

Composite indicators of scientific research: the robustness of university rankings based on composite measures

Michela Gnaldi, Maria Giovanna Ranalli

Dipartimento di Economia Finanza e Statistica
Università di Perugia, Italy

Abstract

Composite indicators (CIs) which compare institution (e.g. schools, universities etc.) performances are useful tools in policy analysis and public communication. However, CI development involves stages where a number of subjective judgements have to be made, from the selection of individual indicators, to the choice of normalisation methods, weighting schemes, aggregation models etc. All these choices are potential sources of uncertainty which should be addressed when constructing a CI because they can have a significant impact on the composite index and rankings of the individual units considered. In this work - after a revision of the main university research evaluation methods and an illustration of the performance benchmarking system implemented in Italy - we address some of the methodological challenges implied in the construction of CIs and investigate the degree to which composite measures are an appropriate metric for evaluating and ranking the research performance of Italian universities.

Keywords: University research performance measurement, Composite indicators robustness.

Introduction

Over the past two decades university research evaluation exercises have become embedded in several systems. Evaluation methods have developed and each system underwent redesign, suggesting that the state-of-the-art is evolving. Depending on the method adopted and the level of complexity employed, assessment exercises yield to university ranking, department grading or research group grading. In Italy, university research assessment is performed by the Steering Committee for University Research Evaluation (CIVR). The outcomes of this first large-scale research evaluation process, based on the model of the British RAE, were published in 2006 by the CIVR. The Italian government used the results of the CIVR exercise to inform the distribution of part of the research portion of general university funds. In 2009, a ranking of Italian universities was issued based on a composite indicator built as a weighted combination of a number of variables and bibliometric indicators produced by the CIVR. In 2009, 7% (at least) of all support in Italian universities was allocated based on the composite indicator.

Composite indicators (CIs) integrate a large amount of information in a format that is easily understood and are, therefore, a valuable tool for conveying an overall assessment of performance in priority areas. However, the construction of composite measures creates specific methodological challenges. It is known that CI development involves stages where a number of subjective judgements have to be made, from the selection of individual indicators, to the choice of normalisation methods, weighting schemes, aggregation models etc. It is expected that the choice to include a particular indicator in the composite, or the choice to employ a normalisation scheme (rather than a different one) can have an impact on the rankings of the individual units (e.g. universities) within the composite. Thus, all these potential sources of uncertainty should be addressed when constructing a CI because they can have a significant effect on the composite index

and rankings of the units considered. Moreover, in a system - such as the Italian one - where universities are rewarded according to the outcome of a CI, great attention should be paid to the origin and nature of such sources of uncertainty because subtle changes to them can noticeably impact on the amount of financial support allocated to each unit.

In this work, we first propose a review of the main university research evaluation systems at international level (paragraph 1) and, within this international frame, illustrate the research assessment method and performance benchmarking system implemented in Italy (paragraph 2 and paragraph 3, respectively). Afterwards, in the light of the methodological challenges involved in the construction of composite measures (paragraph 4), we discuss some of the technical and analytical issues involved in the design of CIs and their policy implications (paragraph 5).

1. Evolving systems of university research evaluation at international level

University research evaluation exercises have become embedded in several systems and performance-based research funding has been increasing around the world. Evaluation methods have developed to include peer components as well as bibliometric measures and the exercises tended to become more and more elaborate.

In the United States, the research quality of university departments is publicly assessed by magazines and independent bodies. One of the most credited and qualified of these exercises is the one conducted by the National Research Council (NRC). The NRC was undertaken in 1983, 1995 and 2007–2008. Up to 2007, no bibliometric measures were used and university departments were ranked on the basis of a reputational survey. The 2007–2008 NRC method was more elaborate than the previous versions and was based on a ranking formula. Such formula was given by a CI constructed using a weighted mix of quantitative variables. A small reputational survey was conducted, and a regression analysis used to identify a weighted number of quantitative variables that best predicted reputational judgments. Departments were ranked based on this weighted mix of variables. The NRC thus required from all departments information on 48 variables to be included in the ranking formula. The 48 variables concern institutional characteristics (i.e. total research expenditure, characteristics of library, university housing for PhD students etc.); doctoral program characteristics (i.e. size, financial support, facilities for PhD students, support provided, employment destinations etc.), and program faculty (size, demographics, awards, bibliometrics etc.). For the bibliometrics, the NRC compiled full bibliographies of SCI indexed papers and their citations from Thomson-Scientific and used this information to calculate three bibliometric indexes: (1) % of faculty publishing, (2) publications/faculty, (3) citations/faculty.

