Questioning the Shanghai Ranking methodology as a tool for the evaluation of universities: An integrative review

Abstract: This integrative review reports on methodological questions about the Shanghai Ranking as a tool for the evaluation of universities, questions that are extensible to other rankings. The paper presents a list of methodological problems that are the result of both a review of the literature and the authors' knowledge, with the aim of improving and refining the ranking in line with the Berlin Principles. The second section makes proposals and provides explanatory notes for improving the evaluation of university institutions. A final inference is that any educational changes undertaken based on conclusions drawn from an institution's ranking position must be considered highly controversial and questionable.

Key words: Shanghai Ranking; integrative review[1]; evaluation methodology; universities; Berlin Principles.

Basic information about the Shanghai Ranking

The Academic Ranking of World Universities -2017- (hereafter referred to by its acronym ARWU), also known as the Shanghai Ranking, was created and first published in 2003 by the *Center for World-Class Universities* (CWCU) of Shanghai's Jiao Tong University.

[1] The integrative review is a methodology that provides a synthesis of knowledge and the applicability of results of significant studies to practice. Bibliographic research and the authors' personal reflection are needed and rationally combined. Six stages are necessary when preparing a research review: forming the central question, in this case, explicitly about questioning the Shanghai methodology, searching for the relevant literature, data collection, critical examination of the studies included, discussion of results, and writing the report. For reasons of concision, this paper only contains the final stage.

Since 2009, however, it has been compiled and published by the *Shanghai Ranking Consultancy*, an independent organization devoted to research into higher education.

The ARWU was "created with the initial purpose to find the global standing of top universities in China[2]" (quote at http://www.shanghairanking.com/aboutarwu.html). However, "it has attracted great attention from universities, governments and public media worldwide", to such an extent that it has become the most used, reputed and influential ranking of its kind (Margison 2014). Its southern neighbor, the Union of India[3], accepted ARWU's relevance as a clear indication for political planners (Virk 2016). Its acceptance in other countries such as Spain has been rapid and scarcely questioned (Docampo 2013; Docampo and Cram 2017) as if it were an undisputable truth, in what could be a manifestation of uncritical acceptance; a topic well researched in Psychology (Chanowitz and Langer 1981), Medical Innovation (Grimes 1993), Information Systems (Bagozzi 2007) or Political Science (Davidov 2009). The ranking has been designed with a simple black box approach, and its creators consider only a few outputs of universities (sensu stricto) to be globally important. As a result, it receives widespread critical acceptance, but ARWU's off the peg use is not suitable for all countries

The first ARWU (2003) used data from the previous year and included the best 500 universities in the world of the 1200 it had considered. For students and their families

[2] China's concern to internationalize its research and obtain recognition through, for example, the winning of Nobel prizes, has reached the point of obsession. Cao (2004, 2014) talks about the Nobel Prize complex or "Nobelmania" that existed in the absence of Chinese born scientists, with Chinese nationality at the time of the concession, working in a Chinese institution, until in 2015 the scientist Tu Youyou won the Medicine and Physiology prize for her contribution to the treatment of malaria.

So then, as Huang (2015) illustrates, the Chinese way is still receptive to Western influence and external international ranking systems or organizations, and it has made impressive progress in selecting elite universities.

[3] Notwithstanding, Indian researchers (Basu, Banshal, Singhal and Singh 2016) propose the application of a multidimensional "Quality-Quantity' Composite Index" to rank India's central universities, and there is a plethora of national ranking systems that seek ideographic contextualization. See Cakur, Acarturk, Alasehir and Cilingir, (2015) for a systematic comparison of national and global university ranking systems.

this system's social influence is undoubted and it has acquired a predominant role in determining the policies of both university administrations and national governments. The ARWU has improved from one annual edition to the next, in the sense of incorporating specialties within each field, to the extent that the latest edition is disciplinarily quite complete, as it covers 52 academic subjects in its 5 fields.

Behind the rankings of universities lies the emerging, yet already powerful, phenomenon of the internationalization of research and higher education, and the race to attract the most talented students and most highly qualified academic team, with the economic implications that this entails. As Bouchard (2017) states, the production of rankings constitutes a multidimensional market, which has also had a powerful ally: the media.

