The Natural Sciences and Engineering
Research Council of Canada

The National Natural Science
Foundation of China

PROCEEDINGS OF THE
SECOND INTERNATIONAL SYMPOSIUM
ON RESEARCH FUNDING

OTTAWA, CANADA

SEPTEMBER 13-15, 1995

The publication of these proceedings has been made possible
through the generous contribution of the

AECL
Atomic Energy
of Canada Limited

PERFORMANCE INDICATORS: KEEPING THE BLACK BOX OPEN

Yves Gingras
Université du Québec à Montréal, Canada

Since 1 have been asked, in this short presentation, to discuss the "limitations" of the use of bibliometric indicators for performance evaluation, let me stress first that I totally agree with the view that research must be evaluated. The fact that this is indeed a difficult task cannot be used as an excuse for concluding that it is not possible to evaluate research in any useful way. 1 think on the contrary that good indicators can be used to gain information on the impact of research, and that those indicators should be used as one source of information, among many, as an input into the decision-making process affecting the system of scientific research at a given level -institution, discipline, department, etc.

Govemment interest in the evaluation of scientific research, be it at the level of programs or institutions, has increased over the last ten years. Whereas in periods of economic growth it was easy to justify new programs of research, budget cuts have more than ever forced administrators and researchers promoting new programs not only to justify these new programs, but also to find ways to cut existing ones. In these circumstances, performance indicators are seen as a way to inform the decision-making process.

In practice, there is a strong tendency to reify a limited number of indicators that seem intuitive and appealing, but are in fact ill-conceived because they are not adequately weighted or normalized. We must resist this tendency to transform indicators into "black boxes" that are taken for granted and used uncritically. The use of quantitative indicators will never replace a decision-making process, which is in the end political. Performance indicators can shed light on the dynamics of scientific research, but they cannot serve as "expert systems" generating automatic decisions. Decision-makers must live with that fact, and assume their responsibilities.

The "Impact" of Science on What?

Yesterday, we talked a lot about the "impact" of research without clearly identifying the "target" of the "impact". Implicitly, most were referring to the impact on the economy, but one could also talk about the impact of research on science itself: does every research project contribute to the advance of science? One could also refer to the impact of scientific research on technological development or even, more globally, on society. After all, Albert Einstein or Stephen Hawking may not have contributed to a better economy (more efficient and productive, to use the fashionable language), or to "innovation," the new buzzword of politicians, but they have arguably transformed our vision of the universe. This is, if 1 am not mistaken, an important aspect of scientific research: knowledge for its own sake. In these times of economic disintegration, where governments want to limit scientific research to the solution of short-term problems, it is not superfluous to recall this long-term goal of university research.

So, there are different levels of the impact of scientific research, and I will limit myself to a discussion of the bibliometric methods used to assess the impact of scientific research on science itself. Francis Narin will show that one can also construct indicators of the relationship

between scientific research and technological development.

What Is an Indicator?

Since the use of indicators is always subject to controversy, let us first recall that an indicator is an index constructed to give "indirect" access to a complex (given or presumed) reality that cannot be apprehended directly. So, by its very definition, an indicator is never a direct and complete measure of that reality. This should be kept in mind by those who, being the reluctant subjects of evaluations, constantly repeat that the reality is more complex than the numbers suggest, that it is not a complete picture, and so on, as if we (the evaluators) were not aware of that. Of course, an indicator is partial, but nonetheless necessary since there is no direct access to the "phenomena" which one wishes to evaluate. Even the actors involved in the process to be evaluated do not have a "complete picture."

It is for this reason that one must develop a variety of indicators assessing different aspects of the "picture": there are indicators of "output," of "impact" (the two are different), of collaboration, networking, etc. Such indicators have been developed over the last 25 years, and they are well-known to the community of scholars devoted to evaluation studies. Their usefulness and limitations have been, and still are, discussed widely in the relevant journals. What we have been witnessing over the last ten years is simply the "discovery" of these techniques by decision-makers in search of evaluation methods. Although what we will say now about bibliometric methods is not new to experts, it is useful to remind those who intend to use such indicators for practical purposes, rather than for research, of the limitations of these methods, and thus pre-empt the use of a "black-box" approach to the use of these methodologies.

