Research Policy 40 (2011) 165–177

Contents lists available at ScienceDirect

Research Policy

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

Rickety numbers: Volatility of university rankings and policy implications

Michaela Saisana*, Béatrice d'Hombres, Andrea Saltelli

Econometrics and Applied Statistics, Joint Research Centre, European Commission, Enrico Fermi 2749, 21027 Ispra, Italy

ARTICLE INFO

Article history:
Available online 13 October 2010

Keywords:
University
Ranking
Composite indicator
Uncertainty analysis
Sensitivity analysis

ABSTRACT

The most popular world university rankings are routinely taken at face value by media and social actors. While these rankings are politically influential, they are sensitive to both the conceptual framework (the set of indicators) and the modelling choices made in their construction (e.g., weighting or type of aggregation). A robustness analysis, based on a multi-modelling approach, aims to test the validity of the inference about the rankings produced in the Academic Ranking of World Universities (ARWU) of Shanghai Jiao Tong University and those produced by the UK's Times Higher Education Supplement (THES). Conclusions are drawn on the reliability of individual university ranks and on relative country or macro regional performance (e.g., Europe versus USA versus China) in terms of the number of top performing institutions. We find that while university and country level statistical inferences are unsound, the inference on macro regions is more robust. The paper also aims to propose an alternative ranking which is more dependant on the framework than on the methodological choices.

1. Introduction

Higher education and its relation to economic growth have figured prominently on the European political agenda in recent years (Acemoglu et al., 2006; Aghion et al., 2008; Sapir et al., 2004). The debate on the quality and performance of higher education systems in Europe has been very much stimulated by the annual publication, since 2003, of Shanghai Jiao Tong University's Academic Ranking of World Universities (henceforth ARWU), which compares university research performance across the world and tends to support the evidence that the USA is well ahead of Europe in terms of cutting-edge university research. The ARWU's main rival ranking, which has been produced annually by the UK's Times Higher Education Supplement (henceforth THES) since 2004, confirms the USA's lead over European institutions. Both these rankings attract worldwide attention.[1]

Traditionally, universities – as the name suggests – have tended to comprehend a broad range of purposes and dimensions; today, given the increasing vertical and horizontal differentiation of universities and their greater diversity in focus and mission, a comparison of overall university performance is arduous. Still, for reasons of governance, accountability and quality assurance, there is increasing interest among policy-makers, stakeholders and practitioners in measuring and benchmarking "excellence" across universities. The growing mobility of students and researchers has also created a market for these measures among prospective students and academics.

Such composite indicators meet the growing appetite for statistical information, which the Stiglitz report (2009, p. 7) attributes to our increasing statistical literacy and ease of access to data.[2] The International Ranking Expert Group (IREG[3]), established in 2004, regularly organises conferences on academic rankings. Recent papers (Aghion et al., 2008; Veugelers and van der Ploeg, 2007; Jacobs and van der Ploeg, 2006) have examined the determinants of university performance using the ARWU as a proxy. Hazelkorn (2007) offers a cross-country analysis of the impact of international rankings on the decisions taken by university leaders.

University rankings excite the controversy that typically attends the publication of league tables. The annual publications of the ARWU and the THES ranking are extensively covered by the media and feed into national debates about the quality of the university system. In France the publication of the ARWU generally meets with a surge of media coverage either bemoaning the lacklustre performance of French universities or denying the capacity of the ARWU

* Corresponding author. Tel.: +39 0332 786572; fax: +39 0332 785733.
E-mail address: michaela.saisana@jrc.ec.europa.eu (M. Saisana).

[1] Besides these two international university rankings, there are more than 30 national university rankings in existence around the world (European Commission, 2008; Hendel and Stolz, 2008; Usher and Savino, 2007). Other international rankings based on bibliometrics are emerging, such as Webometrics' Ranking of World Universities or Performance Ranking of Scientific Papers for World Universities (http://www.webometrics.info/). Ederer et al. (2007) have recently proposed a ranking of university systems at the country level.

[2] "In the "*information society", access to data, including statistical data, is much easier. More and more people look at statistics to be better informed or to make decisions.*" (Stiglitz et al., 2009, p. 7)

[3] For additional information see http://www.ireg-observatory.org/.

0048-7333/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.respol.2010.09.003

to judge adequately the quality of French higher education (e.g., Le Monde, 14 August 2008; Les Echos, 7 August 2008). In Italy, public opinion has been more inclined to self-flagellation regarding the absence of an Italian university in the top 100 of the ARWU, as this appears to substantiate the public perception of the perilous state of the national educational system. In Spain, by contrast, the mere fact that a Spanish university could attain the top 200 of the ARWU is hailed as a great national achievement. Whether intended or not by their proponents, university rankings are now a tool of public and political discourse on national university systems.

However, several studies call into question the relevance of the indicators used in these rankings. The ARWU is mainly based on research performance, with no attempt to take into account other important dimensions of university activity (teaching in particular). It measures past and not current excellence (van Raan, 2007) and is biased towards large, English-speaking, hard-science institutions (van Raan, 2005; Zitt and Filliatreau, 2007; Williams, 2007). The THES ranking relies heavily on reputation indicators derived from expert opinion. Such indicators may be mere "symptoms" of excellence: they favour old institutions and do not represent current research performance (Taylor and Braddock, 2007; Marginson, 2007).

Nevertheless, David Hand, president of the Royal Statistical Society, stresses the pragmatic aspect of these measures: "*League tables [..] are not perfect, and they can never be [. . .] but they are certainly better than nothing [. . .]*", Hand (2004). In addition, as observed by a practitioner, "*because they [university rankings] define what "world-class" is to the broadest audience, these measures cannot be ignored by anyone interested in measuring the performance of tertiary education institutions*" (Salmi, 2009). Further, these measures have brought the debate on university performance into the public domain, which has at least the virtue of serving the principles of transparency and accountability.

It is hence worth investigating whether the political relevance of the ARWU and the THES ranking is supported by the methodological foundations of the indices; that is, whether these statistical constructions conceal any significant methodological flaws which might have implications for policy decisions. The purpose of this paper is to examine the impact of the methodological assumptions made in constructing the rankings on the actual placement of individual institutions. By methodological choices we mean how the selected indicators have been statistically treated to arrive at the composite measure.

International statistical organisations have made progress in establishing good practice in the construction of composite indicators and ranking systems (OECD, 2008) and practitioners strongly recommend undertaking a robustness analysis before making the results of a model-based analysis public (Kennedy, 2007; Saisana et al., 2005; Saltelli et al., 2008). However, although the limitations of the frameworks underlying the ARWU and THES ranking have been extensively discussed, to the best of our knowledge there has been very little examination of the statistical methodologies employed to arrive at these indices (two recent exceptions are Dehon et al., 2009; Billaut et al., 2009).

We make use of these tools to analyse the methodological robustness of ARWU and THES. More precisely, we test whether those two rankings, within the limits of their frameworks, are statistically robust enough to:

(1) Compare the ranks of individual universities;
(2) Compare and contrast the performance of national university systems;
(3) Inform about macro regional differences between North America, Europe and Asia in terms of the number of top performing universities.

We adopt a multi-modelling approach for this study (Saisana, 2008; Saisana and Munda, 2008), whereby different combinations of aggregation and weighting are taken as different models within the same theoretical (and normative) framework. Applying these models to the underlying indicators also allows us to produce a median ranking for both the ARWU and the THES ranking which is more dependent on the framework of indicators than on the methodological assumptions. With this new measure we can also contrast country or macro-regional performance in terms of the number of top performing universities.

While the robustness analysis undertaken in this study makes it possible to quantify the degree to which uncertainty in the rankings results from uncertainties in the methodological assumptions, it does not inform on the overall relevance of the two international university rankings. This rests rather on the reliability of all the earlier steps in the construction of the indices, and in particular on the development of a consistent and coherent theoretical framework. The two frameworks largely reflect the normative assumptions of their developers – acting within the constraints of internationally comparable data – and are not the result of a consensus in the academic community. This limitation is an important context for the findings of the robustness assessment.

The paper is organised as follows. Section 2 describes the main features of the 2008 ARWU and THES ranking, including the underlying indicators and the methodological assumptions on weighting and aggregation. It also presents an overview of the main criticisms raised in the literature about the choice of indicators for the two rankings. Section 3 briefly looks into the two frameworks of indicators by means of correlation analysis. Section 4 presents a robustness assessment of the two international rankings, which involves the simultaneous activation of various sources of uncertainty, e.g., modifying weighting and aggregation rules. Section 5 discusses the reliability of the THES and ARWU rankings and their fitness as a guide for higher education policy and concludes with some recommendations.