In the United Kingdom, the Research Assessment Exercise (RAE) evolved over the years and grew increasingly complex. The exercise began with simple submission requirements in 1986. Departments described their research achievements in two pages, listed their five best publications and provided data on research income, prizes etc. This method was criticized (as for example favouring large departments) and, in response, it evolved so that submissions included greater detail on research strategy and listed four publications per individual and other data, and the original four point scale was elaborated to seven. In 2008, 68 panels of reviewers were convened to assess 180.000 departmental publications and to assign them grades. Thus, over the years, the exercise became increasingly complex and burdensome, and the UK government tried to reduce costs both reducing the exercise frequency (from 3 years to 4, 5 and 7 years) and incorporating in the 2008 exercise a “shadow metrics exercise” (the so-called RAE HERO), that is, a formula based on research grant funding and bibliometric indicators.

In Australia, from 1995 to 2004, the government evaluated the research in its universities using a composite index. The Composite Index was a formula at the university level (not at the departmental level as the RAE or NRC rankings). The formula was a ranking of universities (not an

assignment of grades like the RAE) and the components of the formula were research funding (grants from government, other public sector and industry) and outputs (number of publications and graduate degrees completed). In 2004, the formula based system was left behind after that some studies (Butler, 2003) provided evidence that the Composite Index did not reward research excellence. Thus, the Research Quality Framework (RQF), adopted in 2004, was a broader assessment of research quality. The RQF was RAE-like in that 13 subject area panels were convened to assess scientific quality of research products. However, differently from RAE, the panel assessment was based not only on research product quality, but also on a number of other submissions and metrics. The submissions were to comprise (among others) evidence of collaboration, awards won, students' employment destinations, grant income, the four best outputs for each researcher, a full list of outputs. The metrics were to report the citations per publication and the proportion of work that fell into the top citation percentiles in its field. After considering this information the panels were to assign each research group, within each department, scores on two five point scales representing research quality and research impact.

2. The university research evaluation system in Italy

The analysis of the different research evaluation systems identifies a tension – still far from being resolved – between complexity on one hand and practicality on the other hand. Increasing complexity is encouraged to assure the highest possible accuracy and fairness across heterogeneous academic disciplines (Hicks D., 2009) but raises questions about the cost/benefit ratio which, in turn, advocates a move to quantitative formulas (or CIs) in order to simplify the whole process.

Comparing the main research evaluation systems at international level allows to identify three research evaluation scenarios, characterized by increasing complexity:

1. The first, the simplest one, is the *formula-based* system, which uses as a unique evaluation instrument a CI, obtained as a weighted combination of a number of variables and/or bibliometric indicators;
2. the second is the *peer-review*, namely a system based on peer panels evaluation of the scientific quality of research products with no formal metrics components;
3. the third, the most complex one, is the *peer-informed metric based*, an evaluation process which relies not only on peer panels evaluation but also on bibliometric variables and indicators to assign grades of scientific quality.

Depending on the method adopted and the level of complexity employed, assessment exercises yield to university ranking, department grading or research group grading. In Italy, university research is evaluated through a peer-review system which produces a ranking of universities and can be therefore considered as a system of intermediate complexity. University research assessment is performed by the Steering Committee for University Research Evaluation (CIVR); the outcomes of this first large-scale research evaluation process, based on the model of the British RAE, were published in 2006. The aim of the CIVR exercise was an assessment of the scientific production of 77 Italian universities (in the period included between 2001 and 2003), 12 public research institutions and 13 private research institutions. The object of the exercise was an evaluation of research products intended in *latu sensu*: not only books, chapters, conference proceedings and scientific articles, but also patents, spin-offs, projects, design and drawings, performances, expositions and exhibitions, manufactured products and works of art. The exercise was performed by panels of experts and each product had been evaluated by at least 2 experts, in terms of their quality, relevance/importance, originality/innovation and international competitiveness. The expert

judgment (or score) on each research product had been unique and each product had been judged either excellent, or good, adequate or poor.

The data used in the exercise includes a number of individual indicators (e.g. number of publications, ordinary funds, impact factor of the journal in which each publication is included, number of spin-off created, etc.). Some of these individual indicators are evaluated with reference to university disciplinary areas, whereas others are aggregated by the CIVR at a university level (see for a full reference Table 1 and Table 2).

