; rather the indicators are to be used to add extra information and to complement peer review. It seems ironic that so many scientists should be unhappy with the development of science indicators when measurement is so central to science. It was the physicist Lord Kelvin who said: "When you can measure what you are speaking about and express it in numbers you know something about it."

BIBLIOMETRIC INDICATORS

Science policy analysts have mostly used bibliometric measures; Pritchard, who first used the term bibliometrics, described it as

"all studies which seek to quantify the processes of written communication" and defined it as "the application of mathematical models to books and other media of communication."[19] The idea behind scientific bibliometrics is the simple one that a piece of scientific work will result in a publication. Such techniques have been used for 70 years, including by De Solla Price in his classic essays of the '60s.[20,21] The enterprise became truly important, however, with the development of the *Science Citation Index* by Eugene Garfield in the '60s[1] and then of Computer Horizon Inc by Francis Narin.[2] (The database used by Computer Horizon Inc has been derived from the *Science Citation Index* by "cleaning it up" and systematising it—for instance, the *Science Citation Index* had about 140 different entries for Harvard University, whereas the new database has just one.)

Publication counts—Bibliometric techniques are now developing rapidly, but at their simplest they comprise a count of all publications by an individual or group and thus reflect productivity. The deficiencies in this simple measure are shared by many of the more complex ones.[3] Their first and biggest deficiency is that they are measures of past performance. This limits their value in decisions about future funding—except that past performance may, unsurprisingly, be the most powerful determinant of future performance. Secondly, they ignore methods of communication that do not depend on journals, which may be particularly important with more applied science, where the aim is a product or technique rather than an item of knowledge. Thirdly, publication practices vary among subjects, journals, and countries, so that a comparison of radio-astronomers in Italy with endocrinologists in California is likely to be meaningless. Fourthly, retrieving all the relevant papers and defining the boundaries of a subject are difficult: thus many bibliometric measures are heavily biased towards North America because they use American databases, which are more likely to collect data from North American (or at least English language) journals. Fifthly, difficulties are being created by papers having an increasing number of authors, some of whom have not contributed to the research.

Citation analysis—The particular deficiency of simple counts of numbers of published papers is that they give no idea of quality. Some measure of quality, however, is given by counts of the number of times a particular article is cited in other papers. (Most science policy analysts desist from suggesting that citations reflect the "quality" of an article—rather they prefer the term "impact." What exactly is being measured is one of the problems that occupies the analysts.)

The first problem with citation analysis comprises a series of technical difficulties such as only first authors being listed, variations in names, and authors with identical names. Secondly, for some work citation will begin quickly while for others it may be long delayed. Thirdly, some papers may be repeatedly cited because they are so wildly wrong and scientists are busy demolishing them. Fourthly, some papers and authors are cited not because they are directly relevant but simply because they give an aura of excellence. Fifthly, there is wide variation between subjects in the amount of citation—biochemistry papers tend to have about 30 references whereas mathematical papers have only 10. Sixthly, papers about methods are among the most widely cited papers while important theoretical developments may be quickly absorbed into the body of scientific knowledge. Seventhly, scientists may cite themselves and may even form what have been called "citation cartels." This last problem has particularly vexed critics of bibliometrics, but research has not shown this to be an important deficiency. Indeed, all of these deficiencies have been addressed and systems have been changed to accommodate them. None is now of great importance.

Validation of bibliometric measures—Studies that have attempted to validate bibliometric measures—by comparing them with subjective measures, the judgments of peer review systems, and other quantitative measures—have shown consistent and high correlation.[2] Nobel prize winners, for instance, are highly cited,[22] and subjective rankings of academic departments and institutions by peers correlates highly with their ranking by bibliometric methods.[2] Narin has reviewed 28 papers that compared bibliometric and non-bibliometric measures of research productivity and found high correlations.[2] He then did his own study on biomedical research and again found that bibliometric measures correlated highly with subjective measures, but he warned that bibliometric measures work best with large aggregates of data and are least reliable when judging the work of one scientist or a small group.[2] Science managers are, however, most interested in measuring the performance of individuals or small groups, and much work is now going on to develop such measures.

Measures of impact or influence—One of the advantages of bibliometric measurements over peer review is that once computer databases are established (as they have been) judgments can be generated much more quickly and cheaply. Counting all the citations of papers published by a group is, however, laborious, and it also necessitates waiting three to five years after the publication of a paper. One way round both these problems is to assign a paper a score based on the journal in which it is published, recognising that some journals are much more important than others.