2. Two world university rankings. Features, criticism and policy impact

2.1. ARWU university ranking – main features

The Academic Ranking of World Universities has been published annually by the Shanghai Jiao Tong University since 2003.[4] The sample includes every university in the world which can boast Nobel or Fields laureates among its alumni or staff, has highly cited researchers or articles published in Nature and Science. Universities with significant numbers of articles indexed in the Science Citation Index Expanded and Social Sciences Citation Index are also added to the sample. In total, more than 2000 institutions are considered, 1000 are ranked and the ranking of the top 500 is published. The top 100 universities are assigned a single rank. The remaining 400 universities are assigned to ranking bins: 101–151, 152–200, 201–302, 303–401, 402–503.

The 2008 ARWU is based on four criteria: (1) quality of education, (2) quality of faculty, (3) research output and (4) academic performance. Six indicators are employed (Table 1). The criterion of quality of education is described by a single indicator: the number of alumni of an institution having won Nobel Prizes or Fields Medals, with various sub-weights according to when the alumni obtained their degrees. This indicator is assigned a 10% weight. The

[4] See http://www.arwu.org/ranking.htm for additional information. In 2008 the ARWU was also published by broad subject area in five categories, including Natural Sciences and Mathematics; Engineering, Technology and Computer Sciences; Life and Agriculture Science; Social Sciences; Clinical Medicine and Pharmacy.

Table 1
Shanghai Jiao Tong University Rankings (ARWU), 2008.

Criteria	Indicator	Weight
Quality of education	Alumni of an institution winning Nobel Prizes and Fields Medals	10%
Quality of faculty	Staff of an institution winning Nobel Prizes and Fields Medals	20%
	Highly cited researchers in 21 broad subject categories	20%
Research output	Articles published in Nature and Science	20%
	Articles in Science Citation Index Expanded, Social Sciences Citation Index	20%
Academic performance	Academic performance with respect to the size of an institution	10%

Table 2
Times Higher Education Supplement Rankings (THES), 2008.

Criteria	Indicator	Weight
Research quality	Academic opinion: peer review, 6354 academics	40%
	Citations per faculty: total citation/full time equivalent faculty	20%
Graduate employability	Recruiter review: employers' opinion, 2339 recruiters	10%
International outlook	International faculty: percentage of full-time international staff	5%
	International students: percentage of full-time international students	5%
Teaching quality	Student faculty: full-time equivalent faculty/student ratio	20%

quality of faculty is captured by two indicators: (a) the number of Nobel or Fields laureates among the staff of an institution, and (b) the number of highly cited researchers in 21 broad subject categories over the period 1981–1999.[5,6] Each of these indicators is assigned a 20% weight. The research output is quantified by three indicators: the number of articles published in (a) Nature or Science over the period 2003–2007, (b) Science Citation Index Expanded and (c) Social Sciences Citation Index for 2007. Research output is awarded 40% of the total weight. Finally, the fourth criterion of academic performance is the weighted scores of the above five indicators divided by the number of full-time equivalent academic staff. This indicator, which is the only one to be size-adjusted, can be taken as a measure of efficiency and is assigned the remaining 10% weight. The raw data are normalised by assigning to the best performing institution a score of 100 and all other institutions receiving percentage points away from the leader. The ARWU score is a weighted average of the six normalised indicators, which is finally re-scaled to a maximum 100.

2.2. THES university ranking

The THES World University Ranking has been published annually by the Times Higher Education Supplement since 2004.[7] It initially considered the top 500 universities in terms of research impact. Single-faculty institutions or postgraduate-only institutions were later removed from the sample.[8] The THES currently publishes the ranking of the top 200 institutions.

In the THES World University Ranking, the opinion of scientists and international employers plays a crucial role. In 2008 the THES ranking was based on four criteria: (1) research quality, (2) graduate employability, (3) international orientation and (4) teaching quality (Table 2). Research quality is measured by (a) the opinion of a sample of academics and (b) the number of citations divided by full-time equivalent faculty.[9] For the indicator on academic opinion, 6354 respondents are asked to identify both their subject area of expertise and their regional knowledge.[10] They have then to name up to 30 institutions in their region(s) which they consider to be the best in the relevant field of expertise. This peer review indicator counts for 40% of the total weight. In addition, a 20% weight is assigned to the number of papers published and citations received by research staff over the period 2003–2007.[11] The criterion of graduate employability (10% weight) is based on a single indicator also derived from a survey: a sample of 2339 employers of relevant national or international status is questioned as to the universities from which they would prefer to recruit graduates.[12] The international orientation of the institution is captured by two indicators: first, the percentage of overseas staff at the university, and second, the percentage of overseas students. Each indicator receives a 5% weight. Finally, teaching quality is described by a single indicator, i.e. the ratio between the full-time equivalent faculty and the number of students enrolled at the university. This is assigned the remaining 20% weight.

For each indicator in the THES ranking a z-score is calculated by subtracting the indicator mean from the raw value and then dividing by the standard deviation. The standardised indicator scores are then scaled against a score of 100 for the best performing institution. The final THES score is the weighted average of the six normalised indicators, which is finally re-scaled to a maximum 100.

2.3. Good or bad criteria: literature overview

While both the ARWU and the THES ranking are clear about their normative assumptions, and hence do not expose themselves to the critiques of non-transparency at times directed at composite indicators (see Stiglitz et al., 2009, p. 65), both rankings have been extensively criticised, as discussed in the comprehensive study of Taylor and Braddock (2007). Both rankings rely heavily on bibliometric indicators, which, as noted by Abramo et al. (2009), van Raan (2007), Zitt and Filliatreau (2007), tend to be biased towards English-speaking and hard-science oriented institutions. Journal coverage by Scopus or Thomson-ISI is still not satisfactory for social sciences and humanities, and publishing in peer-reviewed jour-

[5] 'Staff' is defined as those who work at an institution at the time of winning the prize.
[6] See http://www.isihighlycited.com for additional information on the 21 broad subject categories.
[7] The THES ranking also publishes a faculty ranking in the following areas: science; technology; biomedicine; arts and humanities; social sciences.
[8] It also implies that non-university higher education institutions are not taken into account.
[9] The sample size of active academics increases every year. In 2008 it included respondents from 2008, 2007 and 2006. Only the most recent response is taken into account if the respondent has filled in the questionnaire more than once. See http://www.thes.co.uk/worldrankings/ for additional information.
[10] QS Quacquarelli Symonds, an independent consultancy agency, is commissioned with assembling a sample of field-specific experts.
[11] Scopus is the citation data supplier since 2007.
[12] The number of respondents for 2008 also includes respondents from 2007 to 2006 who have not updated their responses. The same consultancy used for the experts (QS) is used for the recruiters.

nals is not the only accepted practice in either of these scientific domains. Finally, citation habits in the various scientific disciplines vary greatly, with a bias in favour of hard sciences. Notwithstanding these criticisms of bibliometric indicators, experts still favour objective indicators of research output over subjective measures based on peer opinion, and so tend to prefer the ARWU over the THES.[13]

The two major objections to the ARWU ranking can be summarised as follows:

- In the ARWU only the research dimension of universities is taken into account, although the relationship between research performance and teaching quality is not straightforward. Given that a great proportion of students do not actually pursue an academic career, the ranking – which ignores dimensions such as employability – might be of little use to the university's "customers". Even if research-based indicators could be accepted as good proxies of overall performance, the use of rare and potentially lagged achievements such as Nobel Prizes (which are, furthermore, only awarded in a limited number of fields) is questionable. Finally, the ARWU's bibliometric approach somehow undervalues the importance of social sciences and humanities, as discussed above.
- Five of the six indicators (representing 90% of the total weight) are size-dependent indicators and only one, academic performance, is normalised by size (Zitt and Filliatreau, 2007; Williams, 2007). This strongly favours – *ceteris paribus* – large institutions and does not give information on the real productivity of the staff of the institution. In 2008, for example, the University of Basel ranks 87th overall but 34th on the academic performance indicator. By contrast, Johns Hopkins University is in the top 20 overall but drops by more than 100 positions on the academic performance indicator. Almost half of the universities in the top 100 shift more than 20 positions on the academic performance indicator compared to the overall rank. The most extreme case is the University of Boston, which moves from 83rd to 294th position after correcting for the impact of size. This simple example shows that the choice of size-dependent versus size-adjusted indicators can make an enormous difference.

