Journal of Informetrics 10 (2016) 224–237

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Nondeterministic ranking of university departments[☆]

Andrea Bonaccorsi[a], Tindaro Cicero[b,*]

[a] DESTEC, School of Engineering University of Pisa Largo, Lucio Lazzarino 2, 56125 Pisa, Italy
[b] ANVUR Italian Agency for the Evaluation of Universities and Research Institutes, Via Ippolito Nievo 35, 00153 Rome, Italy

ARTICLE INFO

Article history:
Received 3 October 2014
Received in revised form 8 January 2016
Accepted 10 January 2016
Available online 30 January 2016

Keywords:
Departments ranking
Ranking
Research quality
National research assessment

ABSTRACT

Rankings in higher education are largely used to summarize a huge amount of information into easily understandable numbers. They are also used by governments in order to allocate funding. Nevertheless, they are often criticized. One stream of criticism refers to the fact that rankings build up an ordinal order by considering only the mean of the distribution of indicators and not their variability. Using the micro-data from the Italian evaluation of the quality of research (VQR, Valutazione della Qualità della Ricerca), we examine whether difference in performance between departments with different position in the ranking are distinguishable from random effects. We obtain a robust clustering of departments in a limited number of groups. The number of groups is in the range 3–7, while in most cases it is 4–6. The implications of these findings for evaluation and research policy are explored.

1. Introduction

Rankings of universities or university departments are largely used by the media, public opinion, and governments in order to examine the position of individual institutions or raise issues about the national performance. They are considered useful because they summarize a huge amount of information into easily understandable numbers (Hazelkorn, 2011; O'Connell, 2013). Rankings are however heavily criticized by scholars in social sciences, ranging from statistics to sociology and political sciences.

One stream of criticism refers to the fact that rankings build up an ordinal order by considering only the *mean* of the distribution of indicators, disregarding other moments of the distribution. Whatever the metrics adopted, each university or department receives a score that is the result of the activity of all its members, or at least its active part. However, the distribution of results of all members is entirely lost and only a single score enters into the composite indicator, hence the final ranking.

We follow this line of criticism by examining the relation between individual scores of quality of research and average scores assigned to departments. We make use of micro-data from the largest evaluation exercise ever carried out, i.e. the evaluation of the quality of research (in the rest of the paper: VQR, Valutazione della Qualità della Ricerca), completed in July 2013 by the Italian Agency for the Evaluation of Universities and Research Institutes (ANVUR).

VQR has produced a ranking of departments and a ranking of universities, following a mandatory provision of the national Italian legislation. At the same time, it has published all data aggregated at the level of departments and scientific areas,

[☆] *Disclaimer*: This paper was prepared or accomplished by Tindaro Cicero and Andrea Bonaccorsi in their personal capacity. The opinions expressed in this article are the authors' own and do not reflect the view of the ANVUR.

[*] Corresponding author. Tel.: +39 06 58317321.
E-mail address: Tindaro.cicero@anvur.it (T. Cicero).

http://dx.doi.org/10.1016/j.joi.2016.01.007
1751-1577/© 2016 Elsevier Ltd. All rights reserved.

allowing actors in the research system to reconstruct possibly different rankings, for example by changing the criteria or weights of aggregation. It has recognized that, while the raw indicators associated to individual publications should be considered reliable, rankings of departments depend crucially on the definitions of the perimeter of comparison and on the treatment of the heterogeneity within the department. We bring this recognition a step further, by using anonymized micro-data on *individual* publications of *individual* researchers, and examining whether departments rankings are statistically robust.

The paper is organized as follows. Section 2 introduces the VQR and explains the source of data. Section 3 review the current debate on university rankings. Section 4 introduces the statistical approach to rankings, as discussed by Lubrano (2009). Section 5 discusses the potential and limitations of the method suggested by Lubrano with respect to the Italian data and applies the statistical analysis of rankings. Section 6 discusses the results while Section 7 concludes.

2. The ranking of departments in the Italian evaluation of quality of research (VQR)

The VQR is the largest research assessment exercise ever carried out. The results of this exercise have been published (in Italian) in July 2013 and are currently under translation into English for a wider circulation. In 2015 a new exercise, covering the period 2011–2014, has been launched, confirming that research assessment is a permanent feature of the institutional landscape.

Following a Ministry of Research decree in 2011, VQR has been managed by the newly created Agency for the Evaluation of Universities and Research Institutes (ANVUR), an independent public institution in charge of research evaluation and higher education quality assurance. VQR has covered the period 2004–2010, involving all researchers of all Italian universities (including private universities and distance universities) and Public Research Organizations (PROs). Other research organizations not subject to evaluation by law have also applied for voluntary evaluation.

Researchers at universities have been asked to submit three research products; researchers at PROs have submitted six products. Research products include publications (articles, books, chapters, proceedings, critical editions, translations and commentaries) but also patents, design, drawings, performances, exhibitions, artifacts, prototypes, databases, software, and maps. 95 universities, 12 PROs and 26 voluntary research consortia and other research organizations have been submitted to evaluation.

VQR has followed a common set of rules that have been dictated by the Ministry of Research. Research products should be evaluated in terms of their relevance (value added for the advancement of knowledge and of science in general), originality/innovativeness (contribution to the state of the art in the field) and internationalization. While the metrics for evaluation could differ across disciplines, all products should be eventually rated into four merit classes (A, B, C and D), which are defined in terms of the position in the overall distribution of merit, as follows: A (Excellent): more than 80%, B (Good): 60–80%, C (Fair): 50–60%, D (Limited): less than 50%. For each class the Ministerial decree introduce a score, as follows: 1 (Excellent), 0.80 (Good), 0.50 (Fair), and Zero (Limited). The scores can also be negative: −1 for not admissible products, −2 for plagiarism, −0.5 for each product missing with respect to the expected number of products (three for university researchers, six for PROs, respectively). Consequently, university researchers have a maximum score of 3, if all their products are considered excellent.

ANVUR nominated in 2011 a panel of 450 experts, divided in 14 disciplinary areas[1]. Expert panels have debated and published evaluation criteria, following a mix of bibliometric approaches and informed peer review. Bibliometric criteria have been based on a combination between number of citations (citations received by an article from the date of publication until end 2010, expressed as quantile of the world distribution of citations for articles in the same journal subject category) and journal impact score (using various measures available from JCR and/or Scopus).

All disciplinary panels in science, technology and medicine, including psychology, opted for bibliometric methods as the dominant methodology. However, in all disciplines a sample of papers have been evaluated with both bibliometric and peer review methods, in order to compare the results.

On the contrary, disciplinary panels in architecture, humanities and social sciences have adopted peer review as the main approach, in most cases at 100%. In the case of peer review, each research product has been submitted to two external referees, chosen independently and without mutual information by members of the panel. Scores of referees have then been reconciled within consensus groups; in case of severe disagreement a third referee has been recruited. Overall, more than 14,000 external referees have been mobilized.

Scores of research products have then been aggregated at the level of disciplines and departments. These data have been published at the lowest possible level of aggregation, which in most cases goes down to more than 300 fine-grained disciplines (SSD, Settori Scientifico Disciplinari). Only when the number of researchers in a university was smaller than four, data have been aggregated at a higher level, in order to preserve statistical confidentiality. Individual scores have been communicated via mail to all researchers but have neither been disclosed, nor transfered internally to the Ministry or anyone else, but kept strictly confidential.