Garfield has developed for journals a measure called the "impact factor," which is the ratio of the number of citations a journal receives to the number of papers published over a particular time period. The *Journal Citation Report* publishes impact factors for journals covered by the *Science Citation Index* and depends on the number of citations in that year to articles published over the previous two years. The problems with this measure are that it does not allow for the fact that as well as original research papers journals also contain varying amounts of review articles (which get highly cited) and other non-research articles (which usually get cited little), does not recognise that a citation in some journals is much more important than a citation in others, and does not cope with citation practices varying between different disciplines. Computer Horizons Inc has tried to get round these deficiencies by developing for each journal a "total influence indicator," which is the product of the "influence weight" of the journal (the weighted number of citations each article, note, or review in the journal receives from other journals normalised by the number of references that that journal gives to other journals) and the "influence per publication" (the weighted number of citations each article, note, or review in the journal receives from other journals).[3] Readers will see that this is becoming complicated, but bibliometric techniques have gone way beyond this measure in their sophistication and complexity. A recent clear and useful review of what they can achieve has been published by Jean King of the Agricultural and Food Research Council.[3]

OTHER SCIENCE INDICATORS

Many other measures of research productivity have been and continue to be proposed.[3] One obvious output of applied research is a patent, and patent counts have been used to compare the research productivity of different countries. Counts have also been made of the citations in patent applications and in the examiners' reports on them.

Another group of science indicators are called "esteem measures," and they include counts of honours (such as becoming a fellow of the Royal Society) and prizes (such as the Nobel prize), counts of invitations to give papers at international conferences or to edit journals, analyses of the migration of scientists, and measures of the ability to attract funding.

"FORESIGHT" INDICATORS

The great deficiency of science indicators is that they are retrospective, and much work is now being done to develop what have been called "foresight indicators," measures that will identify "hot" or "strategic" subjects. Bibliographic techniques—such as "cocitation analysis," "coword analysis," and "bibliographic coupling"[3]—have been developed to map the subjects and spot the topics that are developing fastest. Any analysis that depends on published papers is likely, however, to lag behind the fastest moving and perhaps most important developments—for instance, work on superconductivity. Science policy analysts are therefore looking at

studying conference proceedings, research proposals (most of which end up unstudied in the wastepaper bin), and requests to libraries for reprints. The British Library, for instance, receives about four million requests each year for scientific articles, and an analysis of these might show where science is pointing.

Putting it all together

Among science policy researchers peer review has been taking a hammering whereas science indicators are in the ascendant. Among active scientific researchers, however, there is affection for peer review and distrust of science indicators and their use in the allocation of research funds. The route forward is almost certainly to use, experiment with, and combine many different methods of measuring research productivity and allocating funds. Martin and Irvine have introduced the notion of "converging partial indicators," the idea of using several performance measures, including peer review and bibliometric measures, to make decisions on research funding.[23]

The table shows the deficiencies of some of the indicators and suggests ways in which they might be minimised.[3 23] Much attention has already been paid to improving peer review. The National Science Foundation, for instance, conducted an inquiry into its system and made nine recommendations including that the term "peer review" be replaced by "merit review" to acknowledge better that the decision to award a grant depended on more than the intrinsic technical excellence of the proposal; that the process be speeded up; that reviewers be given more feedback; and that the data system for tracking the process be improved.[24] King, meanwhile, in her review suggests that peer review systems be improved by giving researchers the right of reply, using peers from other disciplines and countries, giving clearer guidelines on the criteria the system is using, and using objective indicators to complement the process.[3]

Some institutions in the United States that fund research have been using a combination of peer review and objective indicators for some time; the Dutch have conducted an evaluation of their national performance in health research using several different indicators (and concluded that it is "solid but not brilliant")[25]; and in Britain the Agricultural and Food Research Council has begun to use other measures in addition to peer review. The Department of Trade and Industry is developing a science indicators network, and the new Centre for Exploitable Areas of Science and Technology is likely to have to depend heavily on science indicators to spot the areas to be exploited. Then the Science and Technology Assessment Office, which is part of the Cabinet Office, is very keen to encourage the new science indicators.

The Medical Research Council, meanwhile, continues to depend on peer review, but the appointment as deputy secretary of Dr David Evered, who has declared his interest in evaluation,[26] suggests that the MRC is likely to begin soon to take evaluation and the use of quantitative measures of research performance much more seriously.

Main problems with the various partial indicators of scientific progress and details of how their effects may be minimised (from Martin and Irvine, reproduced with permission of Elsevier Science Publishers)

