Scientific performance indicators: a critical appraisal and a country by country analysis

Michel Gevers

Department of Mathematical Engineering,
ICTEAM, Louvain University, Avenue Georges Lemaître 4,
B-1348 Louvain-la-Neuve, Belgium
Email : Michel.Gevers@uclouvain.be
http://perso.uclouvain.be/michel.gevers/

Abstract

This chapter consists of two parts. First we give a critical appraisal of the scientific performance indicators that have been introduced for the evaluation of individual researchers, journals, scientific institutions and countries. Their uses and abuses are discussed, as well as the negative effects they have on the research strategies developed by the scientific community. In particular, we highlight the perverse strategies that some journal editors have developed in recent years to increase their impact factor. Finally, we attempt to describe what the qualities of a performance indicator should be. In the second part, we compare the performances of a range of countries on the basis of normalized citation numbers. This comparison includes a cost-benefit analysis, in the sense that the number of citations per country is normalized with respect to the budget invested by that country in scientific research at higher education institutions.

Introduction

The evaluation culture has invaded the university environment over the last twenty years. University administrators have often adopted standard “business models” based on competition and comparison, as if university education and research was to be treated with the same rules and criteria as any private company on the market. Students are described more and more as customers, and universities compete to attract academics that have the best potential to increase their position in the international rankings. Academics and administrators of universities, and research funding agencies spend more and more time doing evaluation and selection, and they want tools that simplify their tasks. As a result, a whole new research area has opened up for the production of performance indicators at all levels of these evaluation procedures.

In the scientific world, performance is evaluated and comparisons are made at essentially four different levels: the individual researcher, the scientific journal, the university or research organisation, and the country. At each of these levels, performance indicators have been produced that are supposed to evaluate the quality, with the view of facilitating the tasks of the overburdened evaluators. A frantic search is going on at these different levels for the design of a single number that would characterize quality. These indicators are continuously being revised, refined and improved. “Any number beats no number” seems to be the new motto. The availability of these indicators is leading to an oversimplification of the quality assessments that are performed at these different levels. In addition, the emergence of every new indicator brings with it the development of feedback strategies whereby the researchers, the journals, the institutions and the countries aim at increasing their relative position as reflected by the indicator rather than pursuing sound and long-term research strategies.

In the first part of this chapter, we examine the leading indicators that are presently used at these four evaluation levels and illustrate some of the obnoxious feedback strategies that they have generated and their consequences. We then attempt to describe what the qualities are of a good performance indicator, and spell out a number of precautions and recommendations that should be adopted when using these indicators.

In the second part of this chapter, we perform a comparative analysis of the scientific performance of 17 countries that are known to be very active in research, where the performance is measured by citations. In order to compare countries that have vastly different populations and financial means, we introduce three different normalizations. The first consists of computing the number of citations per document, the second relates the number of citations to the budget invested by each country in fundamental research, while the third is based on a so-called Normalized Index (NI) that takes account of the size of the institutions, the discipline and the time period.

The four levels of performance evaluation

Scientific performance is evaluated at essentially four different levels: the researcher, the scientific journal, the university or research organization, and the country. Here we briefly present these four levels of performance evaluation, and the most prevalent performance indicators that have been introduced at these levels. In the next section, we describe the feedback effects that these indicators have induced in the scientific community, and the uses and abuses of these indicators.

The researcher is evaluated first and foremost at the various steps of his/her scientific career: at the hiring stage for a post-doc or academic position, at the promotion stage to a higher grade or position, at the application for a grant or for a scientific prize. Whereas in the old days most researchers tended to accomplish their career at the place where they first obtained a tenured position, nowadays an ever-growing fraction of the academic community tends to shift allegiances and to move along to positions considered to be more rewarding or more prestigious. In this pursuit for "excellence" the market value of the researcher becomes the key ingredient. Presently this market value is essentially assessed by the number of publications, the "quality" of the journals in which these papers are published, and the impact of the publications as measured by the number of citations. To facilitate the task of the assessors, some indicators have been devised that are supposed to aggregate with just one number the research output and performance of the candidate. The most famous of these indicators is the Hirsch index [1], best known as the *h-index*. A researcher has an h-index of 20, say, if 20 of his/her publications have been cited at least 20 times.