[1] The Italian Research System is composed by 370 SDSs, grouped in 14 CUN (Italian National University) areas; every researcher belongs to one and only one SDS. See http://www.cun.it/media/116411/settori_scientifico_disciplinari_english.pdf (accessed November 20, 2015).

Average scores differ across disciplines, either because of intrinsic differences in quality and because of the evaluation mix (bibliometric methods invariably give higher scores than peer review). In order to control for differences in the subject mix, data have been aggregated at the level of departments and universities by normalizing the scores of researchers within the respective disciplines. Therefore the ranking of a department depends on the average score of the affiliated researchers, normalized against the national average for the discipline to which they belong.

The production of rankings was mandated by the Ministerial decree. In fact, data aggregated at the level of universities have been used at the end of 2013 and 2014 by the Ministry of Research to distribute the performance-based share of funding, called "quota premiale", a share that increases stepwise up to a target of 30% of the total. Because of the significant financial impact, the notion of university and department ranking has received enormous attention in Italy, a country in which, on the contrary, there had been no tradition in university rankings. Consequently, the methodological foundations of rankings have also been scrutinized carefully in the academic and policy debate following the publication of the results of VQR.

3. The methodological debate on university rankings and the nondeterministic approach

The literature that has addressed university rankings with a critical approach is quite large and has been examined elsewhere with the aim of introducing a new methodology based on conditional ranking (Daraio & Bonaccorsi, 2014; Daraio, Bonaccorsi, & Simar, 2015). In this paper we take a narrow focus on the issue of deterministic assumptions underlying the construction of composite indicators.

Although the practice of university rankings dates back (Hattendorf, 1986), it is in the last decade that a large number of review articles and critical essays have examined the methodology of university rankings and its effects on higher education at large (Bowden, 2000; Provan & Abercromby, 2000; Brooks, 2005; Dill & Soo, 2005; Turner, 2005; van Dyke, 2005; Usher & Savino, 2006; Buela-Casal, Gutiérrez-Martínez, Bermúdez-Sánchez, & Vadillo-Muñoz, 2007; Ioannidis et al., 2007; Taylor & Braddock, 2007; Hazelkorn, 2007; Hazelkorn, 2009; Harvey, 2008). More recently, important methodological improvements have been proposed (Aguillo, Ilan, Levene, & Ortega, 2010; Chen & Liao, 2012; Safon, 2013; Bornmann, Mutz, & Daniel, 2013; Cantwell & Taylor, 2013; Freyer, 2014). A few dedicated books (Kehm & Stensaker, 2009; Salmi, 2009; Dehon, Jacobs, & Vermandele, 2009; Shin, Toutkoushian, & Teichler, 2011; Hazelkorn, 2011; Erkkilä, 2013) summarize the state of the art in the academic literature.

The pervasiveness and importance of the topics is witnessed by the fact that it is the object not only of specialized academic articles, but also of policy documents. Policy publications and official reports have been produced directly or under the auspices of UNESCO (Unesco, 2013), World Bank (Salmi, 2009) or the OECD (Salmi & Saroyan, 2007), as well as by large representative associations such as the Association of Universities in Europe (Estermann & Nokkala, 2009; Estermann, Nokkala, & Steinel, 2010; Rauhvargers, 2011; European University Association, 2012) and the League of European Research Universities (Boulton, 2011), and by specialized policy research centres (Institute for Higher Education Policy, 2007, 2009).

Within this literature, a small subset of authors have addressed technical issues of the statistical assumptions underlying the construction of composite indicators. There are two main lines of investigation here.

Saisana and D'Hombres (2008) and Saisana, D'Hombres, and Saltelli (2011) have developed a methodology to test the robustness of rankings. Being based on elementary indicators aggregated into composite indicators, rankings utilize only one of a number of possible combinations of indicators and of aggregation rules. One problem, often raised in the literature, is that the weights used for the aggregation of individual indicators are arbitrary and lack theoretical foundation (Provan & Abercromby, 2000; Brooks, 2005; Lukman, Krajinc, & Glavic, 2010). Using a simulation technique, Saisana and co-authors show that, in general, rankings are robust in the top positions but less reliable elsewhere, that Shanghai rankings are more robust than Times Higher Education Supplement rankings, and that for a certain numbers of universities the variability induced by changes in the construction of the composite indicator is so large that all existing rankings are de facto meaningless. This line of investigation addresses the statistical issues underlying the construction of composite indicators, taking for granted the elementary indicators that are aggregated in order to come up with a single ranking.

Yet another line of investigation questions the statistical properties of elementary indicators. According to Lubrano (2009) an important methodological problem of rankings is that they assume a deterministic setting, in the sense that they ignore the underlying process of generation of indicators that are then aggregated. In particular, underlying indicators are average values from distributions. Each researcher is attributed a score for his publications, according to agreed bibliometric criteria, such as the normalized number of citations, or the cumulated impact factor of the journals. Having said that, the crucial point is to understand that the score of an author is not a deterministic but a random variable.

As he puts it:

> The score of an author is a random variable for which we have observations. The randomness can be attributed to several facts that are essential to understand. There is a variable time between submission and publication. The probability of acceptance of a paper is influenced by the choice of the names of the referees. The choice of the journal to which the paper is submitted is not necessarily optimal. The contribution of each author to an article is not necessarily the same. The length of a paper is not always strictly related to its quality and impacts. Some of the notes published in *Econometrica* receive much more citations that regular articles. The production of an author might vary a lot across the year either because of cyclical productivity or because the author is at the beginning or at the end of his professional life cycle (Lubrano, 2009, 81).

This point is crucial for the overall discussion, because any aggregation of a random variable is itself a random variable. As the author puts it:

> If scientific production is a random variable, then the measure used for ranking departments and universities is also a random variable. Consequently, rankings cannot be taken at face value. (...) Random variables have standard deviations. Whatever the criterion used for rankings institutions, a standard deviation has to be taken into account and this means that rankings are not deterministic. Two or more institutions can be statistically indistinguishable. The second point is that a standard deviation can be diminished by increasing the number of observations. In order to rank departments properly, we have to use criteria for which there are many observations. If the criterion used is too elitist, for example counting the number of Nobel Prizes, we simply are not going to have enough observations. The Shanghai ranking is subject to this type of criticism (Lubrano, 2009, 85; 98).

In this perspective, the relative position of any university with respect to others cannot be assumed as "objective", because the difference with universities above or below might be statistically indistinguishable from zero. This criticism is not new in the theoretical literature on rankings. As Goldstein and Spiegehalter (1996) have reminded, "the mean has no special status" (Goldstein & Spiegehalter, 1996, p. 395). In other words, rankings suppress the intrinsic variability of indicators at lower levels of aggregation, giving an impression of stable hierarchies among universities, without explicitly testing for the statistical representativeness of differences. Consequently "an overinterpretation of a set of rankings where there are large uncertainty intervals, can lead both to unfairness and to inefficiency and unwarranted conclusions about changes in ranks" (Goldstein & Spiegehalter, 1996, p. 405).

In order to adopt a statistical approach to rankings, however, we first have to address two methodological problems. The first, which we examine in the next section, is whether we can safely assume normality in the distribution of elementary indicators. The second is whether the construction of rankings associated to simple indicators of variability (such as confidence interval, or standard deviation) might not be sufficient to address the problem.