Problem	How effects may be minimised
Publication counts	
(1) Each publication does not make an equal contribution to scientific knowledge	Use citations to indicate average impact of a group's publications and to identify very highly cited papers
(2) Variation of publication rates with specialty and institutional context	Choose matched groups producing similar types of papers within a single specialty
Citation analysis	
(1) Technical limitations with *Science Citation Index*:	Not a problem for research groups
(*a*) Only first author listed	
(*b*) Variations in names	Check manually
(*c*) Authors with identical names	
(*d*) Clerical errors	Not a serious problem for "Big Science" journals
(*e*) Incomplete coverage of journals	
(2) Variation of citation rate during lifetime of a paper—unrecognised advances on the one hand and integration of basic ideas on the other	Not a problem if citations are regarded as an indicator of impact rather than quality or importance
(3) Critical citations	
(4) "Halo effect" citations	
(5) Variation of citation rate with type of paper and specialty	Choose matched groups producing similar types of papers within a single specialty
(6) Self citation and "in house" citation	Check empirically and adjust results if the incidence of self citation or in house citation varies between groups
Peer evaluation	
(1) Perceived implications of results for own centre and competitors may affect evaluation	(1) Use a complete sample or a large representative sample (25% or more)
(2) People evaluate scientific contributions in relation to their own (very different) cognitive and social locations	(2) Use oral rather than written surveys so evaluator can be pressed if a divergence between expressed opinions and actual views is suspected
(3) "Conformist" assessments (for example, "halo effect") accentuated by lack of knowledge on contributions of different centres	(3) Assure evaluators of confidentiality (4) Check for systematic variations between different groups of evaluators

References

1 Garfield E. *Citation indexing: its theory and application in science, technology, and humanities.* New York: John Wiley and Sons, 1979.
2 Narin F. *Subjective versus bibliometric assessment of biomedical research publications.* Bethesda: National Institutes of Health, 1983.
3 King J. A review of bibliometric and other science indicators and their role in evaluation. *Journal of Information Science* 1987;**13**:261-76.
4 Narin F. Bibliometric techniques in the evaluation of research programs. *Science and Public Policy* 1987 April: 99-106.
5 Roy R. Alternatives to review by peers: a contribution to the theory of scientific choice. *Minerva* 1984;**22**:316-28.
6 Horrobin D. Glittering prizes for research support. *Nature* 1986;**324**:221.
7 Roy R. An alternative funding mechanism. *Science* 1981;**211**:1377.
8 Lock S. *A difficult balance: editorial peer review in medicine.* London: Nuffield Provincial Hospitals Trust, 1985.
9 Cole S, Cole JR, Simon GA. Chance and consensus in peer review. *Science* 1981;**214**:881-6.
10 Peters DP, Ceci SJ. Peer review practices of psychological journals: the fate of published articles, submitted again. *Behavioral and Brain Sciences* 1982;5:187-95.
11 Cole S, Rubin L, Cole JR. *Peer review in the National Science Foundation: phase I of a study.* Washington: National Science Foundation, 1978.
12 Small HG. *Characteristics of frequently cited papers in statistics.* Philadelphia: Institute for Scientific Information, 1974.
13 Roy R. Peer review of proposals—rationale, practice, and performance. *Bulletin of Science and Technology in Society* 1982;2:405-22.
14 Lock S. Fraud in medicine. *Br Med J* 1988;**296**:376-7.
15 Altman L, Melcher L. Fraud in science. *Br Med J* 1983;**286**:2003-6.
16 Szilard L. *The voice of the dolphins and other stories.* New York: Simon and Schuster, 1961.
17 Turney J. Showdown at Swindon: a peer review. *The Times Higher Education Supplement* 1988 January 8:6.
18 House of Lords Select Committee on Science and Technology. *Civil research and development.* London: HMSO, 1986.
19 Pritchard A. Statistical bibliography or bibliometrics. *Journal of Documentation* 1969 December 25:348-9.
20 Cole FJ, Eales B. The history of comparative anatomy. *Science Progress* 1917;**11**:578-96.
21 DeSolla Price DJ. *Little science, big science.* New York: Columbia University Press, 1963.
22 Zuckerman H. *Scientific elite: Nobel laureates in the United States.* New York: Free Press, 1977.
23 Martin BR, Irvine J. Assessing basic research: some partial indicators of scientific progress in radioastronomy. *Research Policy* 1983;**12**:61-90.
24 National Science Foundation Advisory Committee on Merit Review. *Final report.* Washington: NSF, 1986.
25 Rigter H. Evaluation of performance of health research in the Netherlands. *Research Policy* 1986;**15**:33-48.
26 Anderson J, Evered D. Why do research on research? *Lancet* 1986;ii:799-802.
27 Ceci SJ, Peters DP. Quoted in: Roy R. Peer review of proposals—rationale, practice, and performance. *Bulletin of Science and Technology in Science* 1982;2:405-22.

ANY QUESTIONS?

Ground linseed is used as a laxative. Is it safe and how does it act?

Linseed oil is used as a purgative in horses and cattle, and linseed itself has been used as a laxative in man in a dose of one or two 5 ml spoonfuls. The seeds are thought to act mainly as a bulk laxative but the fact that linseed oil itself has a laxative effect suggests an additional mechanisn. I can find no evidence that the oil has a stimulant effect on the bowel so perhaps it acts merely as a lubricant and faecal softener. Linseed does not seem to have any adverse effects when used medicinally.—LINDA BEELEY, consultant clinical pharmacologist, Birmingham.

Pharmaceutical Society of Great Britain. *Martindale. The extra pharmacopoeia.* 28th ed. London: London Pharmaceutical Press, 1982:696, 957.