Bibliometric

In the case of university research, it is generally agreed that the main output is still publications. For that reason, bibliometric indicators are generally used to assess the value of university research. Of course, universities also train graduate students and produce inventions and patents, but I will not discuss here the kind of indicators that could be used to assess these aspects of university research.

The first problem that we encounter with the use of the two major bibliometric indicators (publications and citations) as it has developed over the last 25 years, is that their construction occurs in reverse to the "textbook approach" to indicators. Usually, one starts with a concept (impact, quality, etc.), identifies its dimensions, and then searches for indicators with those dimensions and combines them to get an indicator. This indicator is thus a "constructed" index of the concept to be measured. In the case of papers and citations, we have numbers, as it were pre-constructed, to which we then assign a meaning. The number of papers is of course a direct measure of output, but is certainly not a direct measure of impact or quality. Furthermore, although a good indicator ought to have a consistent meaning, sociologists continue to debate the meaning of citations.[1]

Once we decide to use bibliometric indicators to assess scientific research, we must take into account the objectives of the research program or institution being evaluated. Strategic research dedicated to understanding fish behaviour in the ocean in order to better manage the fish industry may of course lead to important publications in reputable scientific journals, but such publications are certainly not the primary objective of the research program, so that one cannot blindly use publication data to evaluate the impact of that research. However, it does make sense to use publication data to evaluate on a comparative basis long-term fundamental

research in astrophysics or elementary particle physics.

Once the objectives of the research to be evaluated have been established, one must choose an appropriate database for conducting the bibliometric analysis. Though most people are familiar with the ISI database, it is far from being the only one, or the most comprehensive in terms of journals covered. Depending on the field under study, one could use INSPEC, MEDLINE or the French database PASCAL, to name but a few of the existing databases, as the source of bibliographic information to be analysed.

Since each database has its own characteristics, it will never be possible to use pre-packaged datasets to evaluate an institution, a discipline, or a program. This explains why it is always costly and time consuming to collect data. And since the quality of the evaluation depends on the quality of the data used to construct the indicators, one must be vigilant in collecting data for the evaluation process. In these matters, one should not forget that everything is in the details - this is why we must keep the "black box" open. As Sylvan Katz showed yesterday for the case of scientific research in Britain, a large amount of cleaning had to be done on the ISI data base before it yielded useful indicators; the cleaning process took them two years. The message to would-be users of indicators is clear: there are no fast and ready-made reliable indicators that can be bought off-the-shelf.

Though papers may measure sheer output, quality is the most important aspect, and the most elusive and difficult to measure using indicators. Nevertheless, a large amount of research over the last 25 years has shown that citation analysis can be used to assess the relative impact of research in the scientific community. In fact, a whole journal, *Scientometrics,* has been devoted to that enterprise since 1979.

Since the citation process is very complex and multifaceted, the use of citation data to assess "quality" and "impact" is much more delicate than the use of publication data. In this case, the quality of the indicator depends crucially on the way it is weighted and normalized to arrive at a comparative evaluation among comparable groups or institutions.

To put the matter in concrete terms, let us take the example of the list of universities ranked according to the impact of their research, often found in *The Scientist* or *Science Watch,* both published by the Institute for Scientific Information (ISI).[2] The most often-used indicator is the "Impact factor" (IF), defined as the ratio of the number of citations (c_i) received by a given university (i) divided by the number of publications (n_i) produced by that university, for a given period of time.

$$IF_i = c_i / n_i$$

This, however is a very crude indicator, and it is not adequate because it does not take into account the fact that the impact factor is specific to a domain of research and varies greatly from one speciality to another, reflecting the level of activity and competition in each speciality. For example, Table I gives the 8-year impact factor for cited papers in some physics specialities for physicists in Quebec universities over the period 1945-1978.