4. The nondeterministic approach to university rankings: Can we assume normality in the distributions?

The notion that rankings should be based on a statistical definition has been applied to departments in Lubrano, Kirman, Bauwens, and Protopopescu (2003a), Lubrano, Bauwens, Kirman, and Protopopescu (2003b), Lubrano and Protopopescu (2004) and is crucial in the recent reformulation of rankings as stochastic dominance relations (Carayol, Filliatreau, & Lahatte, 2013; Carayol & Lahatte, 2014).

It is well known that university rankings generate a huge attention, because readers interpret positional information as representing differences in the underlying quality. To the readers of rankings, the information that university A is better than university B because A is ranked 12 and B ranked 15 is not cast in doubt. Because they are expressed in crude numbers, ranks convey unquestionable truth to the users.

The nondeterministic approach suggests that differences between any pair of universities should be subject to statistical testing. To put it simply: "Are two departments which are ranked differently, really that different? A statistical test for equality of the scores can be easily devised" (Lubrano, 2009, 85).

The score of an university department is the average of the scores of the authors affiliated to it. As suggested in Lubrano et al. (2003a, 2003b) we could test statistical differences between the scores of two departments (named A and B). If X_{Ai} is a random variable denoting the score of researcher i in department A and n_A denotes the number of researchers affiliated to it, we assume that different occurrences of X_{Ai} are mutually independent in probability and identically distributed according to a distribution with mean μ_A and variance $\sigma^2{}_A$. We make the same assumptions for department B and we also assume that the score of an individual in A is independent of the score of any individual in B. This last assumption is a realistic one, since we rely on data derived from VQR 2004–2010 (as described in the previous section) and rules governing this research assessment stipulated that every research paper co-authored by two or more authors of the same department could be submitted only once. Collaborations between researchers of different departments are more unusual, so we assume that the relevant correlation is negligible. From a practical point of view, all researchers affiliated to the same department have the same probability of getting the top score; certainly, several factors can influence the performance of each researcher (including, but not limited to, his/her age, academic rank, gender etc.), yet these factors only account for a small part of entire variability of performance (Cicero, Malgarini, & Benedetto, 2014).

In principle there is no reason to expect that individual assessment scores are normal, because according to the adopted definition in VQR, each research outcome was evaluated to be excellent (score = 1), good (score = 0.8), acceptable (score = 0.5), limited (score = 0), not evaluable (−1) and cases of plagiarism of fraud (−2). The distributions of the researcher scores are often skewed and they are far from normality. It is interesting to observe that, while the distributions of individual productivity are well known to be skewed on the right, the distributions of mean scores are unimodal and skewed to the left. This comes from the fact that researchers submit their *best* products and not a random sample of their production. In order to apply the Central Limit Theorem, we excluded from the analysis all departments with less than 10 researchers in the scientific areas, or 30 research products. For remaining departments, we verified that the Theorem applies, so that the score of a department can be approximated by a normal distribution in large samples. As example in Fig. 1 we represent the researchers distributions in the largest (n = 99) and the smallest department (n = 10) in mathematics and computer sciences. We observe that the

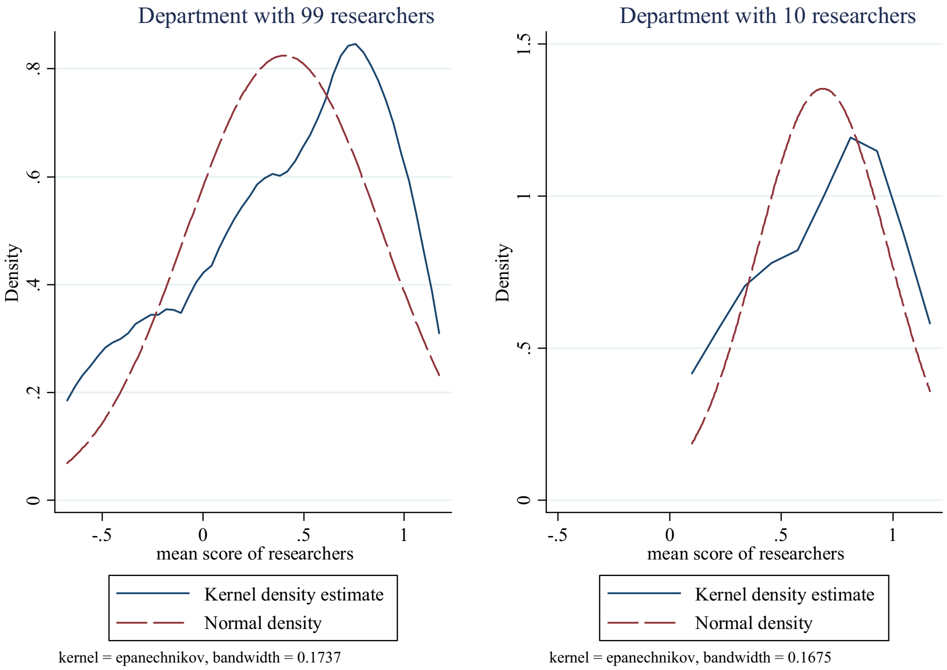

Fig. 1. Distribution of the mean scores of researchers in two departments in area 01 (mathematics and computer sciences).

two distributions are left skewed but we do not observe significant differences between departments of different size[2]. The example supports the convergence to a symmetric distribution in department with at least 10 researchers.

The distributions of random variables can be described by its mean and variance.

The mean score of department A is estimated by

$$I_A = \frac{\sum_{i=1}^{N_A} X_{Ai}}{n_A}$$

And the variance by

$$s_A^2 = \sum_{i=1}^{N_A} \frac{(X_{Ai} - I_A)^2}{n_A - 1}$$

The test is the very simple t-test (Student t) for the equality of the means. The test is distributed as follows:

$$t_{(n_A+n_B-2)} = \frac{I_A - I_B}{\sqrt{\left(s_A^2/n_A\right) + \left(s_B^2/n_B\right)}}$$

where $I_A - I_B$ is the difference between means of the two departments compared. s_A^2 and s_B^2 are the variances of the first and the second department, respectively. n_A and n_B are the sizes of the first and the second department, respectively.

For any pair of departments we are interested to verify if the null hypothesis $I_A = I_B$ cannot be rejected under our assumptions. The procedure we have devised, following Lubrano, is practical: rank all departments according to a chosen criterion, start with the department ranked first and test whether the score is statistically different from the score of the department ranked second, using the t-statistics. If the p-value is not low (that is, it exceeds 0.05), they are part of the same group. Iterate the procedure as follows: compare the department ranked third with the first and place it in the same group if the difference is not significant, then move to ranked fourth, fifth and so on, until you find a statistically significant difference with the first. At this point a different group is identified. The top member of this group is statistically different from the top member of the first group, although it will not be different, in general, from those at the bottom of the first group. Iterate the procedure until you find another significant difference, at which point a new group is created.

The choice of 0.05 is arbitrary but is a "compromise" between the need to have typical values of probability accepted in human and social sciences studies and to obtain a reasonable number of clusters. In fact a very low p-value decreases the

[2] Figures for all scientific fields are available from the authors upon request.

degree of discrimination between departments and involves the risk to obtain an unique partition; instead a high *p*-value increase the possibility of discrimination and induce the proliferation of clusters.

