Research evaluation: improvisation or science?

Giovanni Abramo*[1] and Ciriaco Andrea D'Angelo†[2]
*Laboratory for Studies of Research and Technology Transfer, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy, Italy
†Department of Engineering and Management, University of Rome 'Tor Vergata', Italy

Introduction

The proliferation of bibliometric indicators over recent years has generated a type of disorientation among decision-makers, who are no longer able to discriminate the pros and cons of the various indicators for planning an actual evaluation exercise. Performance-ranking lists at national and international levels are published with media fanfare, influencing opinion and practical choices. My impression is that these rankings of scientific performance, produced by 'non-bibliometricians' [1–3] and even by bibliometricians [4,5], are largely on the basis of what can easily be counted rather than 'what really counts'.

Performance, or the ability to perform, should be evaluated with respect to the specific goals and objectives to be achieved. The objectives must therefore be stated in measurable terms representing the desired outcome of production activity. The principal performance indicator of a production unit (whether this is an individual, research group, department, institution, field, discipline, region or country) is its productivity, or simply speaking the ratio of the value of the production output to the value of the inputs required to produce it. From this point of view, we will see that the renowned crown indicator and h-index (Hirsch-index), with its innumerable variants, and a wide variety of other publication-based and citation-based indicators, are inadequate to measure research productivity. As a consequence, all of the research evaluations based on these indicators and their relative rankings are at best of little or no value, and are otherwise actually dangerous due to the distortions embedded in the information provided to the decision-makers.

In the present chapter, we operationalize the economic concept of productivity for the specific context of research activity and propose a measurable form of productivity. From an economic perspective, we demonstrate the limits of the most commonly used performance indicators and present the indicator FSS (Fractional Scientific Strength), which better approximates the measure of research productivity. We present a methodology for measuring FSS at various levels of analysis. Finally, we will measure the performance of Italian universities by both FSS and the h-index, and will show the distortion inherent in the use of the h-index for assessing institutional productivity.

[1]Email: giovanni.abramo@uniroma2.it
[2]Email: dangelo@dii.uniroma2.it

Research productivity

Generally speaking, the objective of research activity is to produce new knowledge. Research activity is a production process in which the inputs consist of human, tangible (scientific instruments, materials, etc.) and intangible (accumulated knowledge, social networks, economic rents, etc.) resources, and where output, the new knowledge, has a complex character that is both tangible (publications, patents, conference presentations, databases, etc.) and intangible (tacit knowledge, consulting activity, etc.) in nature. The new-knowledge production function has therefore a multi-input and multi-output character. The principal efficiency indicator of any production is productivity, i.e. the ratio of the value of output produced in a given period to the value of production factors used to produce it. To calculate research productivity one needs to adopt a few simplifications and assumptions.

On the output side, a first approximation derives from the imposition of not being able to measure any new knowledge that is not codified. Secondly, where new knowledge is indeed codified, we are faced with the problem of identifying and measuring its various forms. It has been shown [6] that in the so-called hard sciences, the prevalent form of codification for research output is publication in scientific journals. Such databases as Scopus and WoS (Web of Science) have been extensively used and tested in bibliometric analyses, and are sufficiently transparent in terms of their content and coverage. As a proxy of total output in the hard sciences, we can thus simply consider publications indexed in either WoS or Scopus. With this proxy, those publications that are not censused will inevitably be ignored. This approximation is considered acceptable in the hard sciences, although not for the arts, humanities and a good part of the social science fields.

Research projects frequently involve a team of researchers, which shows in co-authorship of publications. Productivity measures then need to account for the fractional contributions of single units to outputs. The contributions of the individual co-authors to the achievement of the publication are not necessarily equal, and in some fields the authors signal the different contributions through their order in the byline. The conventions on the ordering of authors for scientific papers differ across fields [7,8], thus the fractional contribution of the individuals must be weighted accordingly. Following these lines of logic, all performance indicators based on full counting or 'straight' counting (where only the first author or the corresponding author receives full credit and all others receive none) are invalid measures of productivity. The same invalidity applies to all indicators based on equal fractional counting in fields where co-author order has recognized meaning.

Furthermore, because the intensity of publications varies across fields [9–11], in order to avoid distortions in productivity rankings [12], we must compare organizational units within the same field. A pre-requisite of any productivity assessment free of distortions is then a classification of each individual researcher in one and only one field. An immediate corollary is that the productivity of units that are heterogeneous regarding the fields of research of their staff cannot be directly measured at the aggregate level, and that there must be a two-step procedure: first, measuring the productivity of the individual researchers in their field, and then appropriately aggregating these data.