Table I

Specialty	**8-Year Impact Factor**
Meteorology	3.01
Atomic Physics	5.58
Nuclear Physics	5.70
Optics	4.93
Particle Physics	6.99
Mathematical Physics	4.41
Astrophysics	8.74
Solid State Physics	7.07
Plasma Physics	4.65
Total Average	**5.92**

This table shows that, on average, research in meteorology is less frequently cited than research in astrophysics or elementary particle physics. What is true for different specialities inside a given discipline is also true for different disciplines, so that a direct comparison of the citations received by a researcher in mathematics to those received by a medical researcher has no meaning.

In order to obviate the problem of the differential activities of universities in different disciplines, we can define a normalized and weighed impact factor ($NWIF_i$) of a university (i) in a given discipline as the sum over specialities (j) of the ratio (c_{ij}) /n_{ij}) of the impact of a university (i) in a speciality (j) to the average impact of that speciality for all universities, weighed by the proportion of papers published by that university in each speciality (a_i).

$$NWIF_i = \Sigma_j \frac{\left\{\frac{c_{if}}{n_{if}}\right\}}{\left\{\frac{\Sigma_i c_{if}}{\Sigma_i n_{if}}\right\}} a_{ij}$$

This indicator is a measure of average relative activity of a university over all the specialities in which it is active. If a university is active in a sector with a low impact factor, but its impact in this sector is above average, that university is performing well in that sector, and its NWIF will be larger than 1.00. A university having a below-average impact in a sector with a high impact is not performing as well as the first university, and its NWIF in the specific sector will be less than 1.00. If a university has an average impact factor in all the sectors in which it is active then its total NWIF will be 1.00.

If we were looking at many different disciplines instead of specialities, the same indicator could be defined. Table 2 shows the effect of thus correction on the ranking of the three main universities for the 8-year period mentioned earlier.

Table 2

University	IF	NWIF
McGill	6.32	1.05
Montreal	5.97	0.89
Laval	5.29	0.95

Whereas McGill takes first place using either factor, the correction reverses the order of Montreal and Laval. We could construct this measure for five-year periods to obtain the evolution of the impact over the whole period.
This was just a technical example to show that the construction of a homogeneous indicator is at the heart of the process of evaluation, and that quick and dirty indicators can be obtained but can give a biased view. And when decisions are based on such biased indicators inequity, if not disaster, is bound to follow.

Conclusion

The main objective of this short presentation was simply to alert decision-makers to the fact that the validity of performance indicators depends on mundane technical details like the quality of the data bank used (liability of names, address, journal title, etc.) and, more importantly, on the way in which the indicator is constructed. It would be a major error to think indicators can be obtained quickly and cheaply. Each time one looks at a given indicator, one should always begin by asking how it was obtained. Though the tendency to reify numbers and forget about the way they are processed and produced is very strong, it is only by keeping the "black box" of performance indicators open that one will make sure that results are meaningful in a given context and that they can be discussed and improved in a rational manner, rather than presented as if they were the ultimate words of a judge pronouncing a verdict. In short, it is worth remembering the old motto of data analysts: garbage in, garbage out.

Notes

1. For a recent discussion, see Cozzens, Susan E., "What do Citations Count? The Rhetoric-First Model," *Scientometrics*, vol. 15, Nos. 5-6, 1989, pp. 437-047; Amsterdamska, 0. and L Leydesdorff, "Citations: Indicators of Significance?", *Scientometrics*, vol. 15, Nos. 5-6, 1989, pp. 449-071.
2. See for example Science, vol. 256, 10 April 1992, p. 175, and for Canadian universities and chemistry see *The Globe and Mail*, March 30, 1992. Conscious of the limitations of this indicator, the authors of the analysis of Canadian chemistry add the caveat that "the study tends to undervalue contributions of schools with strong chemical engineering departments whose work was generally cited less than half as frequently as papers from other branches of chemistry." This kind of reservation does not give more credibility to the results.