In this way what we obtain is a partition of the list into meaningful groups, each of which is relatively homogeneous within, and different from other groups. Within groups, the ranking has not any statistically meaningful value. This means that rankings may be produced under the conditions that some decisions are made with reference to the construction of the score and the aggregation rule, and have a validity strictly defined by these assumptions.

5. The nondeterministic approach to university rankings: Potential and limitations

A further methodological consideration has to do with whether we might stay happy with a simple examination of confidence intervals, instead of adopting the test of hypothesis approach. To start with, in scientometric studies the use of null hypothesis statistical significance tests (NHST) is controversial and there are different opinions on their merits in the literature. In particular, many null hypotheses are considered implausible, while the crucial assumption of randomness is questioned, and the typical dichotomous application in decision making is criticized (see among others Schneider, 2013, 2015). Schneider stressed the importance of human judgment in order to establish the presence of real differences in rankings or impact factors and criticized the adoption of decision making procedures based on the mechanical application of tests that are logically flawed.

Let us address these criticisms. In our case, we believe that the null hypothesis has a clear meaning: we want to understand to what extent the position of a department in a rank list is due to random effects. Indicators of variability greatly improve the meaning of rankings of departments. With respect to a simple analysis of the standard deviation or confidence intervals, however, the procedure suggested by Lubrano offers a great advantage: it allows the reliable construction of relatively homogeneous groups.

In order to understand this point, let us go back to the role of university rankings. In the Italian case, the publication of a full ranking based on cardinal measures was made mandatory by the legal provision that it would be used for the allocation of financial resources by the Ministry of Research, i.e. following a continuous variable in monetary terms. Given the large differences in size among universities, hence in the annual funding from the Ministry, a cardinal measure that might be easily translated in monetary terms was a clearly preferred choice.

However, if one wants that rankings are also utilized systematically as strategic tools in university governance and administration, the question of reliability becomes crucial. From this perspective, rankings are challenged on methodological grounds, as we have shown above, for their lack of consideration for the underlying variability and poor robustness. For this reason, we believe that a robust classification of departments in homogeneous groups conveys more accurate and strategically useful information than rankings, which, although based on cardinal measures, are unable to give adequate recognition to the underlying variability. The apparent objectivity of cardinal measures leading to rankings, in fact, is a source of acute media and users interest, but is, on the contrary, a very poor guide for strategic decisions of universities. The literature on university rankings has repeatedly observed that they do not provide useful guidelines for all universities below the top (say, below the 50th or 100th position worldwide), while for all those at the top the main strategic guideline is quite simply to try to remain in the top league. For ranking information to be strategically useful, it is crucial that changes in the ranking actually represent *real* changes in performance. If there is variability around the indicators, it is not obvious whether an improvement (or a deterioration) in the ranking is the result of deliberate action, or is largely a random effect. But if a causal effect cannot be reliably stated, then the ranking information is useless as a guide for strategic decisions.

For these substantive reasons we believe that the creation of statistically different ordered groups is a powerful way to make rankings a bottom up strategic tool, in addition to their role as a device for top down financial allocations. Working with a relatively small number of ordered groups makes strategic decisions more intelligible and supports the adoption of clearly communicated and shared objectives. For example, moving up from one group to the higher one is a significant strategic achievement. On the contrary, improving the cardinal measure of the ranking but remaining in the same group does not add too much to the strategy of the university. This is of great importance from the point of view of university academic leaders and administrators. Many important decisions, from allocation of internal resources to hiring and promotion, could be based on a relatively small number of ordered and robust groups, maintain strategic clarity and vision and minimizing the opportunistic behaviors associated to number gaming.

Note that this approach, differently from the one based exclusively on ratings (e.g. the star-system used in the UK), is indeed based on scores (numbers) that may lead to cardinal measures and rankings. It recognizes a difference between the top down approach in financial allocations at Ministry level and the more diffused utilization for strategic decisions at university level.

Having said that, it is important to recognize the limitations of the procedure suggested by Lubrano. Its main limitation is that it compares among themselves only the top department in each group with lower-ranked departments in the ranking. In practice, it might be that the top department is statistically different from the top department of the group ranked higher, but the second-to-the-top or the third or others are not. This counterintuitive possibility depends, of course, on the size of the variability of the departments ranked below the top of the respective group. While this cannot be excluded altogether, a robustness analysis could be carried out on the departments close to the top and the bottom of each group. This could be done by publishing, together with the average score, the individual standard deviations, so that each case could be controlled in a transparent way. Indeed, in the future rankings might be published not only with the associated standard deviations,

Table 1
Descriptive statistics of average score of departments in disciplinary areas.

Area code[*]	Number of departments	Mean score	Median score	Standard deviation	Coefficient of variation	Skewness
1	68	0.615	0.648	0.178	0.289	−1.063
2	42	0.810	0.806	0.086	0.106	−1.171
3	70	0.804	0.813	0.081	0.101	−0.127
4	30	0.603	0.585	0.125	0.207	−0.061
5	143	0.635	0.668	0.174	0.274	−0.828
6	184	0.477	0.486	0.220	0.461	−0.302
7	55	0.616	0.644	0.168	0.273	−0.619
8	75	0.541	0.540	0.121	0.224	−0.294
9	97	0.720	0.748	0.150	0.208	−1.650
10	125	0.650	0.672	0.126	0.194	−1.890
11	134	0.575	0.583	0.136	0.236	−0.072
12	111	0.491	0.514	0.150	0.305	−1.190
13	110	0.327	0.315	0.174	0.532	0.375
14	44	0.440	0.432	0.125	0.284	−0.183
Total	1.288[*]					

[*] Every department is counted as many times as the number of scientific areas in which it is active (with more than 30 products).
[‡] 1. Mathematics and computer sciences; 2. Physics; 3. Chemistry; 4. Earth sciences; 5. Biology; 6. Medicine; 7. Agricultural and veterinary sciences; 8. Civil engineering and architecture; 9. Industrial and information engineering; 10. Antiquities, philology, literary studies, art history; 11. History, philosophy, pedagogy and psychology; 12. Law; 13. Economics and statistics; 14. Political and social sciences.

but also with an embedded simple tool to calculate the statistical significance of groupings (for example by changing the p-value), allowing full interactivity of the information.

On the contrary, the procedure used in this paper does not solve the issue of ranking of departments or universities based on measures that aggregate the individual performance *at the origin*, such as the h-index at departmental level, or are based on percentile bibliometric measures. In addition, it cannot be applied to all cases in which standard deviations of individual scores are not published, as it happens in the Research Assessment Framework in UK.

6. Applying the statistical ranking procedure to Italian data

The dataset used for the analysis is derived from the VQR and is summarized in Table 1[3]. We consider 1.288 departments with at least 10 researchers in the area.

On the basis of the methodological discussion in Sections 4 and 5, we applied the procedure suggested by Lubrano (2009) to all departments subject to VQR. We first calculated the mean values of scores assigned to individual products submitted by researchers affiliated to all departments. As stated above, the number of products submitted by university professors and researchers was three, with a limited number of exceptions due to individuals who were recruited late within the 2004–2010 time window, or enjoyed on leave periods. Using the individual values, we then calculated the means and the standard deviation in every department.