The enormous pressure to "*publish or perish"* has led to an explosion of the number of published papers together with an explosion of the number of journals, of widely varying quality. The competition between the journals has become fierce, and the need has arisen to try and evaluate the "quality" of these journals. This has led to the creation of the infamous *journal impact factor*. It has also led some large research organizations to propose a classification of all journals in different categories. For example, the Australian Research Council (ARC) has proposed a classification of thousands of scientific journals in three categories. This classification has been widely used around the world, but after a couple of years and widespread criticism, the ARC has abandoned its use.

Ever since the publication of the first Shanghai rankings in 2003, the universities have also entered into the game of competition and comparison. A number of competing international rankings have flourished, most of them focusing on research performance with a heavy bias towards natural sciences and medicine, at the expense of humanities and social sciences. The most widely cited rankings nowadays, besides the Shanghai ranking, are those established by the Times Higher Education, Leiden University, US News and World Report, as well as the new U-Multirank whose development is funded by the European Union. One measure of the failure of these rankings to create a consensus about a quality measure for the universities is that they are constantly being revised, making it very difficult to distinguish a trend in the evolution of the performance of any given university.

Finally, a fourth level of performance evaluation has recently made its mark: the comparative analysis of scientific performance of countries, based on aggregates of the performance indicators of their scientific institutions. These comparative analyses are being used by some governments to reorganize their scientific policies, or to reshape the landscape of their scientific institutions. The second part of this chapter will be entirely devoted to such country-by-country performance evaluation.

Engineering the indicators: the feedback effects and their consequences

Every new indicator, at whatever level, brings with it adaptation strategies. As soon as the research community finds out that a particular performance criterion becomes dominant in the evaluation committees, some of its members adapt their research and publication strategy to maximize its impact on the newly adopted criterion rather than aiming at producing the most highly creative research results. This *feedback effect* can have very negative consequences, to the extent that it may threaten the credibility of the research community and the foundations on which the pursuit of scientific research are established. Even though most of these indicators are very recent (the famous h-index was invented as recently as 2005), their widespread use at all levels of performance evaluation has produced, in a very short time, deleterious effects that are already very visible. In this section we analyse some of these effects, and their potential long-term consequences, on the quality of the produced research and on the undermining of the integrity of the research community.

At the individual researcher level, the focus on publication and citation numbers has resulted in an explosion of the number of papers. Authors, particularly at the early stage of their career, rush to publication. They tend to split their research findings into several papers rather than writing a comprehensive paper, thereby harming the pedagogical quality of their publications. Authors revise their accepted papers even if they realize that in the meantime their results have been superseded by some other researcher. The most important objective for a young researcher is not so much to produce a high quality paper but rather to build up a CV and publication list that will impress their next evaluation committee. But the most deleterious effect of this obsession with indicators is that researchers, whether young or more mature, tend to stay away from long-term or risky research topics for two reasons: long-term projects do not lead to large numbers of papers in a short time, while risky and entirely novel topics will typically not be cited widely until years later because the research community in these new topics is still non-existent at the time of publication.

The impact factor (IF) of a journal is based on the number of citations of papers published in that journal over a period of two (IF2) or five (IF5) years after publication. It is an extraordinarily poor measure of "quality" for different reasons:

- It measures the instantaneous (or very short term) impact of a paper, which has very little to do with its quality. Seminal papers are those that have a lasting impact, i.e. that are still being cited ten or twenty years later.
- The impact factor of a journal has more to do with the duration of the reviewing process and the publication process than with the impact on the research community. If most journals within a scientific discipline take an average of two years for the reviewing plus publication process (as happens in a number of disciplines) then these journals will have an IF2 that is close to zero. This, it is impossible to compare disciplines that have very different delays for their reviewing and publication processes. One way to increase the impact factor is to reduce the reviewing time by relaxing the quality criteria, thereby leading to a decrease in the quality of the journal.
- The pressure of the journals to increase their impact factor eliminates visionary or ground breaking research because an author who is ahead of his/her time and ventures into virgin territory will typically not be cited until years later.