In bibliometrics, we have seen the evolution of language where the term 'productivity' measures refers to those based on publication counts, whereas 'impact' measures are those based on citation counts. In a microeconomic perspective, the first operational definition would actually make sense only if we then compare units that produce outputs of the same value. In reality, this does not occur because the publications embedding the new knowledge produced have different values. Their value is measured by their impact on scientific advancements. As a proxy of impact, bibliometricians adopt the number of citations for the units' publications, in spite of the limits of this indicator (negative citations, network citations, etc.) [13]. Citations do in fact demonstrate the dissemination of knowledge, creating conditions for knowledge spillover benefits. Citations thus represent a proxy measure of the value of output.

Comparing units' productivity by field is not enough to avoid distortions in rankings. In fact, citation behaviour too varies across fields, and it is not unlikely that researchers belonging to a particular scientific field may also publish outside of that field (a typical example is statisticians, who may apply statistics to medicine, physics, social sciences, etc.). For this reason, we need to standardize the citations of each publication with respect to a scaling factor stemming from the distribution of citations for all publications of the same year and the same subject category. Different scaling factors have been suggested and adopted to field-normalize citations (average, median, *z*-score of normalized distributions, etc.).

On the side of production factors, there are again difficulties in measuring that lead to inevitable approximations. The identification of production factors other than labour and the calculation of their value and share by fields is not always easy (consider quantifying the value of accumulated knowledge or scientific instruments shared among units). Furthermore, depending on the objectives of the assessment exercise, it could sometimes be useful to isolate and examine the contribution to output of factors that are independent of the capacities of the staff for the units under examination (for example, returns to scale, returns to scope, available capital, etc.).

The FSS indicator as a proxy of labour productivity

The productivity of the total production factors is therefore not easily measurable. There are two traditional approaches used by scholars to measure the total factor productivity: parametric and non-parametric techniques. Parametric methodologies are based on the *a priori* definition of the function that can most effectively represent the relationship between input and output of a particular production unit. The purpose of non-parametric methods, on the other hand, is to compare empirically measured performances of production units (commonly known as Decision Making Units or DMUs), in order to define an 'efficient' production frontier, comprising the most productive DMUs. The reconstruction of that frontier is useful to assess the inefficiency of the other DMUs, based on minimum distance from the frontier.

The measure of total factor productivity requires information on the different production factors by unit of analysis. Instead of total factor research productivity, most often research administrators are interested in measuring and

comparing simply labour productivity, i.e. the value of output per unit value of labour, all other production factors being equal. In measuring labour productivity then, if there are differences of production factors other than labour available to each unit, one should normalize for these. Unfortunately, relevant data are not easily available, especially at the individual level. Thus an often-necessary assumption is that the resources available to units within the same field are the same. A further assumption, again unless specific data are available, is that the hours devoted to research are more or less the same for each individual. Finally, the cost of labour is likely to vary among research staff, both within and between units. In a study of Italian universities, Abramo et al. [14] demonstrated that the productivity of full, associate and assistant professors is different. Because academic rank determines differentiation in salaries, if information on individual salaries is unavailable then one can still reduce the distortion in productivity measures by differentiating performance rankings by academic rank.

Next, we propose our best proxy for the measurement of the average yearly labour productivity at various unit levels (individual, field, discipline, department, entire organization, region and country). The indicator is named FSS, and we have previously applied it to the Italian higher education context, where most of its embedded approximations and assumptions are legitimate.

As noted above, for any productivity ranking concerning units that are non-homogeneous for their research fields, it is necessary to start from the measure of productivity of the individual researchers. Without it, any measure at aggregate level presents strong distortions [12]. In their measures of these data, the authors gain advantage from a characteristic that seems unique to the Italian higher education system, in which each professor is classified as belonging to a single research field. These formally defined fields are called 'Scientific Disciplinary Sectors' (SDSs): there are 370 SDSs, grouped into 14 UDAs (University Disciplinary Areas). In the hard sciences, there are 205 such fields grouped into nine UDAs.

When measuring research productivity, the specifications for the exercise must also include the publication period and the 'citation window' to be observed. The choice of the publication period has to address often-contrasting needs: ensuring the reliability of the results issuing from the evaluation, but also permitting frequent assessments.