Scientometric indicators in the Shanghai ranking

The six scientometric indicators used are available and commented on at http://www.shanghairanking.com/ARWU-Methodology-2016.html. Its use as metadata makes the Shanghai Ranking a robust multivariate estimate, in Freyer's opinion (2014). For the purposes of clarity, the indicators are as follows:

- ALUMNI: The quality of teaching, measured by the number of students who have won Nobel Prizes and Field Medals (in mathematics), adjusted to seniority-decades of their stay (Weight: 10%).

- AWARD: The quality of teaching, measured by the number of professors who won Nobel prizes and Fields Medals (mathematics) while they were at that university, adjusted to the number of prize winners and the seniority for decades of their stay (Weight: 20%).

- HICI: The quality of the teaching staff, measured by the number of highly cited researchers in 21 broad thematic categories based on Web of Science (WoS) data, and

published in the document Highly Cited Researchers by Thomson Reuters and, from 2016 on, by Clarivate Analytics (Weight: 20 %).

- N&S: The quality of research, measured by the number of articles published in *Nature* and *Science* journals (Weight: 20%), adjusted to whether the author is the corresponding author, first author or following (Weight: 20%).

- PUB: The quality of research, measured by the number of articles published in journals indexed in the Science Citation Index Expanded (SCIE) and Social Sciences Citation Index (SSCI) databases of Clarivate Analytics and disseminated online through the WoS with a special weight of 2 for documents indexed in the SSCI database (Weight: 20%).

- PCP: Per capita research performance related to the size of an institution; that is, research output in the previous five indicators adjusted to the number of members of each university working full time (Weight: 10%).

This paper might be considered an exercise of meta-evaluation in the sense that Scriven (2009) defined: any evaluation of an evaluation, evaluation system, or evaluation device, and as a professional obligation of evaluators, similar to the consultant's version of a peer review. We are aware that we could be accused of an incomplete meta-evaluation, of only providing a methodological evaluation developed and presented as an integrative review; however, as Scriven (2009) states, "a partial meta-evaluation is better than none".

Consequently, it is not our intention to find fault with this ranking, which is probably the most valid of the many in existence, perhaps because it is the most credible (Berlin Principle 14); but rather to offer guidelines for its possible improvement in line with the Berlin Principles (CEPES-Institute for Higher Education Policy 2006; Barron 2017) as a legitimizing practice to institutionalize the rankings and align them critically and symbolically with academic values and evaluation systems.

Alongside ARWU, the other four most respected global rankings are the Leiden Ranking (CWTS 2017), the Quacquarelli Symonds (QS 2017), University Ranking by Academic Performance (URAP 2017), and the THE-Times Higher Education (THE 2017). A SWOT analysis of these four taken as a set can be read in Ferreira and Vidal (2017). Bougnol and Dulá (2014) recognize that "the one that emerges as the most successful at avoiding mistakes is CWTS Leiden Ranking". There are many to choose from, nevertheless, among the six rankings studied by Shehatta and Mahmood (2016) there are moderate to high correlations. In general, Aguillo, Bar-Ilán, Levene and Ortega (2010) show that there are reasonable similarities between the rankings, even though each applies a different methodology.

Reviewing criticism of the ARWU ranking

The discussions and proposals for the preparation of university rankings have been numerous and vehement (see Macri and Sinha 2006). O'Connell (2013) discusses the antagonistic discourse surrounding global university rankings that emanates from the contributions of research studies structured from discrepant perspectives. It could be said, as expressed by Daraio, Bonaccorsi and Simar (2015), that rankings are subject to a paradox in that the more they are criticized by social scientists and experts in methodology, the more attention they receive from the media and normative policy makers. However, Daraio et al. (2015) present four criticisms of university ranking systems, namely: a one-dimensional versus multifactorial structure; a lack of statistical robustness; dependence on the size of the institution, and lack of consideration of an input-output structure.