Table 1: CIVR Simple indicators - University Disciplinary Area level

<i>Variable name</i>	<i>Label</i>	
Total number of submitted research products	T	
i. (of which) Excellent	VE	
ii.(of which) Good	VB	
iii.(of which) Appropriate	VA	
iv. (of which) Poor	VL	
v. (of which) total number of evaluated research products	V	$V = VE + VB + VA + VL$
Percentage of Excellent products		$RE\% = 100 VE / V$
Weighted products	W	$W = VE + 0,8VB + 0,6VA + 0,2VL$
Product rating	R	$R = W / T$
Product quality	Q	$Q = R / V$
Number of authors (of research products) belonging to a given University/Total number of authors		
Number of authors (of research products evaluated Excellent) belonging to a given University/Total number of authors of research products evaluated Excellent		
Total research funds	F	
vi. (of which) from the Government	FM	
vii. (of which) from International Institutions	FE	
viii. (of which) from the Athaeneum itself	FS	
ix. (of which) from other sources	FA	
Ability to get funds		$A = 100(FM/FM_{tot} + FE/FE_{tot} + FA/FA_{tot})/3$
Number of professors, lecturers and research fellows moving to/from abroad	MI	
Number of Phd and postdoc students	NR	

3. The performance benchmarking system in Italy

The Italian government employed the results of the CIVR exercise to inform the distribution of part of the research portion of general university funds. In 2009, a ranking of Italian universities was issued based on a composite indicator built as a weighted combination of a number of variables and bibliometric indicators, some of which were derived from the CIVR dataset. In 2009, 7% (at least) of all support in Italian universities was allocated based on this composite indicator. Overall, 523.5 millions of Euros were allocated to Italian universities; 1/3 of the Government support was

distributed based on indicators related to education quality and 2/3 on account of indicators related to research quality.

Table 2: CIVR Simple indicators - University level

<i>Variable name</i>	<i>Label</i>	
Total number of professors, lecturers and research fellows	RT	
Total number of administrative employees	TAE	
Administrative employees over professors, lecturers and research fellows		TAE / RE
Expenditure (in 000 Euro)	I	
Earnings (in 000 Euro)	O	
Number of patents	B	
i. (of which) total number of evaluated patents	BV	
ii.(of which) Excellent	BE	
iii.(of which) Good	BB	
iv. (of which) Appropriate	BA	
v. (of which) Poor	BL	
Costs for patents	CB	
Number of international patents	BO	
Patent weighted index	BW	$BW = B + 1,5 BE$
Number of active patents	TB	
Number of Italian patents active abroad	TO	
On market patent weighted index	TBW	$TBW = TB + 1,5 TO$
Patents earnings	OB	
Number of spinoff	TS	
Number of partnerships	TP	
Economicity Index		$E = (OB-CB)/TB$
Valorizzazione applicativa Index		$AV = (BW+TBW+2*OB +4*TS+2TP) / 10$

As to the first quota (1/3) of public funds, this was based on 5 indicators of education quality:

1. For each study programme, average number of permanent teachers belonging to the basic subject areas which qualify the educational profiles offered by the programmes;
2. Ratio between number of enrolled students to the second academic year (a.a. 2008-09) who acquired at least 2/3 of the available credits and total number of enrolled students in the academic year 2007-08;
3. Ratio between number of credits acquired by students enrolled to the academic year 2007-08 and total number of available credits;
4. Ratio between number of study programmes which collected data on “Students’ opinions on study programmes” and total number of study programmes;
5. Percentage of graduated students in the (solar) year 2004 who were employed three years after graduation.

The second portion (2/3) of public funds was allocated on account of two weighted blocks of indicators and two further simple indicators, all related to research quality. The first group of indicators derived from the CIVR dataset included:

- Product quality (weight 4/9);

- Excellent product quality (weight 2/9);
- Number of professors, lecturers and research fellows moving to/from abroad (weight 1/9;)
- Number of Phd and Post-doc students (weight 1/18);
- Total research funds from Government, International Institutions and other sources (weight 1/9);
- Total research funds from the athenaeum itself (weight 1/18).

The second group of indicators derived from the CIVR dataset included:

- Number of patents submitted in 2001-2003 (with weight 1/10 for Italian patents and 1.5/10 for international patents);
- Number of patents activated in 2001-2003 (with weight 1/10 for Italian patents and 1.5/10 for international patents);
- Patents earnings (weight 2/10);
- Number of spin-off (weight 4/10);
- Number of partnerships (weight 2/10).

The third indicator included in the formula was calculated as a ratio between number of permanent teachers taking part in “Projects of National Interest” (PRIN) and the total number of permanent teachers. The ratio was expressed as a weighted percentage, for each scientific area, based on the percentage ratio between number of PRIN “won” by the athenaeum and total number of PRIN at the national level. The fourth indicator was obtained as a mean average, for each athenaeum, of the percentage of EU projects “won” by the athenaeum (out of the total number of EU projects funded at the national level), and the percentage of EU funds received by the athenaeum (out of the total EU funds at the national level). The weights applied to the two blocks of indicators and the other two indicators were, 0.49, 0.01, 0.15, 0.35, respectively.