Criticism of the THES ranking is focussed on the use of expert-based indicators (50% of the total weight) and the instability of the rankings arising from periodic changes in methodology over the six editions of the ranking. In particular:

- The lack of transparency surrounding the process of selecting the experts has been a cause of concern.[14] Additionally, peer review indicators generally measure the historical reputation of a university rather than current research performance (van Raan, 2007; Taylor and Braddock, 2007). As we show below, expert-based indicators and citation-based indicators exhibit low degrees of correlation.
- The annual changes in the THES methodology have been significant: (a) the sample sizes of the academic and recruiter polls have increased over the years; (b) the indicator on citations, which was based on the previous 10 years' citations in the first two editions, is calculated on the previous 5 years in the latest editions; (c) Scopus has replaced Thomson Scientific as data supplier for citations in the last two editions. Finally, while in the first three editions each institution's score was calculated as a percentage of the best performance, since 2007 the indicators have first been standardised and then converted to the 0–100 scale. Even though the THES team argues that these changes were necessary to improve the quality of the ranking, it is difficult for users to disentangle time variations in the performance of universities from changes which are the result of a statistical artefact. For example, Washington University in St. Louis rises in the overall ranking by more than 100 positions between 2007 and 2008. It is not clear if this shift is the consequence of an improvement in the university's performance or the result of statistical changes implemented in the 2008 THES ranking.

2.4. 2008 ARWU and THES rankings: top universities and cross-country comparisons

A positive feature of the 2008 ARWU and the THES ranking is that seven of the top 10 universities are common to both: Harvard, Cambridge, Caltech, MIT and Columbia, Chicago and Oxford. In other words, the two rankings identify similar world-class universities, despite the diversity of indicators used. However, from the middle to the lower end of the ranking, we observe much greater variation. For instance, McGill is in the top 20 in the THES ranking but below 50 in the ARWU. Similarly, the London School of Economics is 66th in THES but in the 201–302 range in the ARWU.

According to the ARWU and THES rankings, 58 and 42 of the top 100 universities, respectively, are located in North America, and only 34 and 36, respectively, in European countries. The US performs particularly well in terms of institutions at the very top, occupying 17 of the ARWU's top 20 positions and 12 of the THES's top 20. Significant imbalances exist across Europe. While 19 of the THES's top 100 universities are situated in UK, only 2 and 3 universities respectively are located in France and Germany.

The uncertainty analysis carried out in Section 4 will allow us to assess the robustness of the rank assigned to each university as well as of the number of top universities at national or regional level.

2.5. Policy impact

In general, global rankings have aroused debate and elicited two main types of policy response at both EU and national level. The first type of response aims to improve the position of national or regional institutions with respect to the existing rankings; the second is to devise new ways to assess quality. For example, the French government has launched a plan to create 10 centres of excellence in higher education which will regroup several higher education institutions and research organisations so as to consolidate and extend the research capacity of French institutions. The French minister of Education has given herself until 2012 to put at least 10 French universities into the top 100 (in the ARWU), while the French president has put French standing in these international rankings at the forefront of the policy debate (Le Monde, 2008). Similarly, with its "Excellence Initiative" Germany wants to strengthen cutting-edge research and make German research more visible on the world stage. A recent OECD study (Hazelkorn, 2007) shows that university leaders' concern about ranking systems has consequences on the strategic and operational decisions they take to improve their institutions' research performance. The European Commission has charged the CHERPA network (Consortium for Higher Education and Research Performance Assessment) to design and test "a new multi-dimensional" ranking system which would constitute an alternative to and overcome the limits of the ARWU and THES rankings.

3. Correlation analysis of the THES and ARWU frameworks

The ARWU and the THES ranking have different objectives. The original purpose of the ARWU was to measure the gap between the

[13] See appendix, Tables A1 and B1, for additional information.

[14] The information provided in the peer review survey and its associated questionnaire is much more detailed in 2008.

Table 3
Pearson correlation coefficients in the THES and ARWU frameworks.

ARWU framework ($n = 503$)	Alumni winning Nobel	Staff winning Nobel	Highly cited researchers	Articles in Nature and Science	Articles in Science and Social CI	Academic performance – size
Staff winning Nobel prizes	0.76					
Highly cited researchers	0.61	0.66				
Articles in Nature and Science	0.68	0.72	0.87			
Articles in Science and Social CI	0.52	0.48	0.68	0.71		
Academic performance – size	0.67	0.72	0.73	0.78	0.59	
ARWU score	0.80	0.85	0.90	0.93	0.79	0.84

THES framework ($n = 400$)	Academic review	Recruiter review	Teacher to student ratio	Citations per faculty	International staff	International students
Recruiter review	0.61					
Teacher/student ratio	0.09*	0.17				
Citations per faculty	0.45	0.13	0.08*			
International staff	0.17	0.34	0.01*	0.04*		
International students	0.22	0.39	0.12	0.10	0.64	
THES score	0.88	0.67	0.43	0.61	0.31	0.40

* Coefficient non-significant ($p \gg 0.05$).

top Chinese universities and 'world-class' universities, particularly in terms of academic or research performance. The THES aimed to compare universities worldwide while abstaining from the use of 'elitist' indicators such as Nobel prizewinners or articles in Nature and Science. The two rankings thus reflect two distinct normative frameworks and narratives.

Before presenting the robustness assessment of the ARWU and THES rankings, we examine briefly the correlation between the underlying indicators themselves and between these and the overall index. Overall, the correlation between the ARWU scores and its six underlying indicators is stronger than in the THES (Table 3) suggesting that in the ARWU framework there is more overlap of information than in the THES.

The ARWU scores have an almost perfect correlation with the number of articles published in Nature and Science ($r = 0.93$). Also very high is the correlation between the ARWU scores and the alumni or staff winning Nobel Prizes and Fields Medals, the number of highly cited researchers, and the academic performance normalised by the size of the institution ($r \geq 0.84$). The correlation between the ARWU scores and the articles in Science Citation Index Expanded and the Social Sciences Citation Index is also high ($r = 0.79$). Relationships among the six ARWU indicators are all positive and significant, and are generally higher than in the THES, implying a greater degree of overlap. The most pronounced correlation is between the number of highly cited researchers and the number of articles published in Nature and Science ($r = 0.88$). There are also strong correlations between the alumni or staff winning Nobel Prizes and Fields Medals, the number of articles published in Nature and Science and academic performance with respect to the size of an institution. This is not surprising, since these indicators capture research quality and are biased towards hard sciences.

All correlation coefficients between the overall THES score and the six underlying indicators are positive and significant (Table 3). The indicators most closely associated with THES are those related to expert opinion: academic review ($r = 0.81$) and recruiter review ($r = 0.71$). This is partly due to the 50% weight attached to the two indicators together. The built-in bias, which is acknowledged by the THES publishers, derives from the fact that large old universities have an established reputation. As a consequence, these universities receive a higher academic review score. Similarly, recruiters' responses, which come mainly from human resources departments, are very predictable, since recruiters aspire to hire graduates from a small selection of universities. The THES scores are less correlated with the citations per faculty ($r = 0.34$). Relationships among the six indicators included in THES are generally low: four pairs of indicators are not significantly correlated and the average correlation of the remaining indicators is only 0.30. This implies that there is limited overlap in what is being measured and that the six indicators account for different aspects of university performance. The most pronounced correlation is between the two expert-derived indicators, namely the academic review and the recruiter review scores ($r = 0.57$). This is not surprising given that the two review-based indicators are both measures of university reputation. However, reputation and current university excellence, as measured by citations, are not necessarily related to each other. In fact, the correlation between citations per faculty and academic review or recruiter review is either fair or low ($r \leq 0.45$).

4. Uncertainty analysis

Notwithstanding recent attempts to establish good practice in the construction of composite indicators (OECD, 2008), "*there is no recipe for building composite indicators that is at the same time universally applicable and sufficiently detailed*" (Cherchye et al., 2007). Booysen (2002, p.131) summarises the debate on composite indicators by noting that "*not one single element of the methodology of composite indexing is above criticism*". This may be due in part to the ambiguous role of composite indicators in both analysis and advocacy (Saltelli, 2007). As the boundaries between the two functions are often blurred, controversy may be unavoidable when discussing these measures.

Enserink (2007) inquires "*Who ranks the university rankers*"? Although some (e.g., Taylor and Braddock, 2007) challenge the indicators employed in the development of the ARWU, a serious reflection on the ranking methodology is still lacking. We show below how uncertainty analysis (UA) can contribute to such a reflection. UA involves assessing the impact of alternative models on the rankings. Each model is in effect a different composite indicator, in which the normalisation scheme, the choice of weights and the aggregation methods have been varied simultaneously within a plausible range. This approach respects the fact that the scores or ranks associated with composite indicators are generally not calculated under conditions of certainty, even if they are frequently presented as such (Saisana et al., 2005).