Billaut, Bouyssou and Vincke (2010) point out that the criteria used are not relevant, that aggregation methodology is plagued with many serious problems, and that

all rankings suffer from insufficient attention to basic structural aspects. One negative effect of the impact of the rankings is that universities prioritize activities and results that have a positive effect on the ranking itself (Elken, Hovdhaugen and Stensaker 2016); to which should also be added neglecting other functions of a university institution such as teaching or community services (university extension).

Specific criticisms have been raised about the appropriateness of this or other rankings for developing countries as such systems may induce a mimetic effect, encouraging these countries to adopt and adapt their national systems of higher education to the process that underlies the ranking. On the other hand, according to Elken, Hovdhaugen and Stensaker (2016), in Nordic countries such rankings have had a relatively modest impact on the decision-making and strategic actions of the Nordic universities studied, since there are few signs that they compromise existing identities in the region's universities.

However, in the case of the Shanghai ranking, its positive aspects seem to outweigh the negative ones; thus, in response to Florian's questioning (2007) its lack of reproducibility, Docampo (2013) states that it can be safely declared that ARWU results are in fact reproducible.

The critical mass of ARWU has facilitated a positive dynamic of derived research, a symptom of heuristic growth, as shown by manifold studies (i.e. Dehon, McCathie and Verardi 2010; Docampo, Egret and Cram 2015; Jeremic, Bulajic, Martic and Radojicic 2011; and Sadlak and Liu 2009).

Methodological considerations that make the Shanghai Ranking questionable

According to the authors of ARWU: "One of the factors that demonstrates the significant influence of ARWU is its scientifically solid, stable and transparent methodology" (quote

from http://www.shanghairanking.com/aboutarwu.html). Van Raan (2005), however, does not criticize the ranking but rather its bibliometric indicators that are insufficiently interpreted by inexperienced people who encourage quick and simplistic analyses, when higher quality indicators exist.

Florian (2007) questioned whether the data are reproducible, since the dependence between the indicator score relating to data of the SCIE database and the weighted number of items considered obeys a power law instead of the proportional dependence suggested in the official methodology. Docampo and Cram (2014) offer a comprehensive explanation as to why linearity might have been modified by the rankers, by replacing the unwanted dynamical effects of the annual re-scaling based on raw scores of the best performers. Discrepancies in proportionality are also detected in some indicator scores given by the number of articles published in *Nature* and *Science* journals and in the size indicators of students and teachers. Billaut, Bouyssou and Vincke (2010) question the Shanghai ranking again, pointing out that the criteria/indicators used are not relevant, that the aggregation methodology is plagued with serious problems and that it pays insufficient attention to basic questions of foundation.

This study is the result of a compilation of previous questions or threats to its validity, together with new considerations that reveal abundant methodologically relevant issues, following the guidelines of the CEPES-Institute for Higher Education Policy (2006). The adoption of this paper's suggestions would improve and refine any ranking, but especially ARWU. The suggestions follow, grouped by methodological category as threats that are not consistently controlled.

Problems with indicators or implementation-related threats

Omitted indicators. The ARWU omits any treatment of the three basic missions of university: training, management and service to the community. Margison (2014) questions these omissions in the ranking, which basically focuses on research since the weight given to teaching is awarded for the high research capacity of the included universities.

As an alternative, the following indicators could be used: employability (degree and level of employment), graduates seeking employment, institutional prestige according to reputation[4] among experts and stakeholders, and the potential for corporate governance of the university (Florez, López and López 2014). There are aspects of the efficiency of a university that are difficult to measure given its eminently qualitative nature (i.e. its ethos, its characteristic style, the personal values instilled in its graduates).

The omission of an indicator relating to patents, their various types and exploitation, is also a serious limitation. Indicators related to the transfer of knowledge generated through patents, models and prototypes could therefore be considered.

Certain types of document such as reviews of the research indexed in SCIE and SSCI are improperly omitted when it is generally acknowledged that these, together with articles, constitute mature research literature (fully fledged research). This is the gravest omission because reviews are usually the most cited documents and ARWU says definitively about the PUB indicator: “only publications of 'Article' type are considered” -see Indictors Definition - Methodology Section (Shanghai Ranking Consultancy 2017).