The use of the Impact Factor as a supposed quality criterion for the evaluation of journals is where the manipulation and abuses is taking the most enormous proportions. To illustrate this, and at the same time highlight the argument in the last item above, let us cite some recent instructions to authors of a well-established engineering journal, the IEEE Transactions on Industrial Electronics:

"Review criteria: (1) Likelihood for citations of the manuscript.
Our current impact factor is 3.439 and this means that each paper is cited by journal papers at least 5 times within a couple of years after publication. Such expectation is currently the major criterion for review.

FAQ: How many references are needed?
We usually expect a minimum of 20 references, primarily to journal papers.... Please be sure that you have current references (last couple of years). If there are no current references, then there are two possibilities:

- *Authors are not following literature*
- *There are no papers because other people are not interested in the subject*

Both reasons are good enough for manuscript rejection."

This has led a number of Fellows of the IEEE to send a protest letter to the Board of Directors of the IEEE.

At the level of universities and research organizations, enormous efforts and large budgets are spent by some universities to improve their positions in the world rankings. Engineering the position in the rankings sometimes takes precedence over the pursuit of the university's stated objectives. For example, some universities are hiring prestigious academics for short-term visits at huge fees, while requesting them to put the university in their affiliation in all their future publications. Other universities are mandating their newly hired academics, who are on tenure-track positions, to publish at least so many articles in their first three years.

The availability and widespread dissemination of international rankings has led some governments to enter into the evaluation frenzy and to reassess their higher education policy and its organizational structure. In Italy, the Ministry of Higher Education has spent large amounts of money to evaluate each Italian researcher using a combination of indicators and peer review assessments. Alarmed by the relatively poor position of the French universities and research organizations in the international rankings, the French government has reorganized the university landscape by creating large conglomerates under the belief that size matters, i.e. that large institutions will score higher in the international rankings.

In this brief review of the uses and misuses of research performance indicators, we have argued that the engineering of these indicators at the various levels of evaluation has a significant impact on the type and quality of the research, and on the integrity and credibility of the scientific community. It also has a huge cost in terms of the man-hours invested in the various evaluation procedures and in the manipulation of the indicators.

Qualities of a good performance indicator

Research performance indicators are the subject of intense discussion and activity within the scientific community of experts in bibliometrics and scientometrics, and within a number of international organizations. Specialized journals are devoted to the topic, such as *Scientometrics, Research Evaluation, Journal of the American Society of Information Science*. International organizations have been set up that are entirely devoted to the topic, such as the International Ranking Expert Group (IREG). They publish policy statements, such as the Berlin Principles on Ranking of Higher Education Institutions [2], established by IREG in May 2006. A more recent initiative is the San Francisco Declaration on Research Assessment [3], also known as DORA, elaborated in December 2012; its main recommendation is that journal-based metrics, such as Journal Impact Factors, should never be used to assess individual scientist's contributions, or in hiring, promotion or funding decisions.

What are the qualities of a good performance indicator? In [4] three criteria have been proposed that are considered as necessary conditions for the validity of an indicator.

1. Adequacy of the indicator to the property it is supposed to measure.

The level of investment in Research and Development in a country is a good measure of the intensity of research activity in that country, but it cannot be used as a measure of the quality of the research. Similarly, the number of Nobel prizes received by graduates or academics of a university does not reflect the quality of the education at that university today, because the last Nobel Prize may have been awarded decades ago.

2. Sensitivity to the intrinsic inertia of the object

Universities have a huge inertia; their quality cannot change dramatically in a year. Therefore an *annual* ranking in which a university moves in a single year by 5 or 10 places shows that the indicator is defective, not that the quality of that institution has plummeted or raised dramatically. As argued by Gingras [4], *annual* rankings of universities can therefore only be explained by marketing strategies of the ranking organizations; they serve no scientific purpose, but they absorb a lot of resources from the universities that have to produce the data.

3. Homogeneity of the dimensions of the indicator

A homogeneous indicator of the research output could be the number of papers produced. However, if one combines the number of papers with a citation measure (as is done in the h-index) then one obtains a heterogeneous indicator. The same occurs with indicators that are based on a weighted average of different indicators. Quoting from [4]: *"Combining different indicators into a single number is like transforming a multi-dimensional space into a single point, thus loosing nearly all the information contained in the different axes".*

The following are two additional criteria that should in my view be applied in selecting indicators.