Uncertainty (or robustness) analysis, as described in OECD (2008),[15] has already been used in the assessment of several com-

[15] This kind of robustness analysis has also been described in the literature as sensitivity analysis (Leamer, 1990). In Saltelli et al. (2008) uncertainty analysis is defined as the study of uncertainty in the inference, while sensitivity analysis is the study of how the uncertainty in the inference can be apportioned to the uncertainty

posite indicators, such as the Composite Learning Index (Saisana, 2008), the Environmental Performance Index (Saisana and Saltelli, 2010), the Alcohol Policy Index (Brand et al., 2007) and the Index of African Governance (Saisana et al., 2009). All of these analyses were performed by the authors together with the developers *ex ante* the publication of their respective indices. We follow here a similar methodology, albeit *ex post*.

4.1. Multi-modelling approach

A multi-modelling approach is applied in the present work for the purpose of robustness analysis. It involves exploring, via a saturated sampling, plausible combinations of three main assumptions needed to build the index: (a) the weights attached to the indicators; (b) the aggregation rule; and (c) the number of indicators included. We carried out a total of 70 simulations for both the ARWU and the THES. In the present exercise the ARWU and THES data are assumed to be error-free. Furthermore, as the raw data for the indicators are not available (data for both the ARWU and the THES are only available after normalisation), our analysis does not address uncertainty in the data themselves and in their normalisation.[16] The uncertainty propagation features in our analysis are described next.

(a) *Assumption on the weighting scheme*: In the ARWU and THES rankings there are no explicit justifications for the selected weights. We tested three alternative and legitimate weighting schemes: factor analysis derived weights (upon factor rotation and squaring of the factor loadings, as described in Nicoletti et al., 2000); equal weighting; and "university-specific weighting". The last alternative, also known as Data Envelopment Analysis, involves choosing the set of weights for each university that maximises that university's performance relative to all other universities.[17] Practitioners use this approach to counter stakeholder objections that a given weighting scheme is not fair because it does not reflect a certain stakeholder's priorities (Cherchye et al., 2008). In fact, in the US and Canada several universities and colleges have refused to participate in ranking exercises on the grounds that rankings did not reflect institution-specific priorities (Enserink, 2007).

(b) *Assumption on the aggregation rule*: The ARWU and the THES rankings are built using a weighted arithmetic average (a linear aggregation rule) of the six indicators (Eq. (1)). Decision-theory practitioners have challenged aggregations based on additive models because of inherent theoretical inconsistencies (Munda, 2008) and because of the fully compensatory nature of linear aggregation, in which an *a%* increase in one indicator can offset a *b%* decrease in another indicator, where "*b*/*a*" depends on the ratio of the weights of the two indicators.[18] Besides the original developers' choice (linear aggregation), we added two alternative approaches to aggregation: a geometric weighted average (Eq. (2)) and a multi-criteria method. In the case of the geometric averaging, we scaled the normalised data onto a 1–100 range to allow for the proper use of the geometric aggregation. The multi-criteria literature offers several methods (Kemeny, 1959; Munda, 2008; Young and Levenglick, 1978). We selected the Borda adjusted score method, suggested by Brand et al. (2007) (Eq. (3)), for two reasons: first, it can deal with a large number of entities (e.g., universities), unlike the other currently available Condorcet-type methods (Condorcet, 1785); and second, it can deal with ties in indicator scores and also incorporate information on weights, unlike the classical Borda method (Borda, 1784). Both these alternative approaches are less compensatory than the linear aggregation.

Thus the three methods employed in our multi-modelling analysis are:

Weighted Arithmetic Average score:

$$y_j = \sum_{i=1}^{n} w_i \cdot x_{ij} \tag{1}$$

Weighted Geometric Average score:

$$y_j = \prod_{i=1}^{n} x_{ij}^{w_i} \tag{2}$$

Borda adjusted score:

$$y_j = \sum_{i=1}^{n} \left(m_{ij} + \frac{k_{ij}}{2} \right) \cdot w_i \tag{3}$$

where y_j: composite indicator score for university j, w_i: weight attached to indicator i, x_{ij}: normalised score for university j on indicator i, m_{ij}: number of universities that perform worse than university j relative to indicator i, k_{ij}: number of universities with equivalent performance to university j relative to indicator i.

(c) *Assumption on the indicators*: We have either retained all six indicators or in some simulations excluded one at a time. This statistical procedure is a tool to test the robustness of inference and should not be understood as a disturbance of either the ARWU or THES framework, but rather as a cross-validation exercise. In fact, we have verified that by eliminating one indicator at a time – while using the original weighting scheme and the linear aggregation method employed by the rankers, and doing this for all the indicators in the framework, the "median" ranking is very similar to the original.

4.2. Uncertainty analysis results – universities

Based on the multi-modelling approach described above, we have generated a total of 70 models for the ARWU and THES rankings. Tables 4 and 5 report the median rank and its 99% confidence interval for the top 50 universities in both rankings. Confidence intervals were estimated using bootstrap procedures (1000 samples taken with replacement – see Efron, 1979). We interpret the 'median' performance across all 70 models as a summary measure of methodological uncertainty and the confidence interval as the volatility of the ranking which can be attributed to a change to the underlying methodology.

For the ARWU, it is beyond doubt that Harvard, Stanford, Berkley, Cambridge and MIT are in the top five, both in the original and in the simulated rank (99% confidence interval for the median rank). However, as soon as we move away from the top 10, the methodological assumptions have a strong effect on the final rank:

in the input assumptions. Sensitivity analysis in this more restricted sense is not tackled in the present paper.

[16] Data, data editing and normalization assumptions can in principle be treated with the same approach (OECD, 2008).

[17] To avoid extreme scenarios in which a large number of universities score 1 as a result of assigning zero weight to many indicators, we attached restrictions to the indicators' shares (product of indicators and weights), as often advised in the recent literature (Cherchye et al., 2008; Wong and Beasley, 1990). The indicators' shares were allowed to range from 10% to 30%.

[18] In certain contexts a compensatory logic leads to aberrant results. An index for the performance of cars could give a high score to a Ferrari with a flat tire as the poor wheel's efficiency would be compensated for by a brilliant engine. Thus in systems where performance depends on the functioning of several dimensions, no dimension should perform below a given threshold for the index to be positive. Purely compensatory rules should preferably be avoided in these contexts.

Table 4
Original and simulated median rank (with confidence interval) for the ARWU top 50 universities.

Country	University	ARWU rank	Median rank	99% confidence interval for the median rank
USA	Harvard U.	1	1	[1,1]
USA	Stanford U.	2	3	[2,4]
USA	U. California – Berkeley	3	3	[3,3]
UK	U. Cambridge	4	4	[2,5]
USA	Massachusetts Inst Tech (MIT)	5	5	[4,5]
USA	California Inst Tech	6	6	[6,7]
USA	Columbia U.	7	7	[6,7]
USA	Princeton U.	8	9	[8,11]
USA	U. Chicago	9	10	[9,12]
UK	U. Oxford	10	9	[9,10]
USA	Yale U.	11	11	[9,11]
USA	Cornell U.	12	12	[11,12]
USA	U. California – Los Angeles	13	13	[13,13]
USA	U. California – San Diego	14	16	[15,18]
USA	U. Pennsylvania	15	14	[13,15]
USA	U. Washington – Seattle	16	18	[17,20]
USA	U. Wisconsin – Madison	17	16	[16,17]
USA	U. California – San Francisco	18	35	[24,49]
Japan	Tokyo U.	19	23	[21,25]
USA	Johns Hopkins U.	20	19	[18,22]
USA	U. Michigan – Ann Arbor (*)	21	47	[27,64]
UK	U. Coll London	22	20	[19,22]
Japan	Kyoto U.	23	21	[20,23]
Switzerland	Swiss Fed Inst Tech – Zurich	24	19	[17,21]
Canada	U. Toronto	24	24	[22,25]
USA	U. Illinois – Urbana Champaign	26	23	[22,25]
UK	Imperial Coll London	27	24	[23,26]
USA	U. Minnesota – Twin Cities	28	30	[28,32]
USA	Washington U. – St. Louis	29	26	[25,28]
USA	Northwestern U.	30	30	[29,31]
USA	New York U.	31	31	[28,32]
USA	Duke U. (*)	32	57	[35,80]
USA	Rockefeller U.	32	32	[29,47]
USA	U. Colorado – Boulder	34	35	[32,38]
Canada	U. British Columbia	35	33	[32,35]
USA	U. California – Santa Barbara	36	60	[46,74]
USA	U. Maryland – Coll Park	37	36	[34,38]
USA	U. North Carolina – Chapel Hill	38	47	[43,51]
USA	U. Texas – Austin	39	45	[42,49]
UK	U. Manchester	40	39	[37,42]
USA	U. Texas Southwestern Med Center	41	38	[35,43]
USA	Pennsylvania State U. – U. Park (*)	42	76	[61,95]
France	U. Paris 06	42	42	[39,47]
USA	Vanderbilt U.	42	38	[36,42]
Denmark	U. Copenhagen	45	37	[35,39]
USA	U. California – Irvine	46	65	[54,78]
Netherlands	U. Utrecht	47	42	[39,45]
USA	U. California – Davis (*)	48	98	[72,116]
France	U. Paris 11	49	49	[43,64]
USA	U. Southern California	50	73	[61,84]