Moreover, publication in journals is not the gold-standard for scientific information in certain fields and disciplines, as is the case of Engineering or Architecture. Neither does the ARWU incorporate the value of high-impact books or monographs that

[4] The other two rankings (Quacquarelli Symonds (QS) World University Ranking and THE - Times Higher Education World University Ranking) could be criticized, however, for the excessive weight, more than 60%, of the institutional reputation generated by the surveys. This makes them even more questionable.

are not necessarily referenced with citations, when as Moksony, Hegedus and Csaszar (2014) indicate, there is a greater preference for books than articles as outlets for publication in qualitative departments, which puts the former at a disadvantage.

The ARWU disregards other indicators, which could be important as explanatory variables that would adjust the metadata for each university, such as the country's Gross Domestic Product, the level of institutional transparency and quality of its democracy (Jabnoun 2015). The two latter indicators particularly alert us to the ethical dimension in higher education. At a minimum, the variable funds allocated to higher education should be considered to adjust the additive metadata of each institution.

Supposed validity of the indicators used. This refers to the supposed validity of the ranking accorded by extended use, i.e. inferring validity due to use (Zeller 1997) or validity by consequences (Lane 2014). This is dangerous because it leads to and induces an unproductive disregard of an institution's other functions. We would argue that validity due to use or consequences, which appears to be what the Shanghai Ranking has been applying since 2003, is not sustainable.

In this sense, the relevance and validity of certain indicators is highly controversial; it could be said that what it is possible to measure is measured, but this is not always what is needed. Thus, quality of teaching is a questionable indicator given its low and very limited inclusion of only Nobel Laureates and/or Field prizes among its professors or alumni. This means that only a few universities are measured, which implicitly provides a very asymmetric distribution with too many structural zeros. Dehon, McCathie and Verardi (2010), while recognizing the excellence of the ARWU, have highlighted the excessive weight of top researchers, high marks, and overall research production of an institution.

The ARWU and the other various rankings that have proliferated in recent years do not provide the validity of any predictive criterion that show a high correlation between ranking distribution and other predictive indicators such as employability, professional success, the income of former alumni or satisfaction in their professional life[5]. In this way, and as the Berlin Principle 7 proposes "indicators must be chosen according to their relevance and validity … and not by availability of data".

Anomalous reliability. The distribution of valuations (position in the ranking) varies considerably from one year to the next, which indicates a strangely low reliability given that such remarkable changes in these institutions are not likely in such a short timespan. Such low reliability is particularly worrying, as noted by Sorz, Wallner, Seidler and Fieder (2015), for universities with low ranking positions, which often show inconclusive fluctuations from one year to the next, thus making the index questionable as an appropriate basis for management purposes. Nevertheless, Docampo (2011) shows consistent reliability when considering the 32 best national systems of higher education.

Questionable weighting of indicators. Obviously, different importance for aggregating performance in individual indicators leads to different rankings, and because final scores are based on weighted indicators, for which raw data and its processing are not publicly available, some differences may be attributable both to small variations on what Piro and Sivertsen (2016) believe are not important indicators, and to substantial variations on what we believe are important indicators. Here, Berlin Principle 9 could be also evoked: "Make the weights assigned to different indicators (if used) prominent and limit changes to them".

[5] López-Martín, Moreno-Pulido and Expósito-Casas (2018) reveal a validity problem of the Spanish U-Ranking (Fundación BBVA-IVIE, 2017), which could be associated with a systematic error in predicting performance criteria through features that are not relevant to it.

The ranking does not justify the special value of 2 given to documents indexed in the SSCI database, which inflates the value of Social Sciences production, although that SSCI scientific production represents only a low percentage (around 15%) of the total scientific production worldwide. Equally, the weights given to the indicators used (10%, 20%, 20%, 20%, 20%, 10%, respectively) to establish a combined indicator or metadata is more than questionable: it is discretional and without a broad pre-established consensus, except that given by the proponents of the ranking. Subsequently, as Safon (2013) determines by factorial analysis, the metadata could be an epiphenomenon of an X factor that has little to do with quality.