4. Insensitivity to small variations in the data: small numbers must be avoided

A good indicator should not change substantially if one of the input data changes by a small number. For example, one additional Nobel Prize or the death of one Nobel Laureate in a university does clearly not change the quality of this university and should therefore not change its ranking substantially. A corollary is that indicators should not rely on small numbers.

5. Normalization with respect to field, time period, and size

It makes no sense to compare research output in different fields or at different periods using the same indicator, because the publication culture varies widely over disciplines and because the level of activity in different topics can change rapidly. In addition, the sizes of the research communities in different disciplines are vastly different. In mathematics, most papers have single authors, while in nuclear physics it is not uncommon to have more than one hundred authors on a paper; the research community on social networks was very modest ten years ago but the field has become enormously popular in the last decade; it makes no sense to compare the h-index of a young post-doc with that of a senior academic. Attempts have been made to address those problems by introducing normalized indices. One such normalized indicator, at the institution level, is the so-called Normalized Impact (NI), introduced by the Karolinska Institute in Sweden. It compares the scientific impact, measured by citation numbers, of an institution with the world average in the same scientific domain and over the same period, while taking account of the size of that institution.

Research performance of countries, measured by citations

Against the above backdrop, we now evaluate and compare the research performance of 17 countries that are considered very active in fundamental research, on the basis of the number of citations of papers produced within these countries. The number of citations of a document is indeed one important indicator of the impact of a research paper or book within a discipline. The 17 countries selected in this study are Australia, Belgium, Canada, China, Denmark, Finland, France, Germany, Israel, Italy, Japan, Netherlands, Spain, Sweden, Switzerland, UK, USA.

In keeping with the remarks made about the qualities of a performance indicator, we introduce three types of normalizations which allows us to compare countries of very different sizes and whose domains of excellence may vary widely. Thus, for the 17 countries selected for this analysis, we look at the number of citations per document, the number of citations of a country versus the budget invested by that country in fundamental research, and finally the Normalized Impact (NI) mentioned above.

The analysis has been performed using data available from the websites of the World Bank[1], the OECD[2] and SCImago[3], a website that is specialized in the ranking of journals and countries on the basis of citations of papers and that is powered by SCOPUS[4], a data-base of scientific documents and citations maintained by Elsevier. This comparative analysis of the scientific performance of countries was inspired by Giuseppe De Nicolao, a founder of the website ROARS (Return On Academic ReSearch)[5], whose help is gratefully acknowledged.

The SCImago website allows one to compute citation numbers of documents published in all possible countries over the period 1996-2011, or separately for each year of that period. The number of citations given for a specific country and for a particular year X is the number of citations of all papers published by authors who work at an institution of that country during the year X and cited during the years X, X+1, X+2, etc, until 2011. Thus, the citation numbers for the year 2009, say, refer to the number of citations of all documents published in 2009 and cited in 2009, 2010 or 2011. When referring to the period 1996-2011, all documents published during that period are considered.

Citations per document

The first normalized data represent the average number of citations per document for the 17 countries involved. Figure 1 shows the ranking of the 17 countries in terms of the number of citations per

[1] http://data.worldbank.org
[2] OECD= Organisation for Economic Co-operation and Development: http://www.oecd-ilibrary.org/
[3] http://www.scimagojr.com
[4] http://www.scopus.com/
[5] http://www.roars.it/

document over the long period (1996-2011). In order to evaluate whether this indicator has evolved over the years, these data have also been computed separately for the years 2010 and 2011. This allows one to check whether the relative positions have evolved over this 15-year period. Of course, one should bear in mind that the variance for these much shorter periods is significantly larger than for