Notes: Universities for which the median rank is highly uncertain (confidence interval >30 positions) are marked with an asterisk.

the University of California–San Francisco ranks 18th in the ARWU but this rank falls outside the confidence interval of the median rank [24,49] in our uncertainty analysis. Similarly, the University of Michigan is ranked 21st in the original ranking while the median rank is 47 and its associated confidence interval spans 27–64.

In the case of THES, the impact of the uncertainties on the ranking is more pronounced. MIT ranks 9th in THES, while the confidence interval of the median rank is [12,24]. Very high volatility is also evident for most of the universities ranked between the 10th and 20th positions, such as Duke University, John Hopkins or Cornell. The ranking of universities is evidently highly volatile after the 30th position for both the ARWU and THES rankings.

Figs. 1 and 2 present the median rank and its 99% confidence interval for all the universities ranked by the ARWU and THES. In addition the name of universities whose original rank does not fall within the interval is reported. The developers might find this information useful. The plots can either be used directly as measures (thus replacing a crisp original score with a median performance) or as part of a robustness analysis. We observe that the ranks of 43 universities in the ARWU's top 100 (e.g., University of Munich, University of Helsinki) fall outside the confidence interval, as do the ranks of 61 universities in the THES's top 100 (e.g., University of Tokyo). For 3 in 10 universities in the ARWU's top 200 and 5 in 10 universities in the THES's top 200, the range for the expected rank (i.e. the confidence interval for the median rank) is greater than 30 positions (across the 70 scenarios).

The main conclusion of this uncertainty analysis is that neither of the two rankings can be used to compare the ranks of individual universities, given that for the majority of universities, the assigned rank is very sensitive to the underlying methodology.

4.3. Uncertainty analysis results – countries

Table 6 shows the number of top universities in selected countries in the original ARWU and THES according to the simulations. Two main comments are in order.

Table 5
Original and simulated median rank (with confidence interval) for the THES top 50 universities.

Country	University	THES rank	Median rank	99% confidence interval for the median rank
USA	HARVARD U.	1	6	[4,6]
USA	YALE U.	2	7	[3,8]
UK	U. of CAMBRIDGE	3	3	[2,3]
UK	U. of OXFORD	4	4	[3,4]
USA	CALIFORNIA Institute of Technology (Calt..	5	6	[5,8]
UK	IMPERIAL College London	6	3	[2,5]
UK	UCL (U. College London)	7	4	[3,5]
USA	U. of CHICAGO	8	8	[8,10]
USA	MASSACHUSETTS Institute of Technology	9	18	[12,24]
USA	COLUMBIA U.	10	22	[12,31]
USA	U. of PENNSYLVANIA	11	11	[10,12]
USA	PRINCETON U.	12	10	[9,12]
USA	DUKE U.	13	30	[17,43]
USA	JOHNS HOPKINS U.	13	32	[18,45]
USA	CORNELL U.	15	30	[15,41]
Australia	AUSTRALIAN National U.	16	11	[10,16]
USA	STANFORD U.	17	35	[23,47]
USA	U. of MICHIGAN	18	26	[20,35]
Japan	U. of TOKYO (*)	19	55	[30,81]
Canada	MCGILL U.	20	23	[21,25]
USA	CARNEGIE MELLON U.	21	23	[21,25]
UK	KING'S College London	22	16	[13,18]
UK	U. of EDINBURGH	23	17	[15,20]
Switzerland	ETH Zurich (Swiss Federal Institute of T..	24	17	[14,24]
Japan	KYOTO U. (*)	25	71	[39,114]
Hong Kong	U. of HONG KONG	26	18	[15,25]
USA	BROWN U.	27	37	[33,49]
France	École Normale Supérieure, PARIS (*)	28	59	[40,75]
UK	U. of MANCHESTER	29	23	[20,27]
Singapore	National U. of SINGAPORE(NUS)	30	24	[20,29]
USA	U. of CALIFORNIA, Los Angeles (U.. (*)	30	81	[52,124]
UK	U. of BRISTOL	32	25	[21,29]
USA	NORTHWESTERN U. (*)	33	55	[40,72]
France	ÉCOLE POLYTECHNIQUE	34	28	[26,34]
Canada	U. of BRITISH COLUMBIA(*)	34	59	[39,75]
USA	U. of California, BERKELEY (*)	36	63	[42,76]
Australia	The U. of SYDNEY	37	29	[25,36]
Australia	The U. of MELBOURNE	38	37	[35,41]
Hong Kong	HONG KONG U. of S & T.	39	21	[18,32]
USA	NEW YORK U. (NYU) (*)	40	73	[53,94]
Canada	U. of TORONTO (*)	41	73	[54,93]
Hong Kong	The CHINESE U. of Hong Kong	42	29	[24,37]
Australia	U. of QUEENSLAND	43	32	[29,40]
Japan	OSAKA U. (*)	44	118	[69,162]
Australia	U. of NEW SOUTH WALES	45	43	[38,47]
USA	BOSTON U. (*)	46	88	[54,103]
Australia	MONASH U.	47	39	[31,46]
Denmark	U. of COPENHAGEN	48	60	[51,74]
Ireland	TRINITY College Dublin	49	38	[34,49]
Switzerland	Ecole Polytech. Fédérale de LAUSANNE..	50	28	[23,39]

Notes: Universities for which the median rank is highly uncertain (confidence interval >30 positions) are marked with an asterisk.

Table 6
Number of top universities in France or Germany in the original THES or ARWU and according to simulations.

	ARWU ranking			THES ranking		
	Original	Simulated median[1]	99% conf. int. for the median[2]	Original	Simulated median[1]	99% conf. int. for the median[2]
Top10 France	0	0	[0,0]	0	0	[0,0]
Germany	0	0	[0,0]	0	0	[0,0]
UK	2	2	[2,2]	4	4	[4,4]
Top20 France	0	0	[0,0]	0	0	[0,0]
Germany	0	0	[0,0]	0	0	[0,0]
UK	2	3	[2,3]	5	6	[6,7]
Top50 France	2	2	[1,2]	2	1	[1,1]
Germany	0	2	[2,2]	0	0	[0,0]
UK	5	6	[6,6]	8	14	[12,15]
Top100 France	3	4	[3,5]	2	2	[2,2]
Germany	6	8	[7,8]	3	2	[2,3]
UK	11	12	[12,12]	17	26	[22,29]
Top200 France	7	8	[7,9]	4	5	[5,6]
Germany	14	15	[15,16]	11	11	[11,12]
UK	22	23	[21,24]	29	38	[35,40]

Notes: [(1)]Simulated median across 70 scenarios; [(2)]confidence interval for the median calculated across 70 scenarios replicated with 1000 bootstrap samples.