Labour productivity at the individual level

At the micro-unit level (the individual researcher level, R) we measure FSS_R, a proxy of the average yearly productivity over a period of time, accounting for the cost of labour. In equation 1 [1]:

$$FSS_R = \frac{1}{s} \cdot \frac{1}{t} \sum_{i=1}^{N} \frac{c_{ii}}{\bar{c}_i} f_i \qquad (1)$$

where s is the average yearly salary of the researcher; t is the number of years of work of the researcher in the period of observation; N is the number of publications of the researcher in the period of observation; c_{ii} is the citations received by publication i; $\bar{c}_i$ is the average of the distribution of citations received

for all cited publications of the same year and subject category of publication *i*; and f_i is the fractional contribution of the researcher to publication *i*.

We normalize by the average of the distribution of citations received for all cited publications because it proved to be the most reliable scaling factor [15]. Fractional contribution equals the inverse of the number of authors, in those fields where the practice is to place the authors in simple alphabetical order, but assumes different weights in other cases. For the life sciences, widespread practice in Italy and abroad is for the authors to indicate the various contributions to the published research by the order of the names in the byline. For these areas, we give different weights to each co-author according to their order in the byline and the character of the co-authorship (intra-mural or extra-mural). If first and last authors belong to the same university, 40% of citations are attributed to each of them; the remaining 20% are divided among all other authors. If the first two and last two authors belong to different universities, 30% of citations are attributed to first and last authors; 15% of citations are attributed to second and last author but one; the remaining 10% are divided among all others.

Calculating productivity accounting for the cost of labour requires knowledge of the cost of each researcher, information that is usually unavailable for reasons of privacy. In the Italian case we have resorted to a proxy. In the Italian university system, salaries are established at the national level and fixed by academic rank and seniority. Thus all professors of the same academic rank and seniority receive the same salary, regardless of the university that employs them. The information on individual salaries is unavailable, but the salary ranges for rank and seniority are published. Thus we have approximated the salary for each individual as the average of their academic rank. If information on salary is not available at all, one should at least compare research performance of individuals of the same academic rank.

The productivity of each scientist is calculated in each SDS and expressed on a percentile scale of 0–100 (worst to best) for comparison with the performance of all Italian colleagues of the same SDS; or as the ratio to the average performance of all Italian colleagues of the same SDS with productivity above zero. In general, we can exclude, for the Italian case, that productivity ranking lists may be distorted by variable returns to scale, due to different sizes of universities [16] or by returns to scope of research fields [17].

Labour productivity at the organizational level

We have seen that the performance of the individual researchers in a unit can be expressed in percentile rank or standardized to the field average. Thus the productivity of multi-field units can be expressed by the simple average of the percentile ranks of the researchers. It should be noted that resorting to percentile rank for the performance measure in multi-field units or for simple comparison of performance for researchers in different fields is subject to obvious limitations, the first being compression of the performance differences between one position and the next. Thompson [18] warns that percentile ranks should not be added or averaged, because percentile is a numeral that does not represent equal-interval measurement. Furthermore, percentile rank is also sensitive to the size of the

fields and to the performance distribution. For example, consider a unit composed of two researchers in two different SDSs (A and B, each with a national total of ten researchers), who both rank in third place, but both with productivity only slightly below that of the first-ranked researchers in their respective SDSs: the average rank percentile for the unit will be 70. Then, consider another unit with two researchers belonging to another two SDSs (C and D, each with 100 researchers), where both of the individuals place third, but now with a greater gap to the top scientists of their SDSs (potentially much greater): their percentile rank will be 97. In this particular example, a comparison of the two units using percentile rank would certainly penalize the former unit.

However, the second approach, involving standardization of productivity by field average, takes account of the extent of difference between productivities of the individuals. In equation 2, the productivity over a certain period for department *D*, composed of researchers that belong to different SDSs [3]:

$$\mathrm{FSS}_D = \frac{1}{RS}\sum_{i=1}^{RS}\frac{\mathrm{FSS}_{R_i}}{\overline{\mathrm{FSS}}_{R_i}} \tag{2}$$

where *RS* is research staff of the department, in the observed period; FSS_{R_i} is the productivity of researcher *i* in the department; and $\overline{\mathrm{FSS}}_{R_i}$ is the average productivity of all productive researchers in the same SDS of researcher *i*.