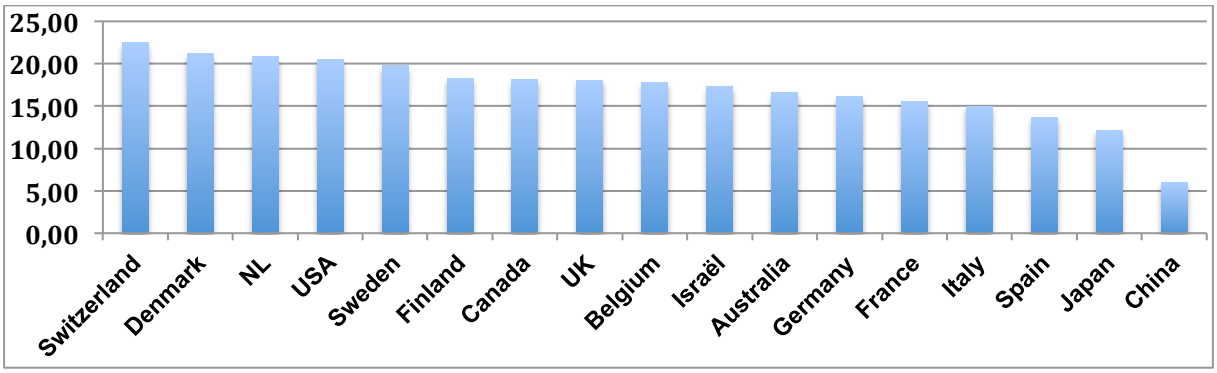

Figure 1: Number of citations per document for the period 1996-2011

the longer period. Figure 2 shows the country rankings, in terms of citation number per document, for the year 2011.

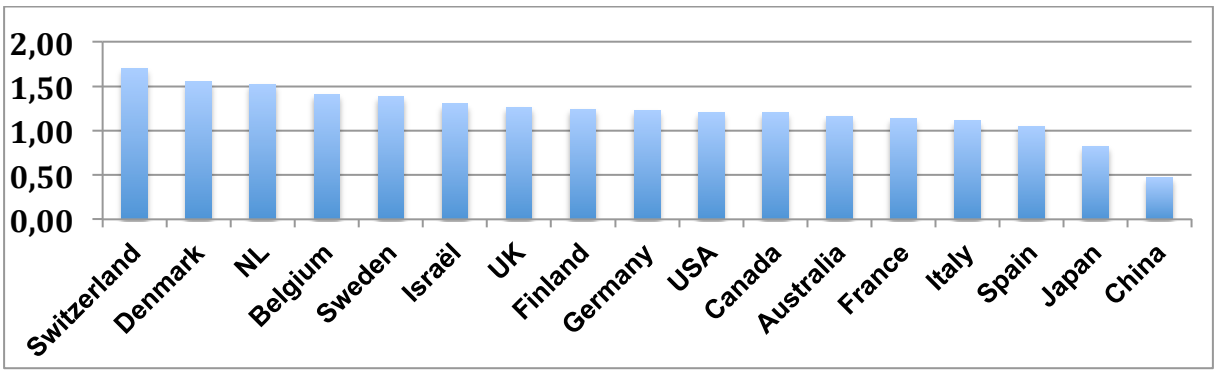

Figure 2: Number of citations per document for the year 2011

The ranking of the countries for the year 2010 is much the same as for 2011 (and remember that the variance is lower than for the ranking in 2011): Sweden is in 4-th position and Belgium in 5-th, while Israël is in 10-th position after Canada.