On the basis of the standard deviation and the degrees of freedom, we iteratively computed the t-test, as suggested by the procedure. The t-test was considered appropriate since: (a) the number of observations is very large; (b) individual scores are distributed with unimodal distributions, although not necessarily normal. In addition to these technical issues, it allows a direct comparison with the results of Lubrano (2009).

To illustrate the procedure, consider Fig. 2, again taken from Area 1 (Mathematics). All departments have their own mean value, with the associated standard deviation of researchers and are ranked from top to bottom. The best department has an astonishing average score of 0.977, the worst one has −0.067. It is clear from the width of the standard deviation that there is large variability *within* departments, so that departments that are close in the rankings might indeed be non-distinguishable in statistical terms. On the right-hand axis the value of the probability associated to the t-test is then reported. Once the value falls below 0.05 a statistically significant difference is identified. In the case of Mathematics, it happens with the department ranked #1 (the only member of Group 1), then the department ranked #7 (Group 2 includes 6 members), ranked #30 (Group 3 includes 23 members), then #50 and #67. There is only one member of Group 6.

The horizontal distance between two successive points in which the t-test line intersects the horizontal line representing the confidence interval gives an indication of the size of the group. The p-value curve intersects the line in five points, generating six groups. Note that in this procedure the number of groups is not imposed externally but is generated by the data themselves.

[3] γ_1 (Gamma) of Fisher is used to estimate the degree of skewness. We use this type of *normalized* index to take into account differences in number of departments among areas.

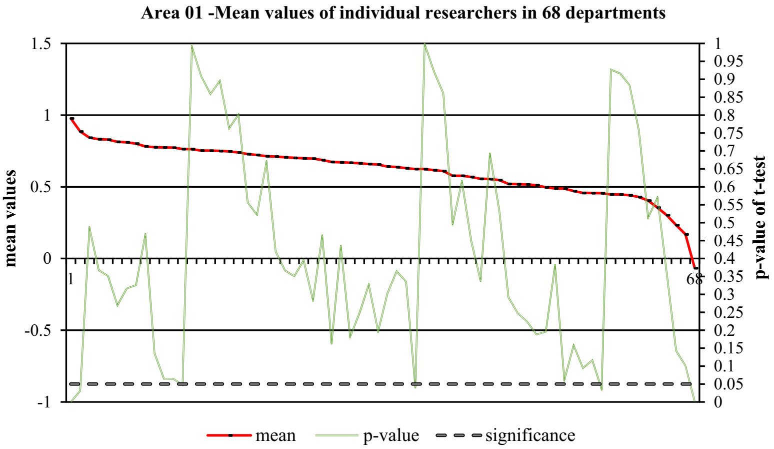

Fig. 2. Mean score across departments and probabilities associated to the *t*-test. Area 1 (mathematics).

As reported in Section 4, the *p*-value of 0.05 is conventional and could be changed in future replications of this study. On the one hand, it is a value commonly accepted in social science studies. On the other hand, it is clear that by changing the *p*-value the null hypothesis would be rejected at different points of the distribution, so that the number of clusters might be different from the one reported here. However, we are not interested in any discussion on the true number of clusters. We are mainly interested in showing how we depart from the starting condition in which all members of the ranking are considered different from each other. By taking a conventionally accepted value of statistical significance we already obtain a remarkable departure from this starting condition.

7. Main results and discussion

Tables 2a and 2b gives the descriptive statistics for the identified groups. We report data separately for bibliometric and non-bibliometric sectors (i.e. for STEM and SSH disciplines), given that the evaluation procedures have been largely different. In these tables we calculate the groups by using the average score of individual researchers, which has a maximum at 1 (all three products are excellent). In the bibliometric sectors (called 1–9 in the administrative jargon and covering science, technology and medicine) the use of peer review has been kept to the minimum, being used for very recent articles (i.e. articles for which the citation window was very short and the number of citations was considered an unreliable indicator of quality) or for those articles for which there was sharp disagreement between the indicators based on citations and on the impact factor of the journal. In addition, a control sample of articles has been submitted to both bibliometric analysis and peer review, in order to learn from the comparison of the two methodologies. Summing up, in these areas research products have been consistently evaluated with bibliometric approaches.

On the contrary, in non bibliometric sectors (called 10–14 and covering Humanities and Social Sciences) peer review has been dominant, up to 100% of cases in Humanities and Arts and in Law. An important exception is Economics and Management, in which articles have been evaluated bibliometrically, while books and chapters of books have been subject to peer review.

There are a number of results that apply equally well to bibliometric and non bibliometric sectors (i.e. for STEM and HSS).

First, the number of statistically distinguishable groups is in the range between 3 and 7. Only three groups have been identified in the smallest scientific areas, such as Earth Sciences ($n = 30$ departments), or in large areas with lower variability in quality of research, such as Chemistry ($n = 70$ departments). Up to seven groups have been identified in the largest areas, such as Medicine, Industrial and information engineering, and History, philosophy, pedagogy and psychology. Thus, in all other disciplines, the typical situation is one in which the number of distinguishable groups falls between 4 and 6. This is a remarkable result. On one hand, it is shown that the differences in rankings originate from differences in the underlying quality distribution that are not always statistically significant. It may be legitimate to produce rankings based on deterministic measures of research quality in order to allocate financial resources using continuous distributions. But from an evaluative point of view it must be kept in mind that the differences that have solid ground in the underlying reality are only a few.

At the same time, departments *do* differ among themselves in a significant way. Even considering the large variability among individual researchers, there are clear differences in the mean values. Good departments have a significantly larger share of high quality researchers. Although good researchers are found almost everywhere (Bonaccorsi & Cicero, 2015), it does not seem that departments are completely random collections of researchers. Rather, depending on their history, the scientific tradition, the recruitment policy, the overall prestige of the university and other contextual factors, departments are found in different quality layers. The possibility for departments to actively change their position over time is an interesting issue, for which we will have to observe the data across subsequent assessment exercises.

Table 2a
Descriptive statistics of quality of research of Italian departments. Bibliometric sectors.

Group	Upper score	Group average score	Number of departments	Percentage of departments (%)
Mathematics and computer sciences				
1	0.977	0.977	1	1.47
2	0.885	0.811	11	16.18
3	0.763	0.701	25	36.76
4	0.624	0.539	20	29.41
5	0.455	0.368	10	14.71
6	−0.067	−0.067	1	1.47
Total			68	100.00
Physics				
1	0.971	0.959	2	4.76
2	0.926	0.888	11	26.19
3	0.851	0.809	18	42.86
4	0.762	0.708	11	26.19
Total			42	100.00
Chemistry				
1	0.959	0.921	12	17.14
2	0.884	0.830	33	47.14
3	0.772	0.715	25	35.71
Total			70	100.00
Earth sciences				
1	0.825	0.770	7	23.33
2	0.692	0.605	14	46.67
3	0.534	0.469	9	30.00
Total			30	100.00
Biology				
1	0.967	0.911	7	4.90
2	0.857	0.791	40	27.97
3	0.718	0.655	51	35.66
4	0.559	0.486	34	23.78
5	0.367	0.320	8	5.59
6	0.103	0.075	3	2.10
Total			143	100.00
Medicine				
1	0.950	0.906	3	1.63
2	0.855	0.767	29	15.76
3	0.701	0.630	37	20.11
4	0.564	0.499	41	22.28
5	0.427	0.368	37	20.11
6	0.273	0.200	25	13.59
7	0.104	0.036	12	6.52
Total			184	100.00
Agricultural and veterinary sciences				
1	0.933	0.892	5	9.09
2	0.812	0.725	17	30.91
3	0.670	0.606	20	36.36
4	0.497	0.440	9	16.36
5	0.365	0.259	4	7.27
Total			55	100.00
Civil engineering and architecture				
1	0.790	0.717	12	16.00
2	0.671	0.602	22	29.33
3	0.547	0.519	22	29.33
4	0.477	0.438	11	14.67
5	0.371	0.310	8	10.67
Total			75	100.00
Industrial and information engineering				
1	0.974	0.943	4	4.12
2	0.918	0.895	7	7.22
3	0.848	0.806	34	35.05
4	0.756	0.711	27	27.84
5	0.645	0.581	21	21.65
6	0.437	0.425	2	2.06
7	0.097	0.087	2	2.06
Total			97	100.00