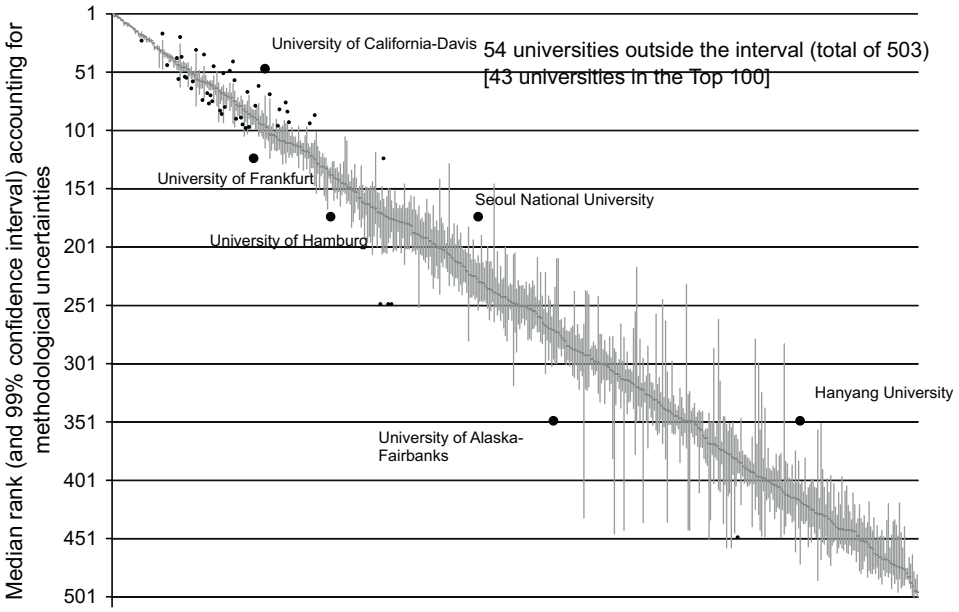

Fig. 1. Simulated median and its 99% confidence interval (across 70 models) for university ranks assessed using the ARWU framework. University ranks that fall outside the interval are marked in black. *Note:* The dots relate a university's ARWU rank to the median rank calculated over the set of scenarios (indicators, weighting scheme, aggregation rule) generated in our uncertainty analysis. Results for all 503 universities are depicted. Uncertainty intervals (vertical bars) for the median rank are estimated using 1000 bootstrap samples. The ARWU rank is precise for the top 100 universities and in bins of 101–151, 152–200, 201–302, 303–401, 402–503. To avoid superimposing universities beyond rank 100, universities per bin have been spread out and assigned in the middle of the respective bin. For example, the university of Hamburg in Germany was originally ranked by the ARWU developers in the 152–200 range, but the uncertainty analysis results reveal that Hamburg University is much better positioned, somewhere between 133 and 146 on average.

First, as far as the ARWU is concerned, the number of German, French and British universities in the top 10 and in the top 20–200 of the original ranking falls inside (or is very close to) the confidence interval of the median number. The number of top British universities in the top 200 amounts to 22 in the original ranking while the median is 23. Similarly, 14 and 15 German universities are in the top 200 according to respectively the original ranking and our simulations. By comparison, the THES ranking is far less statistically robust in the comparison of the performance of national university systems. For example, the number of top 100 British universities is 17 in the original ranking while the confidence interval of the median number is [22,29]. As discussed below this is not a flaw of the THES ranking per se.

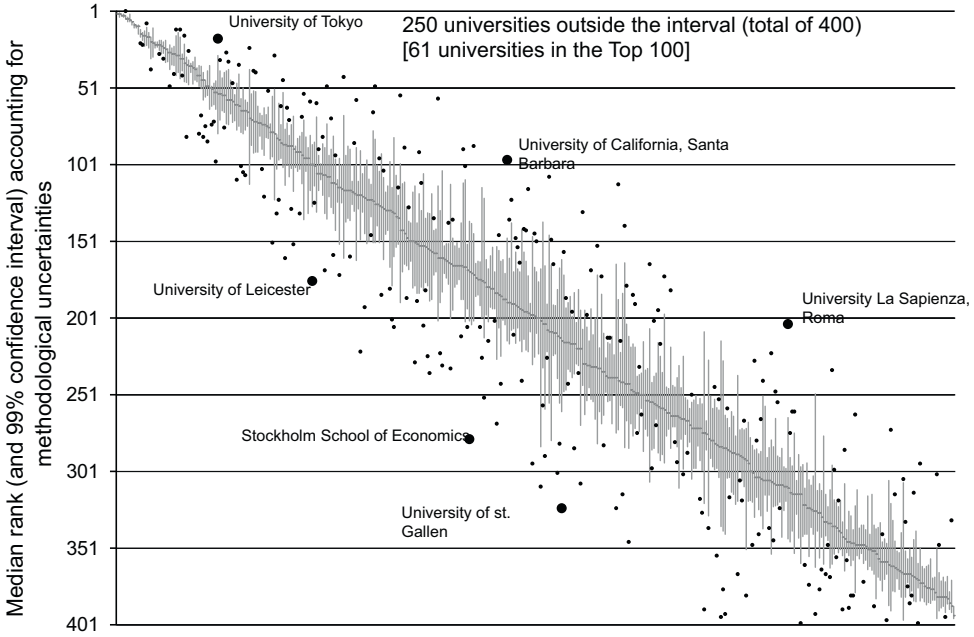

Fig. 2. Simulated median and its 99% confidence interval (across 70 models) for the university ranks assessed using the THES framework. University ranks that fall outside the interval are marked in black. *Note:* The dots relate a university's THES rank to the median rank calculated over the set of scenarios (indicators, weighting scheme, aggregation rule) generated in our uncertainty analysis. Results for all 400 universities are depicted. Uncertainty intervals (vertical bars) for the median rank are estimated using 1000 bootstrap samples. For example, the University of Leicester in the UK was originally ranked by the THES developers in 177th position, but the uncertainty analysis results reveal that this university is much better positioned, somewhere between 88 and 129 on average.

Table 7
Number of top universities according to region (Europe, USA, China) in the original THES or ARWU and according to simulations.

	ARWU ranking			THES ranking		
	Original	Simulated median[1]	99% conf. int. for the median[2]	Original	Simulated median[1]	99% conf. int. for the median[2]
Top10 Europe	2	2	[2,2]	4	4	[4,4]
USA	8	8	[8,8]	6	6	[5,6]
China	0	0	[0,0]	0	0	[0,0]
Top20 Europe	2	3	[3,4]	5	7	[7,9]
USA	17	16	[16,16]	13	8	[7,12]
China	0	0	[0,0]	0	1	[0,2]
Top50 Europe	10	14	[13,15]	12	20	[17,21]
USA	36	31	[30,32]	20	16	[14,19]
China	0	0	[0,0]	4	3	[3,3]
Top100 Europe	34	37	[36,40]	36	47	[42,49]
USA	54	49	[48,52]	37	29	[25,33]
China	0	0	[0,0]	5	4	[3,4]
Top200 Europe	79	82	[79,84]	82	93	[86,96]
USA	90	88	[87,89]	59	53	[50,57]
China	1	1	[1,1]	11	10	[9,11]

Notes: (1)Simulated median across 70 scenarios; (2)confidence interval for the median calculated across 70 scenarios replicated with 1000 bootstrap samples.

Second, assuming that the median of our simulations is a good measure to benchmark countries in terms of the number of top performing institutions (since it acknowledges the uncertainty inherent in the methodology), then it is remarkable that the number of British universities in the top 100 is three times as high as the number of French universities in the ARWU, but 12 times as high in the THES. Among the top 100 and 200 universities, the ARWU puts the UK on a level with both France and Germany; however, the UK performs considerably better in the THES[19].

Additionally, at the level of the top 50 universities France and Germany perform equally well in the ARWU, while the number of German universities in the top 100 is twice the French (8 versus 4). The simulated median rank based on the THES shows that there is the same number of top 100 universities in France and Germany (=2).

The advantage of having tested the results against the median performance here is that the conclusions above are solely dependant on the framework of the indicators and not on the methodological choices (weighting or type of aggregation).

4.4. Uncertainty analysis results – regions

Table 7 shows the number of top universities by region (Europe, USA, China) in the original ARWU and THES and according to our multi-modelling analysis.

As far as the top 10 universities are concerned, the inference based on the original rankings is robust. The simulations based on the ARWU identify that there are, on average, 2 European and 8 US universities in the top 10, as suggested by the original ARWU. Similarly, the simulations based on the THES as well as the original ranking identify that there are 4 and 6 European and US universities in the top 10 respectively. These conclusions are framework-dependant but methodology-independent.

However, conclusions on regional performance are less robust when we move down the rankings, and this is particularly evident in the THES. Indeed, European universities in the top 50 and top 100 number 10 and 34 respectively in the original ARWU, while the median numbers are respectively 14 and 31. By comparison, 12 and 36 of the top 50 and top 100 are European universities in the original THES whereas the median figures are 20 and 47.

These results suggest that the ARWU is quite robust when comparing the performance of university institutions between North America, Europe and Asia. By contrast, similar conclusions drawn from the original THES ranking change significantly when the methodology employed to build the ranking is modified.

Simulations based on the ARWU show that there are more US than European universities in the top 10, top 20, top 50 and top 100, although among the top 200 there are roughly the same number of US and European universities. By contrast, simulations based on the THES ranking show that the numbers of European and US universities are roughly the same in the top 50, but that there are clearly more European than US universities among the top 100 and top 200.