Distortion inherent in the use of the h-index for assessing institutional productivity

In a recent study, Abramo et al. [19] examined the accuracy of the popular h- and g (Egghe)-indexes for measuring university research productivity by comparing the ranking lists derived from their application to the ranking list from FSS. In the present chapter, we report an extract of that study. For every SDS, we identify the Italian universities included in the first quartile of the ranking by FSS, then check

Table 1
Top universities by FSS that are not included in the same subset when performance is measured by h- and g-indexes

	Percentage of top 25% universities by FSS not included in the same set by:	
UDA	h-index	g-index
Mathematics and computer science	45	47
Physics	48	51
Chemistry	49	46
Earth sciences	42	35
Biology	42	36
Medicine	40	35
Agricultural and veterinary science	41	33
Civil engineering	28	26
Industrial and information engineering	40	35
Average	42	38

which of these would not be included in the same quartile under the rankings constructed with h and g values.

Table 1 presents the data aggregated by UDA. On average, the percentage of top 25% universities by FSS that are not included in the same set by h-index is 42%. Among the individual UDAs, the figures for these data vary between a minimum of 28% for the universities in civil engineering and a maximum of 49% for chemistry.

Concluding remarks

Until now, bibliometrics has proposed indicators and methods for measuring research performance that are largely inappropriate from a microeconomics perspective. The h-index and most of its variants, for example, inevitably ignore the impact of works with a number of citations below the h value and all citations above the h value of the h-core works. The h-index fails to field-normalize citations and to account for the number of co-authors and their order in the byline. Last but not least, owing to the different intensity of publications across fields, productivity rankings need to be carried out by field [20], when in reality there is a human tendency to compare h-indexes for researchers across different fields. Each one of the proposed h-variant indicators tackles one of the many drawbacks of the h-index while leaving the others unsolved, so none can be considered completely satisfactory.

The new crown indicator [MNCS (mean normalized citation score)], on the other hand, measures the average standardized citations of a set of publications. It is calculated as follows: one first calculates for each publication the ratio of its actual number of citations and the average number of citations of all publications of the same document type (i.e. article, letter or review) published in the same field and in the same year. One then takes the average of the ratios that one has obtained. The MNCS then cannot provide any indication of unit productivity. In fact, a unit with double the MNCS value of another unit could actually have half the productivity, if the second unit produced four times as many publications. Whatever the CWTS (Centre for Science and Technology Studies) research group [21] might claim for them, the annual world university rankings by MNCS are not 'performance' rankings, unless someone abnormally views performance as average impact of product, rather than impact per unit of cost. Applying the CWTS method, a unit that produces only one article with ten citations has better performance than a unit producing 100, where each but one of these gets ten citations and the last one gets nine citations. Furthermore, the methodology reported for producing the ranking lists does not describe any weighting for co-authorship on the basis of byline order. Similar drawbacks are embedded in the SCImago Institutions Ranking by their main indicator, the Normalized Impact (also known as Normalized Index), measuring the ratio between the average scientific impact of an institution and the world average impact of publications of the same time frame, document type and subject area. We do not consider further any of the many annual world institutional rankings that are severely size-dependent: the SJTU (Shanghai Jiao Tong University), *THES* (*Times Higher*

Education Supplement) and QS (Quacquarelli Symonds) rankings, among others. These seem to represent skilled communications and marketing operations, with the actual rankings resulting more from improvisation than scientifically reasoned indicators and methods.

The great majority of the bibliometric indicators and the rankings based on their use present two fundamental limitations: lack of normalization of the output value to the input value, and absence of classification of scientists by field of research. Without normalization, there cannot be any measure of productivity, which is the quintessential indicator of performance in any production unit; without providing the field classification of scientists, the rankings of multi-field research units will inevitably be distorted, due to the different intensity of publication across fields. An immediate corollary is that it is impossible to correctly compare productivity at international levels. In fact, there is no international standard for classification of scientists, and we are further unaware of any nations that classify their scientists by field at domestic level, apart from Italy. This obstacle can in part be overcome by indirectly classifying researchers according to the classification of their scientific production into WoS or Scopus categories, and then identifying the predominant category. FSS is a proxy indicator of productivity permitting measurement at different organizational levels. Both the indicator and the related methods can certainly be improved, but they do make sense according to the economic theory of production. Other indicators and related rankings, such as the simple number (or fractional counting) of publications per research unit, or the average normalized impact, cannot alone provide evaluation of performance; however, they could assume meaning if associated with a true measure of productivity. In fact, if a research unit achieves average levels of productivity this could result from average production and average impact, but also from high production and low impact, or the inverse. In this case, knowing the performance in terms of number of publications and average normalized impact would provide useful information on which aspect (quantity or impact) of scientific production to strengthen for betterment of production efficiency.

Aside from having an indicator of research unit productivity, the decision-maker could also find others useful, such as those informing on unproductive researchers, on top researchers (10%, 5%, 1%, etc.), top publications, dispersion of performance within and between research units, etc.