Table 2b
Descriptive statistics of quality of research of Italian departments. Non-bibliometric sectors.

Group	Upper score	Group average score	Number of departments	Percentage of departments (%)
Antiquities, phylology, literary studies, art history				
1	0.822	0.782	20	16.00
2	0.756	0.718	41	32.80
3	0.673	0.631	40	32.00
4	0.575	0.494	21	16.80
5	0.295	0.175	3	2.40
Total			125	100.00
History. philosophy. pedagogy and psychology				
1	0.984	0.984	1	0.75
2	0.912	0.867	3	2.24
3	0.794	0.725	26	19.40
4	0.682	0.636	34	25.37
5	0.586	0.538	38	28.36
6	0.459	0.404	30	22.39
7	0.263	0.237	2	1.49
Total			134	100.00
Law				
1	0.729	0.653	26	23.42
2	0.601	0.544	42	37.84
3	0.467	0.402	31	27.93
4	0.308	0.234	10	9.01
5	0.086	−0.037	2	1.80
Total			111	100.00
Economics and statistics				
1	0.790	0.657	9	8.18
2	0.588	0.512	21	19.09
3	0.435	0.347	34	30.91
4	0.257	0.188	38	34.55
5	0.101	0.051	8	7.27
Total			110	100.00
Political and social sciences				
1	0.733	0.629	6	13.64
2	0.568	0.507	15	34.09
3	0.434	0.373	20	45.45
4	0.205	0.173	3	6.82
Total			44	100.00

Second, our results help to reformulate the debate on qualitative or quantitative bases for evaluation. Going back to the literature cited in the introductory section, most studies that are critical toward rankings would be ready to accept the notion that qualitative differences among universities do exist. The argument that these differences are incommensurable by nature, that is, they *cannot* be transformed into a coarse-grain qualitative ordering is held by a small minority of authors, often with an inclination against any evaluation whatsoever. The large majority of scholars accept the notion that qualitative differences do exist, so that a general relation of order, of the type “better than” can be accepted. In general, researchers have an intuitive notion of what is good in research. Still, many scholars, particularly in humanities and social sciences find it difficult to accept that these intuitive qualitative differences can be transformed into complete ordering using quantitative methods. Their position is that a rating of departments or universities into coarse-grain classes or groups would be acceptable, while a ranking would not be acceptable due to the lack of statistical robustness. The statistical ranking methodology indeed delivers, as a result, a small number of homogeneous groups. While this result is clearly based on purely quantitative measures, it offers a summary representation which is intuitively closer to the representation held by scholars who distrust the measurement of research quality. In this sense the methodology builds a bridge between quantitative and qualitative approaches to research assessment. An intuitively acceptable coarse-grain representation is generated on top of a robust fine-grained set of information.

Third, it is interesting to examine the distribution of departments across the groups (Fig. 3). All distributions of the relative frequency of occurrence across the groups are unimodal, irrespective of the number of groups. The modal group is #3 in basic sciences and medicine, while it is #2 in applied disciplines (Agrarian and Veterinary Sciences, Architecture and Engineering) and in all Humanities and Social Sciences. In all cases the groups at the extremes include a small minority of departments.

It is very challenging to stay in Group 1, the excellent group. Counting all departments included in Group 1 we find 185 cases, or 14.4% of the total ($n = 1288$). In all bibliometric sectors, no more than 20% of departments stay in Group 1, with minimum values at 1.47% in Mathematics, 1.63% in Medicine, 4.12% in Engineering. Similarly, the worst group includes a number of departments that exceeds 10% of the total only in Mathematics, Chemistry and Architecture, among departments

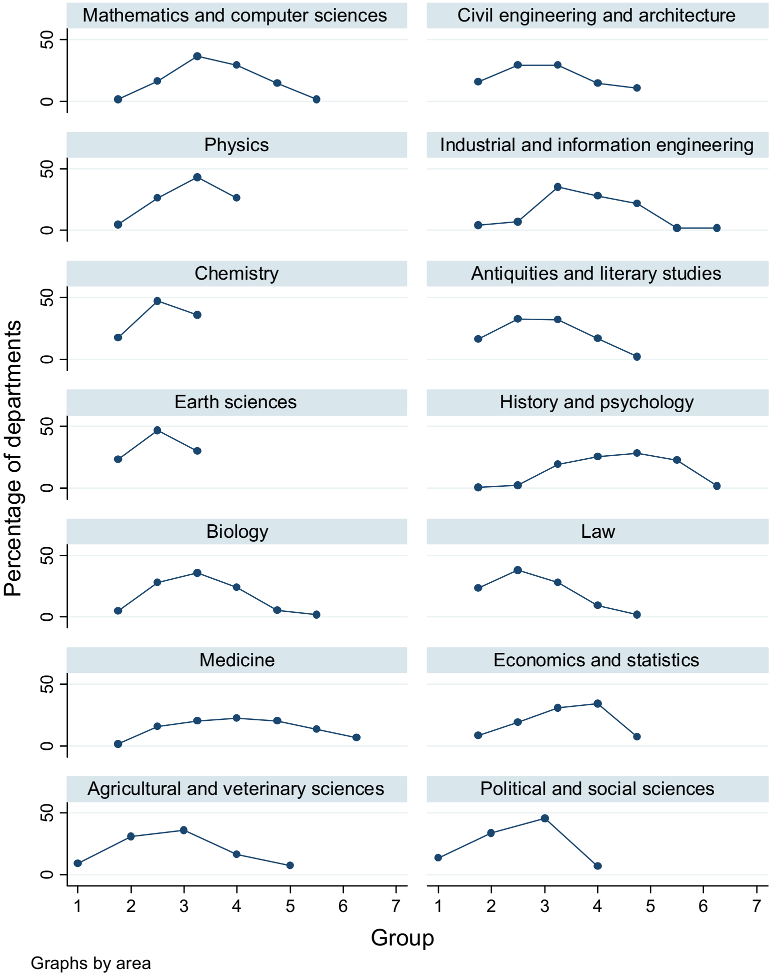

Fig. 3. Distribution of departments in groups.

with at least four groups[4]. Thus the distribution of departments is unimodal, with the mode oscillating between groups #2 and #3, and with upper and lower tails that are not "fat".