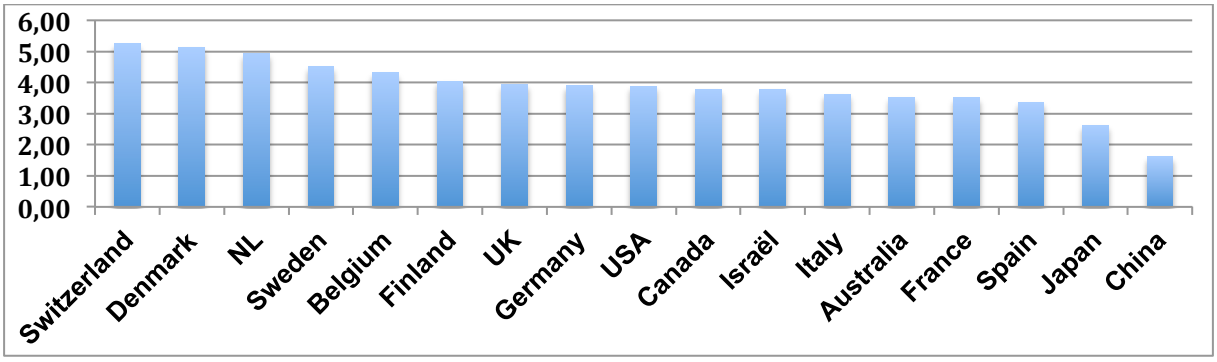

Figure 3: Number of citations per document for the year 2010

In comparing the recent data with those averaged over a 15-year period, it is remarkable to observe that the position of the top three countries has remained unchanged. In addition, Sweden is continuously in the top group. On the other hand, Belgium has moved significantly upward while the position of the USA has declined significantly.

Return on investment: citations versus budget for research

A second way of producing comparisons between countries of different sizes is to examine the return on investment. Given that citation counts are really a measure that reflects the scientific performance of fundamental research (industrial Research and Development is typically less conducive to publications), we have considered the investments of the 17 countries in Higher Education Research and Development (referred to as HERD by the OECD). In order to evaluate the return on investment, the number of citations produced in 2010 are divided, for each country, by the investment in HERD in 2008, thus allowing for some lag between investment and a measure of its return. The results appear in Figure 4.

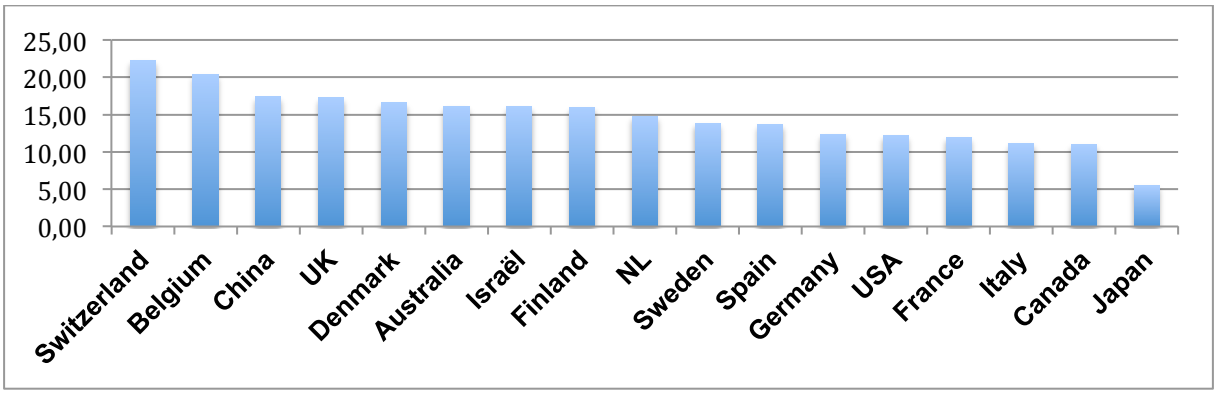

Figure 4: Citation numbers in 2011 versus investment in HERD in 2008

To check the robustness of this ranking we have computed the same graphs for the citations in 2010 versus HERD in 2008, as well as the citations in 2008 versus HERD in 2006; the relative positions remain essentially unchanged. In particular, Switzerland is always in top position and Belgium in second, while the UK is always within the top 4.

Comparing countries by Normalized Impact

The Normalized Impact (NI) has been defined above for the evaluation of a scientific institution. By adding the citation numbers for all institutions of a country, it can also be used to compare countries. Such comparison has been performed by Professor Félix de Moya Anegon, who has produced Figure 5, in which the NI has been computed for 50 countries, with the world average set at one [5]. Figure 5, produced in 2010, shows that the universities that achieve the highest NI are to be found in the UK and the USA, but the USA has by far the largest spread between high-performing and low-performing universities. Another remarkable feature is the very small spread of the NI between the “best” and the “worst” universities in Belgium, Norway, Singapore and New Zealand, with a median that is always well above the world average for these countries. Thus, a PhD student who decides to go to one of these countries knows that he/she will be in a good place, whatever university he/she chooses within those countries.