Interestingly, although the ARWU was originally developed with a view to measuring the distance of Chinese universities from the rest of the world, no local regional bias was introduced in the development of the ranking. In fact, it is the THES that paints a rosier picture for China – 10 Chinese universities appear in the top 200 in the THES compared to just 1 Chinese university according to the ARWU.

5. Concluding remarks and policy implications

The Academic Ranking of World Universities (ARWU) by Shanghai's Jiao Tong University and the UK's Times Higher Education Supplement (THES) ranking have attracted extensive media attention and stimulated a useful debate on higher education in Europe and worldwide. Higher education institutions provide an array of services and positive externalities to society (broad education, innovation and growth, active citizens, entrepreneurs and administrators, etc.) which call for multi-dimensional measures of effectiveness and/or efficiency. A clear statement of the purpose of any such measures is also needed, as measuring scientific excellence is evidently not the same as measuring graduate employability or innovation potential. The conceptual differences between the ARWU and the THES ranking illustrate this clearly.

In league tables and ranking systems, ranks are often presented as if they had been calculated under conditions of certainty. Media and stakeholders take these measures at face value, as if they were unequivocal, all-purpose yardsticks of quality. To the consumers of composite indicators, the numbers seem crisp and convincing.

"Rankings are here to stay, and it is therefore worth the time and effort to get them right" warns Alan Gilbert, (Nature News, 2007). In this paper we have offered a quality check in the form of a statistical robustness analysis based on the multi-modelling approach, which involved activating different sources of uncer-

[19] The UK has 12 and 23 universities in the top 100 and top 200 respectively in ARWU, which is equal to the number of top performing universities in France and Germany put together. By comparison, THES puts 26 UK universities in the top 100, against 4 universities in France and Germany combined; in the top 200 there are 38 UK universities as opposed to 16 French and German universities combined.

tainty simultaneously, in order first to test the validity of inferences associated with the ARWU and the THES, and second to produce a median ranking (and respective confidence interval) which is more framework- than model-dependent.

Our analysis suggests that:

(1) Apart from the top 10 universities, neither the ARWU nor the THES should be used to compare the performance of individual universities. Equally deceptive can be the interpretation of changes in rank for a given university, as indulged in by the media.[20] The position of the majority of the universities is highly sensitive to the underlying statistical methodology chosen by the rankers. Based on 70 scenarios, our robustness assessment shows that the ranks of 43 universities in the ARWU's top 100 fall outside the confidence interval (e.g., University of Munich, University of Helsinki), as do the ranks of 61 universities in the THES's top 100 (e.g., University of Tokyo). Furthermore, for 3 in 10 universities in the ARWU's top 200 and 5 in 10 universities in the THES's top 200, the range for the median rank is >30 positions, intervals that are too wide to draw any conclusions on precise ranks.
(2) If the median rank derived from the 70 scenarios carried out in this study for both the ARWU and THES constitutes a satisfactory measure to assess the performance of universities and is representative of the plurality of stakeholders' views on how the indicators should be aggregated and weighted, then the uncertainty analysis results lead to certain conclusions which are less dependent of the methodology used than either the ARWU or THES rankings. The median performance of Europe is not as good as that of the US at the level of the top 10 according to the simulations based on the ARWU. The THES paints a more favourable picture for Europe; however, given the bias of the THES ranking towards British universities noted in Saisana and D'Hombres (2008) and Taylor and Braddock (2007), it does seem that the 'excellence alarm' over the incapacity of European institutions to compete at the very top levels may have some basis. Furthermore, roughly 40 European universities are in the top 100 in both rankings ([36,40] in the ARWU and [42,49] in the THES). At the level of the top 200, the two rankings disagree on the relative strength of the two regions; the ARWU puts Europe and the USA on a level (roughly 80–90 top universities in each region), whilst the THES strongly favours European over US universities (93 and 53 respectively). Within Europe, only UK institutions figure in the top 10 or top 20. Among the top 50 or top 100 the UK clearly stands out from other European countries, while no difference can be discerned between France and Germany. Among the top 200, however, Germany clearly outperforms France in both the ARWU and THES. These conclusions are based on the simulated median and are hence more dependant on the framework of indicators (whether the ARWU or THES) than on the methodological choices in the aggregation.

While the uncertainty analysis carried out in this paper tests whether selected inferences associated with the ranking are robust or volatile with respect to variations in the assumptions behind the construction, it does not reveal whether the ARWU or THES are legitimate models of university performance. Further, it is likely that the more ambitious a measure in its attempt to capture a multifaceted phenomenon, the more volatile the ranking; therefore, the higher volatility of the THES ranking does not imply that it is inferior to the ARWU. The statistical robustness analysis is indeed only one element of a comprehensive assessment of any composite indicator. The global relevance of a ranking also depends on the reliability of all the steps in the construction of the CI, from the development of a theoretical framework and the selection of variables to the presentation and dissemination of results to the general public and policy-makers through league tables. This is particularly important in the light of the numerous criticisms directed at the choice of indicators in both rankings. The results of this analysis must therefore be read against the context of these limitations.

The analysis implies that the two rankings reflect the perspectives of their developers and do not necessarily meet the practical needs of students or of higher education policy-makers. If the ARWU and THES are being used by stakeholders to inform their choices, e.g., about where to study or how to reform and improve the higher education system, then the messages they convey may result in sub-optimal decisions. Policy-makers in particular would be better advised to exploit these measures only to raise awareness of certain issues in higher education policy and practice among constituencies in civil society. As discussed in Section 2.5, policy initiatives have been taken specifically to improve a country's general performance in the ARWU scores, which should be a cause for concern.

A robustness assessment such as that which has been discussed in the present work should be employed in acknowledgement of the methodological uncertainties that are intrinsic to the development of a ranking system and to test whether the space of inference of the ranks for the majority of the universities is narrow enough to justify any meaningful classification. Although such a robustness check has already proved useful in the development and validation of several composite indicators, it is still not common practice in the field of research evaluation. Should university rankings be "here to stay", and new more comprehensive multidimensional measures developed, as proposed by the European Commission for the EU, uncertainty analysis of these new constructs would be advisable.

Acknowledgments

The authors wish to thank the members of the Unit of Econometrics and Applied Statistics at the Joint Research Centre in Ispra (Italy) as well as the participants of the 4th Conference of the International Rankings Expert Group (June 14–17, 2010, Astana, Kazakhstan) and of the 7th International Conference on Social Science Methodology (September 1–5, 2008, Naples, Italy) for their comments on the paper. We also thank three anonymous referees for valuable suggestions that helped improve the quality of the paper.

[20] See for instance Le Figaro (2008).

Appendix A.

Table A1
2008 THES indicators: what has been said so far.

Indicators	Weight	Problems	Qualities
I – Peer review	40%	Survey: regional bias and lack of transparency (I and II) 1 – Assessors were asked to assess the relative performance of institutions in their own geographical area 2 – Not clear what questions were asked and who was surveyed Reputation indicator (I) 1 – Depends on past performance	
II – Employer review	10%	Selection bias (II) Outstanding universities are initially well connected and recruit good students: nothing to do with excellence produced within the U	
III – Citation per capita	20%	Quantity without taking quality into account (III) 1 – Measures research quantity without specifically rewarding high research quality 2 – Only 20% of total score while research is one of the main components of U excellence 3 – Bibliometric indicators: biased toward English-language journals and downplay the importance of social sciences and humanities	Per capita: no size bias (III)
IV – Student/teaching ratio	20%	Crude measure of teaching quality (IV)	Proxy for teaching quality (IV) 1 – Gives 20% of total score to one important aspect of U activities 2 – Objectively measurable 3 – Very difficult to find other ways to measure teaching quality
V – International orientation		Not real criterion of U excellence (V)	Proxy for university quality (V)
A – % of overseas students	5%	1 – Correlated with the characteristics of the university city's population (multicultural city)	1 – Capacity to attract foreign staff and students
B – % of overseas staff	5%	2 – Say more about the quality of recruitment methods (e.g., good advertising) than about university excellence	2 – International education

Note: This overview is largely drawn from Taylor and Braddock (2007).

Appendix B.

Table B1
2008 THES indicators: what has been said so far.