Fourth, it is useful to examine the differences between groups. This information is important to heads of departments, and more generally to decision makers, in order to evaluate the challenge posed by the goal to improve one's position. Let us assume for a moment that departments may improve their position, keeping other departments constant (which is not obviously the case). In other words, it is not important to move up the ranking, as it is does not mean an improvement that is statistically distinguishable from a random effect. The real improvement would be visible if a department would be able to jump from a group to the next group in a statistically robust ordering. Otherwise what we would obtain is only "Brownian motion" of departments, moving some positions without any significant improvement. We calculated the first difference between the score of the best department in each of the groups. We then calculated the relative difference as the percentage of improvement from the lower group to the upper group (Table 3). Since the relative difference may be exceedingly large in moving from the worst group to the next one, also considering that the worst group is formed by few cases, we dropped all first differences of this kind.

Take for example Law, or Economics and statistics. To move up from Group 4 to Group 3 would require an improvement over the starting position of 51.4% and 69.4%, respectively, which is clearly a difficult challenge. How challenging is, on average across all groups, the upward mobility? Our 1,288 departments are grouped into 72 groups in 14 areas. Considering that first differences in a group of n units are $(n-1)$, this gives us 72–14 = 58 differences. Dropping the worst groups it gives us 58–14 = 44 differences. On average, these differences are 23.2%: in order to move from a group to the next one, it takes an improvement which is 23.2% over the starting score. Once we take away the worst group, moving up from the next worst

[4] In calculating this share we collapse a few cases in which the worst groups is composed of just one department (History and Philosophy, Humanities and Arts, Mathematics).

Table 3
First differences among groups in average score and rates of growth in upward mobility.

Group	Upper score	First difference	Inter-class rate of growth (%)	Number of departments
Mathematics and computer sciences				
1	0.977	0.091	10.3	1
2	0.885	0.122	16.0	11
3	0.763	0.140	22.4	25
4	0.624	0.169	37.1	20
5	0.455	0.522	n.a.	10
6	−0.067			1
Total				68
Physics				
1	0.971	0.045	4.9	2
2	0.926	0.075	8.8	11
3	0.851	0.089	11.6	18
4	0.762			11
Total				42
Chemistry				
1	0.959	0.074	8.4	12
2	0.884	0.113	14.6	33
3	0.772			25
Total				70
Earth sciences				
1	0.825	0.133	19.2	7
2	0.692	0.158	29.5	14
3	0.534			9
Total				30
Biology				
1	0.967	0.109	12.8	7
2	0.857	0.139	19.4	40
3	0.718	0.159	28.5	51
4	0.559	0.192	52.3	34
5	0.367	0.264	256.8	8
6	0.103			3
Total				143
Medicine				
1	0.950	0.095	11.2	3
2	0.855	0.154	22.0	29
3	0.701	0.137	24.3	37
4	0.564	0.137	32.0	41
5	0.427	0.155	56.8	37
6	0.273	0.169	163.3	25
7	0.104			12
Total				184
Agricultural and veterinary sciences				
1	0.933	0.122	15.0	5
2	0.812	0.142	21.1	17
3	0.670	0.173	34.8	20
4	0.497	0.132	36.1	9
5	0.365			4
Total				55
Civil engineering and architecture				
1	0.790	0.119	17.7	12
2	0.671	0.124	22.6	22
3	0.547	0.071	14.8	22
4	0.477	0.106	28.5	11
5	0.371			8
Total				75
Industrial and information engineering				
1	0.974	0.056	6.1	4
2	0.918	0.071	8.3	7
3	0.848	0.092	12.2	34
4	0.756	0.111	17.2	27
5	0.645	0.208	47.5	21
6	0.437	0.340	349.8	2
7	0.097			2
Total				97
Antiquities. phylology. literary studies. art history				
1	0.822	0.066	8.7	20
2	0.756	0.083	12.3	41
3	0.673	0.099	17.1	40

Table 3 (*Continued*)

Group	Upper score	First difference	Inter-class rate of growth (%)	Number of departments
4	0.575	0.279	94.5	21
5	0.295			3
Total				125
History. philosophy. pedagogy and psychology				
1	0.984	0.072	7.9	1
2	0.912	0.118	14.9	3
3	0.794	0.112	16.5	26
4	0.682	0.095	16.2	34
5	0.586	0.127	27.8	38
6	0.459	0.196	74.7	30
7	0.263			2
Total				134
Law				
1	0.729	0.128	21.3	26
2	0.601	0.134	28.7	42
3	0.467	0.158	51.4	31
4	0.308	0.222	258.3	10
5	0.086			2
Total				111
Economics and statistics				
1	0.790	0.202	34.3	9
2	0.588	0.153	35.2	21
3	0.435	0.178	69.4	34
4	0.257	0.155	153.2	38
5	0.101			8
Total				110
Political and social sciences				
1	0.733	0.165	29.0	6
2	0.568	0.134	30.9	15
3	0.434	0.229	112.0	20
4	0.205			3
Total				44

group to the next best group would require on average an improvement of 34%. Since these results must be achieved with the effort of all members of a department, mobility up from the bottom positions seem quite challenging. However, this is exactly the kind of information needed by university leaders if they want to improve the position over time, while keeping high the consensus among the colleagues and the strategic momentum.

8. Conclusions

We have provided an exercise to classify departments of Italian universities into groups that are statistically distinguishable from others. This follows from the theoretical notion that scores of research quality, whatever the methodology by which they are produced, must be interpreted as random variables, following a distribution. Given this premise, the mean value of the distribution, which is the only value assumed in conventional university rankings, misses important information.

By adopting the procedure suggested by Lubrano (2009) we identified in all 14 scientific areas as many as 72 groups in which 1288 departments can be classified. In each area the number of groups is between 3 and 7, but in most cases between 4 and 6. The exercise lends support to the criticism to rankings based on the notion of statistical ranking, as opposed to deterministic. At the same time it identifies a number of groups that allow robust qualitative evaluation of departments, with sound policy and managerial implications.

The number of clusters cannot be aggregated in a single number.

We however suggest that the ordered clustering of departments can also be used as a strategic tool for universities. For example, a simple way of using the clustering of departments is counting how many departments a university has in cluster 1, 2, . . ., 7 across the disciplinary areas, and which share of the total do they represent. Needless to say, this does not transform into a ranking, because there is a need for weighting the clusters and the departments from various areas (Docampo, 2012). Instead of a single ranking, this counting could deliver a number of highly valuable information to be used for strategic considerations.

References

Aguillo, I., Ilan, J., Levene, M., & Ortega, J. (2010). Comparing university rankings. *Scientometrics*, *85*, 243–256.

Bonaccorsi, A., & Cicero, T. (2015). Distributed or concentrated research excellence? Evidence from a large scale research assessment exercise. *Journal of the American Society for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23539

Bornmann, L., Mutz, R., & Daniel, H. D. (2013). Multilevel-statistical reformulation of citation-based university rankings: The Leiden Ranking 2011/2012. *Journal of the American Society for Information Science and Technology*, *64*(8), 1649–1658.

Boulton, G. (2011). University rankings: Diversity, excellence and the European initiative, Procedia. *Social and Behavioral Sciences*, *13*, 74–82.

Bowden, R. (2000). Fantasy higher education: University and college league tables. *Quality in Higher Education*, *6*(1), 41–60.