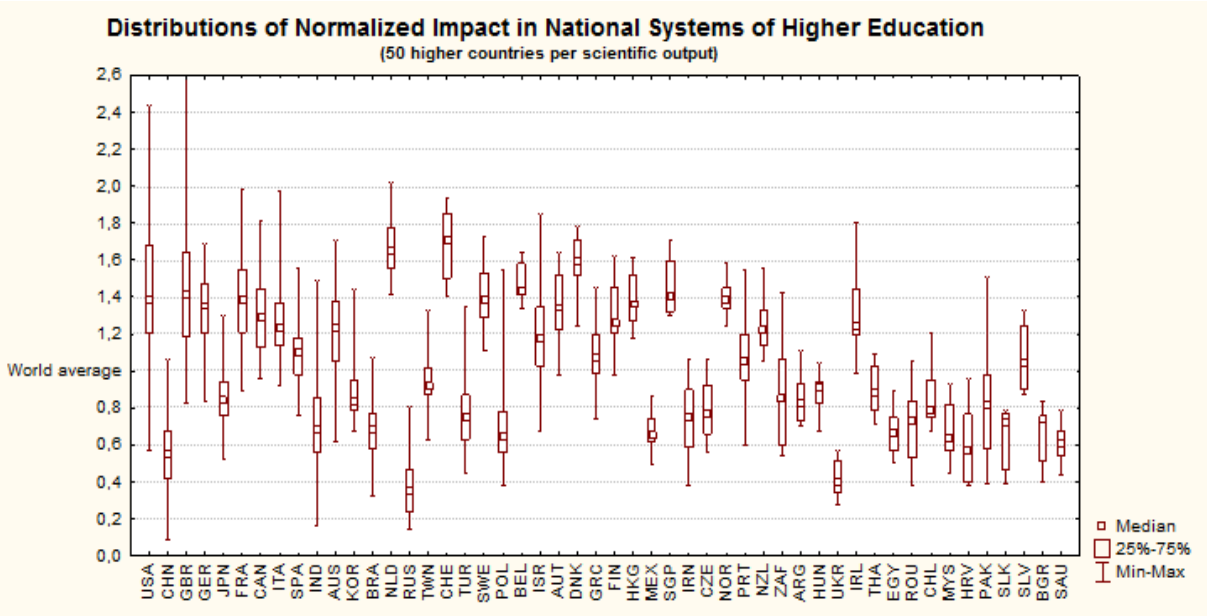

Figure 5: Normalized Impact of Higher Education Institutions in 50 countries

Conclusions

We have discussed the most commonly used research performance indicators at the four levels at which evaluations typically take place: the researcher, the journal, the institution and the country. We have argued that the introduction of every new indicator produces the introduction of feedback strategies whose aims are to increase the value of the indicator, sometimes at the expense of the quality of the research and of the research outputs. These negative effects have been illustrated with examples, at all four levels mentioned above. Nevertheless, indicators are useful provided they are well thought out and used with great care. This has led us to propose a number of quality criteria that performance indicators should possess.

In the second part of this chapter we have proposed an evaluation and comparison of the scientific performance of countries as measured by citation counts. In keeping with the quality criteria defined for performance indicators, we have used indicators that are adequately normalized; in addition our comparative analysis has been based not just on one, but on three normalized criteria. One should bear in mind that citation counts are but one of several indicators of research performance. We leave it to each reader to draw his/her own conclusions concerning the scientific performance of the countries that have been analysed.

References

1. Hirsch, J. E. (November 2005). "An index to quantify an individual's scientific research output". *PNAS* **102** (46): 16569–16572. arXiv:physics/0508025

2. http://www.che.de/downloads/Berlin_Principles_IREG_534.pdf

3. http://am.ascb.org/dora/

4. Gingras, Y. (2013), Criteria for Evaluating Indicators. In *Bibliometrics and Beyond : Metrics-Based Evaluation of Scholarly Research,* Cambridge, MA : MIT Press, to appear 2014.

5. SCImago (2007). SJR – SCImago Journal and Country Rank. Retrieved January 31, 2011 from http://www.scimagojr.com