On the basis of the analyses above, we issue an appeal and recommendation. Our appeal to scholars is to concentrate their efforts on the formulation of productivity indicators more or less resembling the one we propose, and on the relative methods of measurement, aiming at truly robust and meaningful international comparisons. Our recommendation is to avoid producing research performance rankings by invalid indicators and methods, which under the best of circumstances serve no effective purpose, and when used to inform policy and administrative decisions can actually be dangerous. Our undertaking, as soon as possible, should be to develop a roadmap of actions that will achieve international performance rankings that are meaningful and useful to the research administrator and policymaker.

References

1. QS-Quacquarelli Symonds (2013) World University Rankings. http://www.topuniversities.com/university-rankings/world-university-rankings (Retrieved 26 July 2013)
2. SJTU, Shanghai Jiao Tong University (2013) Academic Ranking of World Universities. http://www.shanghairanking.com/ARWU2011.html (Retrieved 16 July 2013)
3. THES, Times Higher Education Supplement (2013) World Academic Ranking 2011–2012. http://www.timeshighereducation.co.uk/world-university-rankings/2011-2012/top-400.html (Retrieved 26 July 2013)
4. CWTS Leiden Ranking (2013) http://www.leidenranking.com/ranking (Retrieved 26 July 2013)
5. SCImago Journal and Country Rank (2013) Country Rankings. http://www.scimagojr.com/countryrank.php (Retrieved 26 July 2013)
6. Moed, H.F. (2005) *Citation Analysis in Research Evaluation*. Springer, Dordrecht
7. Pontille, D. (2004) *La Signature Scientifique: Une Sociologie Pragmatique de l'Attribution*. CNRS Éditions, Paris
8. RIN (Research Information Network) (2009) *Communicating Knowledge: How and Why Researchers Publish and Disseminate Their Findings*. http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/communicating-knowledge-how-and-why-researchers-pu (Retrieved 26 July 2013)
9. Garfield, E. (1979) Is citation analysis a legitimate evaluation tool? *Scientometrics* **1**, 359–375
10. Moed, H.F., Burger, W.J.M., Frankfort, J.G. and Van Raan, A.F.J. (1985) The application of bibliometric indicators: important field- and time-dependent factors to be considered. *Scientometrics* **8**, 177–203
11. Butler, L. (2007) Assessing university research: a plea for a balanced approach. *Science and Public Policy* **34**, 565–574
12. Abramo, G., D'Angelo, C.A. and Di Costa, F. (2008) Assessment of sectoral aggregation distortion in research productivity measurements. *Research Evaluation* **17**, 111–121
13. Glänzel, W. (2008) Seven myths in bibliometrics. About facts and fiction in quantitative science studies. Proceedings of WIS Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting (Kretschmer, H. and Havemann, F., eds), Humboldt-University Berlin, Germany, 29 July–1 August 2008
14. Abramo, G., D'Angelo, C.A. and Di Costa, F. (2011) Research productivity: are higher academic ranks more productive than lower ones? *Scientometrics* **88**, 915–928
15. Abramo, G., Cicero, T. and D'Angelo, C.A. (2012) Revisiting the scaling of citations for research assessment. *Journal of Informetrics* **6**, 470–479
16. Abramo, G., Cicero, T. and D'Angelo, C.A. (2012) Revisiting size effects in higher education research productivity. *Higher Education* **63**, 701–717
17. Abramo, G., D'Angelo, C.A. and Di Costa, F. (2013) Investigating returns to scope of research fields in universities. *Higher Education* doi:10.1007/s10734-013-9685-x
18. Thompson, B. (1993) GRE percentile ranks cannot be added or averaged: a position paper exploring the scaling characteristics of percentile ranks, and the ethical and legal culpabilities created by adding percentile ranks in making "high-stakes" admission decisions. Annual Meeting of the Mid-South Educational Research Association, New Orleans, U.S.A., 12 November 1993
19. Abramo, G., D'Angelo, C.A. and Viel, F. (2013) The suitability of h and g indexes for measuring the productivity of research institutions. *Scientometrics* doi: 10.1007/s11192-013-1026-4
20. Abramo, G. and D'Angelo, C.A. (2007) Measuring science: irresistible temptations, easy shortcuts and dangerous consequences. *Current Science* **93**, 762–766
21. Waltman, L., Calero-Medina, C., Kosten, J. et al. (2012) The Leiden ranking 2011/2012: data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology* **63**, 2419–2443