Brooks, R. L. (2005). Measuring university quality. *Review of Higher Education*, *29*(Fall (1)), 1–22.

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics*, *71*(3), 349–365.

Cantwell, B., & Taylor, B. J. (2013). Global status, intra-institutional stratification and organizational segmentation: A time-dynamic Tobit analysis of ARWU position among U.S. Universities. *Minerva*, (51), 195–223.

Carayol, N., Filliatreau, G., & Lahatte, A. (2015). Impact, dominance et classements des universités. *Revue Economique*, *66*(1), 173–194.

Carayol, N., & Lahatte, A. (2014). Dominance relations and ranking when both quantity and quality matter. Applications to US universities and worldwide economic departments. In *Gretha working paper no. 14*.

Chen, K., & Liao, P. (2012). A comparative study on world university rankings: A bibliometric survey. *Scientometrics*, *92*, 89–103.

Cicero, T., Malgarini, M., & Benedetto, S. (2014). Research quality, characteristics of publications and socio-demographic features of universities and researchers: Evidence from the Italian VQR 2004–2010 evaluation exercise. In *STI 2014 Leiden* (p. 106).

Daraio, C., & Bonaccorsi, A. (2015). Beyond university rankings? Generating new indicators on European universities by linking data in open platforms. *Journal of the Association for Information Science and Technology*,. Forthcoming.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. *European Journal of Operational Research*, *1*, 1–15.

Dehon, C., Jacobs, D., & Vermandele, C. (Eds.). (2009). *Ranking universities*. Brussels: Editions de l'Université de Bruxelles.

Dill, D., & Soo, M. (2005). Academic quality, league tables, and public policy: A crossnational analysis of university ranking systems. *Higher Education*, *49*, 495–533.

Docampo, D. (2012). Adjusted sum of institutional scores as an indicator of the presence of university systems in the ARWU ranking. *Scientometrics*, *90*, 701–713.

Erkkilä, T. (2013). *Global university rankings. Challenges for European education*. Basingstoke: Palgrave Macmillan.

Estermann, T., & Nokkala, T. (2009). *University autonomy in Europe: Exploratory study*. Brussels: European University Association.

Estermann, T., Nokkala, T., & Steinel, M. (2010). *University autonomy in Europe II. The scorecard*. Brussels: EUA.

European University Association. (2012). *Public funding observatory*, http://www.eua.be/eua-work-and-policy-area/governance-autonomy-andfunding/public-funding-observatory.aspx.

Freyer, L. (2014). Robust rankings. Review of multivariate assessments illustrated by the Shanghai rankings. *Scientometrics*, *100*, 391–406.

Goldstein, H., & Spiegehalter, D. J. (1996). League tables and their limitations. Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A.*, *159*(3), 385–443.

Harvey, E. L. (2008). Rankings of higher education institutions: A critical review. *Quality in Higher Education*, *14*(3), 187–207.

Hattendorf, L. (1986). College and university rankings. *RQ*, *25*(3), 332–347.

Hazelkorn, E. (2007). The impact of league tables and ranking systems on Higher Education decision making. *Higher Education Management and Policy*, *19*(2), 89–112.

Hazelkorn, E. (2009). Rankings and the battle for world-class excellence: Institutional strategies and policy choices. *Higher Education Management and Policy*, *21*(1), 1–22.

Hazelkorn, E. (2011). *Rankings and the reshaping of higher education: The battle for world-class excellence*. Houndsmill, UK: Palgrave-Macmillan.

Institute for Higher Education Policy. (2007). *College and university ranking systems: Global perspective and American challenges*. Washington, DC: Author.

Institute for Higher Education Policy. (2009). *Impact of college rankings on institutional decision making: Four country case studies*. Washington, DC: Author.

Ioannidis, J. P. A., Patsopoulos, N. A., Kavvourai, F. K., Tatsioni, A., Evangelou, E., Kouri, I., et al. (2007). International ranking systems for universities and institutions: A critical appraisal. *BMC Medicine*, *5*(30), 1–9, 2007.

Kehm, B. M., & Stensaker, B. (Eds.). (2009). *University rankings, diversity, and the new landscape of higher education*. Rotterdam, Netherlands: Sense Publishers.

Lubrano, M. (2009). A statistical approach to rankings: Some figures and explanations for European universities. In C. Dehon, D. Jacobs, & C. Vermandele (Eds.), *Ranking universities*. Brussels: Editions de l'Université de Bruxelles.

Lubrano, M., Kirman, A., Bauwens, L., & Protopopescu, C. (2003). Ranking economics deparments in Europe: A statistical approach. *Journal of the European Economic Association*, *1*, 1367–1401.

Lubrano, M., & Protopopescu, C. (2004). Density inference for ranking European research systems in the field of economics. *Journal of Econometrics*, *123*, 345–369.

Lubrano, M., Bauwens, L., Kirman, A., & Protopopescu, C. (2003). *Core discussion paper 2003/50*.

Lukman, R., Krajinc, D., & Glavic, P. (2010). University ranking using research, educational and environmental indicators. *Journal of Cleaner Production*, *18*(2010), 619–662.

O'Connell, C. (2013). Research discourses surrounding global university rankings: Exploring the relationship with policy and practice recommendations. *Higher Education*, *65*, 709–723.

Provan, D., & Abercromby, K. (2000). University league tables and rankings. *Research in Higher Education*, *45*(5), 443–461.

Rauhvargers, A. (2011). *Global university rankings and their impact: EUA report on rankings, 2011*. Brussels, Belgium: European University Association.

Safon, V. (2013). What do global university rankings really measure? The search for the X factor and the X entity. *Scientometrics*, *97*, 223–244.

Saisana, M., & D'Hombres, B. (2008). *Higher education rankings: Robustness issues and critical assessment. How much confidence can we have in HE rankings? European Commission Joint Research Centre*. Institute for the Protection and Security of the Citizen. Centre for Research on Lifelong Learning (CRELL).

Saisana, M., D'Hombres, B., & Saltelli, A. (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research Policy*, *40*, 165–177.

Salmi, J. (2009). *The challenge of establishing world-class universities*. Washington, DC: The World Bank.

Salmi, J., & Saroyan, A. (2007). League tables as policy instruments: Uses and misuses. *Higher Education Management and Policy*, *19*(2), 33–70.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*(1), 411–432.

Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics*, *7*(1), 50–62.

Shin, J. C., Toutkoushian, R. K., & Teichler, U. (Eds.). (2011). *University rankings. Theoretical basis, methodology and impacts on global higher education*. Dordrecht: Springer.

Taylor, P., & Braddock, R. (2007). International university ranking systems and the idea of university excellence. *Journal of Higher Education Policy and Management*, *29*(3), 245–260.

Turner, D. R. (2005). Benchmarking in universities: League tables revisited. *Oxford Review of Education*, *31*(3), 353–371.

Unesco. (2013). Rankings and accountability in higher education. In P. T. M. Merope, P. J. Wells, & E. Hazelkorn (Eds.), *Uses and misuses*. Paris: Unesco.

Usher, A., & Savino, M. (2006). *A world of difference: A global survey of university league tables*. Toronto, ON: Educational Policy Institute.

Van Dyke, N. (2005). Twenty years of university report cards. *Higher Education in Europe*, *30*(2), 103–125.