Indicators	Weight	Problems	Qualities
I – Nobel Prizes (P) and Fields Medals (M)		Rough measures of teaching and research quality (I-A, B)	Proxy for research and teaching quality (I-A, B, II)
A – won by alumni	10%	1 – Many U have no N or F laureates: no distinction for those U	1 – Reward research quality and not only research quantity
B – won by faculty members	20%	2 – Attributing N and F laureates to teaching quality is not straightforward because of a self-selection bias	2 – Proxy for university ability to attract outstanding researchers
		3 – N and F prizes are time-specific: not representative of current performance	3 – Quality research output: exclude researchers with "soft" academic publications (in particular II)
		4 – Affiliation at the time of prize is problematic if prize-winning work was done before joining the U (I, B) Hard science bias (I-A, B, II, III) N and F prizes possible in only a limited number of fields	
II – Number of highly cited researchers	20%	Only 2 out of the 21 disciplines belong to social sciences	
III – Number of papers published by staff			
A – in Science and Nature	20%	Covers only hard sciences	
B – in wide academic journals	20%		4 – Focus more on research quantity (III B) 5 – Rewards more articles indexed in the social sciences/arts and humanities to compensate for hard science bias
IV – Academic performance/U size	10%	Scant weight (IV)	Adjustment of size bias (IV)

Note: This overview is largely drawn from Taylor and Braddock (2007).

References

Abramo, G., d'Angelo, C.A., Caprasecca, A., 2009. Allocative efficiency in public research funding: can biblometrics help? Research Policy 38, 206–215.

Acemoglu, D., Aghion, P., Zilibotti, F., 2006. Distance to frontier, selection, and economic growth. Journal of the European Economic Association 4 (1), 37–74.

Aghion, P., Dewatripont, M., Hoxby, C., Sapir, A., 2008. Higher aspirations: an agenda for reforming European universities, Bluegel Blueprint Series N.5.

Billaut, J.C., Bouyssou, D., Vincke P., 2009. Should you believe in the Shanghai ranking? http://hal.archives-ouvertes.fr/docs/00/40/39/93/PDF/Shanghai_JCB-DB_PV.pdf.
Booysen, F., 2002. An overview and evaluation of composite indices of development. Social Indicators Research 59 (2), 115–151.
Borda, J.C.de, 1784. Mémoire sur le elections au scrutiny, Histoire de l'Académie Royale des Sciences, Paris.
Brand, D.A., Saisana, M., Rynn, L.A., Pennoni, F., Lowenfels, A.B., 2007. Comparative analysis of alcohol control policies in 30 countries. PLoS Medicine 4 (4), 752–759.
Cherchye, L., Knox Lovell, C.A., Moesen, W., Van Puyenbroeck, T., 2007. One market, one number? A composite indicator of EU internal market dynamics. European Economic Review 51, 749–779.
Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, 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 Operational Research Society 59, 239–251.
Condorcet, M.de, 1785. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la probabilité des voix, De l' Imprimerie Royale, Paris.
Dehon, C., McCathie, A., Verardi, V., 2009. Uncovering excellence in academic rankings: a closer look at the Shanghai ranking. Scientometrics 83 (2), 515–524.
Ederer, P., Schuller, P., Willms, S., 2007. University Systems Ranking: Citizens and Society in the Age of Knowledge. Lisbon Council Policy Brief, Brussels.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7 (1), 1–26.
Enserink, M., 2007. Who ranks the rankers. Science 317, 1026–1028.
European Commission, 2008. Progress towards the Lisbon Objectives in Education and Training, Indicators and Benchmarks, Commission Staff Working Document, Brussels.
Hand, D., 2004. Measurement Theory and Practice: The World through Quantification. Hodder Arnold Publisher.
Hazelkorn, E., 2007. Impact and influence of league tables and ranking systems on higher education decision-making. Higher Education Management and Policy 19 (2), 87–110.
Hendel, D.D., Stolz, I., 2008. A comparative analysis of higher ranking systems in Europe. Tertiary Education and management 14 (3), 173–189.
Jacobs, B., van der Ploeg, F., 2006. Guide to reform of higher education: a European perspective. Economic Policy, CEPR, CES, MSH 21 (47), 535–592.
Kemeny, J., 1959. Mathematics without numbers. Daedalus 88, 571–591.
Kennedy, P., 2007. A Guide to Econometrics, fifth ed. Blackwell.
Leamer, E., 1990. Let's take the con out of econometrics, and sensitivity analysis would help. In: Granger, C. (Ed.), Modelling Economic Series. Clarendon Press, Oxford.
Le Figaro, 2008. La France, mauvaise élève du classement de Shanghaï, 06 August 2008. http://www.lefigaro.fr/.
Le Monde, 2008. Comment sauver l'université française, 14 August 2008, http://www.lemonde.fr.
Le Monde, 2008. La France veut hisser ses universités dans les classements mondiaux, 3 July 2008, http://www.lemonde.fr.
Les Echos, 2008. Classement de Shanghai: les universités françaises à la traine, 7 August 2008, http://www.lesechos.fr/info/france/4759189.htm.
Marginson, S., 2007. Global university rankings: implications in general and for Australia. Journal of Higher Education Policy and Management 29 (2), 131–142.
Munda, G., 2008. Social Multi-criteria Evaluation for a Sustainable Economy. Springer, Berlin.
Nature News, 2007. Academics strike back at spurious rankings. Nature 447 (May), 514–515.
Nicoletti, G., Scarpetta, S., Boylaud, O., 2000. Summary indicators of product market regulation with an extension to employment protection legislation, OECD, Economics department working papers No. 226, ECO/WKP(99)18.
OECD, 2008. Handbook on Constructing Composite Indicators. Methodology and User Guide. OECD, Paris.
Saisana, M., Annoni, P., Nardo, M., 2009. A robust model to measure governance in African countries, Report 23773, European Commission, JRC-IPSC, Italy.
Saisana, M., 2008. The 2007 Composite Learning Index: Robustness Issues and Critical Assessment, Report 23274, European Commission, JRC-IPSC, Italy.
Saisana M., D'Hombres B., 2008. Higher Education Rankings: Robustness Issues and Critical Assessment, Report 23487, European Commission, JRC-IPSC, Italy.
Saisana, M., Munda G., 2008. Knowledge Economy: measures and drivers, Report 23486, European Commission, JRC-IPSC, Italy.
Saisana, M., Saltelli, A., Tarantola, S., 2005. Uncertainty and sensitivity analysis techniques as tools for the analysis and validation of composite indicators. Journal of the Royal Statistical Society A 168 (2), 307–323.
Saisana, M., Saltelli, A., 2010. Uncertainty and Sensitivity Analysis of the 2010 Environmental Performance Index, Report 24269, European Commission, Joint Research Centre, Italy.
Salmi, J., 2009. The Challenge of Establishing World Class Universities. World Bank Publications, Washington, 136 pp.
Saltelli, A., 2007. Composite indicators between analysis and advocacy. Social Indicators Research 81 (1), 65–77.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis. The Primer, John Wiley & Sons, England.
Sapir, A., Aghion, P., Bertola, G., Hellwig, M., Pisani-Ferry, J., Rosati, D., Viñals, J., Wallace, H., 2004. An Agenda for a Growing Europe: The Sapir Report. Oxford University Press.
Stiglitz, J.E., Sen, A., Fitoussi JP, 2009, Report by the Commission on the Measurement of Economic Performance and Social Progress, www.stiglitz-sen-fitoussi.fr.
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.
Usher, A., Savino, M., 2007. A global survey of university ranking and league tables. Higher Education in Europe 32 (1), 5–15.
van Raan, A.F.J., 2005. Fatal Attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics 62 (1), 133–143.
van Raan, A.F.J., 2007. Challenges in the Ranking of Universities. In: Sadlak, J., Cai, L.N. (Eds.), The World-Class University and Ranking: Aiming Beyond Status. UNESCO-CEPES, Bucharest.
Veugelers, R., van der Ploeg, F., 2007. Towards evidence-based reform of European universities. CESifo Economic Studies 54, 99–120.
Williams, R., 2007. Broadening the criteria: lessons from the australian rankings. In: Sadlak, J., Cai, L.N. (Eds.), The World-Class University and Ranking: Aiming Beyond Status. UNESCO-CEPES, Bucharest.
Wong, Y-H.B., Beasley, J.E., 1990. Restricting weight flexibility in data envelopment analysis. Journal of the Operational Research Society 47, 136–150.
Young, H.P., Levenglick, A., 1978. A consistent extension of Condorcet's election principle. SIAM Journal on Applied Mathematics 35, 285–300.
Zitt, M., Filliatreau, G., 2007. Big is (made) beautiful – some comments about Shanghai-ranking of world-class universities. In: Sadlak, J., Cai, L.N. (Eds.), The World-Class University and Ranking: Aiming Beyond Status. UNESCO-CEPES, Bucharest.