Bowman and Bastedo (2011) question the use of an additive approach, as opposed to a multiplicative one, since it includes different treatments of the student-teacher ratio and potential funding that may vary depending on the inclusion or exclusion of an institution. A multiplicative approach for aggregation would overcome these difficulties and even provide a more transparent interpretation of the weights. In this line, Ding and Liu (2011) integrated the subjective and objective weights by respectively using the additive and multiplicative model to reflect both the subjective considerations of experts and the objective information, and obtained three kinds of integrative weight.

Confusing indicators. The indicator "teacher quality measured by the number of highly-quoted researchers" is confusing because the 2015 ranking itself indistinctly used two *ad libitum* lists. This contravenes both Principles 9 and 16 of the Berlin Principles in terms of not changing the assigned weights (Principle 9, “limit changes to them”) and not reporting errors (Principle 16, “Institutions and the public should be informed about errors that have occurred”).

In the case of highly cited researchers, the ARWU considers only the first affiliation as on the Highly Cited Researchers list by Clarivate Analytics. This criterion is confusing

on both sides, ARWY and Clarivate, because it does not indicate whether it is the first institution where the author worked, the one where s/he produced most articles, or the one where s/he was working in the year the ranking was drawn up. It should be remembered that, unlike the immobility of academics in some universities and higher educational systems (e.g. Spain), teacher mobility is a general characteristic in other geographical areas and generally considered desirable or necessary in order to improve. Clearly, ARWU assumes this real confusion although it was inherited.

As for student and teaching staff quality indicators, higher values are assigned to the most recent university attended, therefore the time of permanence in each institution is not considered differentially (Martínez-Rizo 2011).

The quality of research, measured by the number of articles published in *Nature* and *Science* journals is questionable. The unilateral selection of these two generalist journals, which are not those with the highest impact factor (IF), and their high weight (20%) are two extremely controversial criteria. Thus, for example, a highly-specialized journal such as *CA-A Cancer Journal for Clinicians*, with a large IF (2016) = 187.04, and even a generalist medical journal such as the *New England Journal of Medicine*, IF (2016) = 72.406, make it difficult to understand why primacy is accorded to *Nature* IF (2016) = 40.137 and *Science*, IF (2016) = 37.205.

Although there is a broad consensus on the scope of IF for comparing different journals within a certain field, its use in the comparison between subject-specific and generalist journals is misleading. Moreover, the IF itself appears to be questionable as a gold evaluative standard because it is too limited in time and probably influenced by compliant citation practices, as Fernandez-Cano (1995) indicated long ago. Thus, the Shanghai Ranking project could be more about receiving universal approval than for a sustainable science in general.

Over-emphasized citation indicator. This ranking over-emphasizes the citation of journals indexed in the Journal Citation Reports (JCRs) of the WoS, with JCRs as the predominant indicator, giving a consequent dependence on WoS. Even the Ranking of Innovative Universities (RIU) that is published by Thomson Reuters and based exclusively on citation data, receives a critical view from a methodological perspective from Tijssen, Yegros-Yegros and Winnink (2016).

The data analysis problem or analytic threat

Opaque adjustment. The adjustment of the number of members is somewhat opaque as it does not consider the extent of dedication of each university's members to research, and only considers those devoted to research full-time. Even the ranking itself recognizes that if it is not possible to ascertain full-time teaching staff, it uses the weighted scores of the other five indicators. As Berlin Principle 6 claims, transparency should include the calculation of indicators as well as the origin of data.

If distribution of the statistical data of any indicator presents a significant distortion that determines asymmetric distributions, this ranking does not indicate what standard statistical techniques will be used, when necessary, to adjust the indicator. This is a serious point to consider given the large number of universities evaluated, with a huge number of them close to the lowest score (0%).

It should also be remembered that each individual indicator is first normalized to achieve comparable figures. This ultimately adds even more opacity to the final metadata. Tofallis (2012) presents a detailed technical discussion on how different data can be normalized and how this affects rankings. Jovanovic, Jeremic, Savic, Bulajic and Martic (2012) showed ARWU's great inconsistencies between the university ranks obtained

from the original compared to normalized data, with subsequent wide fluctuations between universities.