4. Methodological challenges in the construction of composite measures

In the development of CIs it is necessary to follow a sequence of steps, from the definition of a theoretical framework to data selection, imputation of missing data, multivariate analysis, choice of normalisation, weighting and aggregation schemes, and sensitivity analysis. A framework is the preliminary step in constructing composite indicators because it should define the phenomenon to be measured, selecting individual indicators and weights that reflect their relative importance and the dimensions of the overall composite. This process should ideally be based on what is desirable to measure and not on which indicators are available (OECD, 2008). Data selection is a second fundamental step in the process of construction of a CI since the strengths and weaknesses of composite indicators largely derive from the quality of the underlying variables. Preferably, variables should be selected on the basis of their relevance, analytical soundness, timeliness, accessibility, *etc.* Detection of missing data is also crucial to assess the impact of the imputation on the composite indicator results, to spot the presence of outliers in the dataset and thus to make it possible the development of robust composite indicators. Multivariate techniques (Factor Analysis, Principal Component Analysis *etc.*) are needed to analyse the underlying nature of the data, the existence of relationships among individual indicators and, therefore, to assess the suitability of the whole data set. Normalisation is required prior to any data aggregation as the indicators in a data set often have different measurement units. Thus, normalisation is needed to allow scale adjustments and to transform highly skewed indicators. The definition of a weighting scheme is also a major issue in constructing a CI as weights are essentially value judgements which can significantly affect the overall composite indicator and rankings. A number of weighting techniques exist and weights can be chosen based on statistical methods and/or expert opinions. Aggregation methods also vary.

It is known (OECD, 2008) that while the linear aggregation method is useful when all individual indicators have the same measurement unit, geometric aggregations are better suited if some degree of non-compensability between individual indicators or dimensions is required (e.g. a deficit in one dimension can be compensated by a surplus in another).

Furthermore, it is agreed that any composite indicator may be considered as a model (OECD, 2008) where the CI is the response variable and the covariates are all the subjective judgements (the sources of uncertainties) which have to be made (e.g. the selection of individual indicators, the choice of normalisation methods, weighting schemes, aggregation model *etc.*). It is argued (Saisana *at al.*, 2005; Munda *at al.*, 2005, Saltelli, 2007) that all these potential sources of uncertainty should be addressed because they affect both the variance of the CIs and the variability of any ranking based on CIs. Sensitivity analysis can be considered as an appropriate tool to assess such uncertainties as it studies how the variation in the output (e.g. CIs and rankings) can be apportioned to different sources of variation in the assumptions, and how the given composite indicator depends upon the information fed into it. Given that its primary aim is to quantify the overall uncertainty in institution rankings as a result of the uncertainties in the model input, it can help to gauge the robustness of the composite indicator ranking and to identify which institutions are favoured or weakened under certain assumptions.

A contribution to the understanding of the methodological issues involved in the construction of composite measures for the benchmarking of universities has been proposed in Gnaldi and Ranalli (2009). In that work, 135 CIs of scientific research for the Italian universities were constructed on a selection of individual indicators put forward by the CIVR and combining (i) five normalisation methods, (ii) a weighting scheme and (iii) two aggregation schemes. The variation in the rankings assigned by the resulting 135 CIs to the given universities was then explored to gauge the robustness of the composite indicator rankings and to assess the contribution of the individual sources of uncertainty to the output variance. The authors showed that the university ranking distribution based on CIs is associated with a certain level of variability, which is very strong for some Universities. In particular, it was illustrated that the groups of university laggards is less sensitive to variations, while the central ranking positions are occupied by the most sensible universities, with a variation in their ranking of around 30 positions (out of 61 positions). It was also highlighted that a very important source of uncertainty is the normalisation scheme, together with the choice to include/exclude a simple indicator from the CIs calculus. A regression analysis run per university also revealed that the same single indicator can be, at the same time, a strong source of strength and weakness for different universities, determining a considerable positive or negative variation in their ranking positions.

5. Conclusions

Since early '90s, national university ranking exercises have developed in several Countries, undergoing a deep redesign denoted by a tension between complexity and practicality. Increasing complexity is encouraged to meet accuracy and fairness but raises issues on the costs of assessment exercises which, in turn, advocate a shift to simplification and, thus, to the use of composite measures (CIs). The need to restrain systems that are becoming increasingly elaborate, and even unwieldy, and to control for costs has persuaded some Countries, including Italy, to employ CIs for the benchmarking of universities. However, the construction of composite measures raises a number methodological issues. It has been underlined that any composite indicator may be considered as a model and that all potential sources of uncertainty (e.g. the selection of individual indicators, the choice of normalisation methods *etc.*) should be addressed because they affect the variance of the CIs. Since any university ranking distribution based on CIs is associated with a certain level of variability, dealing with uncertainty in composite performance measures is essential. Any composite indicator of scientific performance should be published with indications of uncertainty to

communicate the sensitivity of the reported measures. In any performance benchmarking system, such as the Italian one, it is crucial to know an estimate of the degree of variation for each indicator so that definitive conclusions can be drawn about genuine differences in performances among universities. Besides, it has been noted (Gnaldi and Ranalli, 2009) that the choice of a normalisation and aggregation scheme has a significant impact on the rankings of universities within the composite. As a consequence, great attention should be paid to the origin and nature of such sources of uncertainty because subtle changes to them can noticeably impact on the composite index and rankings of universities. As to the weighting scheme, when it is computed by means of Factor Analysis - as in the contribution proposed by Gnaldi and Ranalli (2009) - it gives weights that intervene to correct for overlapping information between two or more correlated indicators, and not as measures of the theoretical importance of the associated indicator. However, as in common practice, a larger weight could be given to components which are considered more significant in the context of a particular composite indicator. Other weights - such as those employed within the Italian university funding formula and based on experts' judgments - can be accounted for to assess the impact of their use on university rankings.

It has been also highlighted that the same single indicator can be, at the same time, a strong source of strength and weakness for different universities, determining a considerable positive or negative variation in their ranking positions. Hence, the choice to include a single indicator in the composite is, again, a crucial phase in the construction of a composite measure and a political choice which strongly affects the composite and the ranking. In a system where universities are rewarded according to the outcome of the composite indicator, these decision rules need to be treated with caution and to be publicly defined.

Acknowledgments: Work supported by the project PRIN 2007 *Models and methods for the evaluation of the social and professional return of the competences formed in a university study programme and analysis of the Italian study programmes in statistics and medicine.*

References

Butler, L. (2003), Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy*, 32, 143–155.

Fabbris L. and Gnaldi M. (2008), Indicatori di valutazione della qualità della ricerca negli atenei: sensibilità, sostituibilità e capacità discriminatoria, in *Professionalità nei servizi innovativi per studenti universitari*, (Eds: Fabbris L., Boccuzzo G. and Martini M.C.), 139-171, CLEUP, Padova.

Gnaldi M. and Ranalli M.G. (2009), Composite indicators of scientific research, *Quaderni di Statistica*, Liguori Editore, 11, 165-181.

Hicks D. (2009), Evolving regimes of multi-university research evaluation, *Higher Education*, 57, 393-404.

Jacobs R., Smith P. and Goddard M. (2004), Measuring performance: an examination of composite performance indicators, *Technical Paper Series 29*, Centre for Health Economics, University of York.

Jacobs R., Goddard M. and Smith P.C. (2006), Public Services: Are Composite Measures a Robust Reflection of Performance in the Public Sector, *CHE Research Paper 16*, Centre for Health Economics, University of York.

Melyn W. and Moesen W.W. (1991), Towards a synthetic indicator of macroeconomic performance: unequal weighting when limited information is available, *Public Economic research Paper 17*, CES, KU Leuven.

Moesen L., Rogge W., Van Puyenbroeck N.T., Saisana M., Saltelli A., Liska R., Tarantola S. (2008), Creating composite indicators with DEA and robustness analysis: the case of the Technology Achievement Index, *Journal of the Operational Research Society*, 59, 239-251.

Munda G. and Nardo M. (2005), Constructing Consistent Composite Indicators: the Issue of Weights, EUR 21834 EN, *Joint Research Centre*, Ispra.

OECD (2008), *Handbook on Constructing CIs – Methodology and user guide*, available at: <http://composite-indicators.jrc.ec.europa.eu/Handook.htm>

Saisana M., Tarantola S. and Saltelli A. (2005), Uncertainty and sensitivity techniques as tools for the analysis and validation of CIs, *Journal of the Royal Statistical Society Series A*, 168, 307-323.

Saltelli A. (2007), Composite indicators between analysis and advocacy, *Social Indicators Research*, 81, 65-77.

Saaty R.W. (1987), The analytic hierarchy process: what it is and how it is used, *Mathematical Modelling*, 9, 161-176.