Simplistic standardization. For each indicator, the institution with the highest score is assigned a value of 100, and the values of the other institutions are calculated as a percentage of the maximum score. This ordinal reporting, due to its simplistic standardization on a common scale of 0-100, disregards the variability of the various indicators/variables, whereas other approaches could transform the data, taking into account the various dispersions of the indicators (Williams and de Rassenfosse 2016). These authors go on to state that transforming data muddies interpretation, and that the choice of which variables should be included is more important than the weights assigned to them. Bougnol and Dulá (2015) talk of isotonic attributes in the sense that a weighting scheme that uses a positive weight for the values of an attribute rewards longer magnitudes independently of their intrinsic quality; This represents another subtle instance of Merton's Matthew effect (1968).

One-dimensional metadata. Moed (2017) argues that the current evaluation systems inferred from rankings are still one-dimensional in that they provide finalized, seemingly unrelated, indicator values rather than offering a dataset and tools to observe patterns in multi-faceted data. Consequently, global rankings such as Shanghai offer a simplistic evaluation.

Selection bias or sample threat

The ARWU discards private research corporations and/or non-university research institutions that collaborate with universities. For example, no relevant corporations appear, such as IBM, Novartis or Vivendi, and highly qualified laboratories, such as Roche or Merck, are not included. It is well known that research of scientific-technical

and especially economic impact is carried out in private corporations (Gupta and Karisiddappa 2000), even when they benefit from the major outputs of science research funded by governments (Comins 2015).

The existence of parallel state research networks determines the bad position of French, Russian and even German universities. The reason for this lies in the fact that these countries have other research institutions that are more relevant than universities that teach at graduate level. In the case of France, the CNRS (*Centre National de la Recherche Scientifique*), in Spain, the Higher Council for Scientific Research (CSIC: *Consejo Superior de Investigaciones Científicas*), in Germany, the Max-Planck Society (*Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V*), and in Russia, the academies. Russia has only three institutions in the ARWU.

Of course, ARWU stands for Academic Ranking of World Universities and not for other private or governmental research corporations that are not educational, but which can obtain more resources to the detriment of universities; while some other systems may emphasize research in universities (Anglosphere countries) other systems, in contrast, could understate the research performed in universities (the Spanish, French, German or Russian cases).

The Shanghai Ranking shows a strong bias in favor of the Anglosphere, which is hardly surprising as the research production considered is usually printed in publications that use English as the *lingua franca.* Consequently, the research recovered from WoS databases feeds this bias; when the Berlin Principle 12 says “data that are collected with proper procedures for scientific data collection. Data collected from representative or non-skewed samples of students, faculty or other parties”.

Some explanatory notes and proposals to improve the evaluation of universities using rankings

In its preamble the Berlin Principles state: "it is important that those producing rankings … hold themselves accountable for quality in their own data collection, methodology, and dissemination." The following notes and proposals are thus intended to improve any evaluative process of higher education institutions, including any ranking. A ranking should only be considered as an additional instrument to advise universities on good practices related to internal evaluation services and policies.

In keeping with to the previous considerations, ARWU could be specifically refined in accordance with the Berlin Principles in the following methodological steps: considering other additional indicators rather than just five, including more consistent reliabilities over time, using a higher consensus in the weighting of indicators, avoiding confusing indicators, deemphasizing the citation indicator, making a clearer adjustment in the number of member indicators, employing a more sophisticated data standardization, incorporating multi-dimensional metadata, and improving the sample selection.

Notwithstanding, there are other considerations from the Berlin Principle as explanatory notes and proposals to improve the evaluation of universities when using rankings. The following are a series of guidelines aimed at contributing to the improvement of the evaluation of university institutions in line with the spirit of the Berlin Principles (CEPES-Institute for Higher Education Policy 2006) using rankings, and with methodological consistency, but which also consider the role of rankings and their plausible impacts.

Towards critical multiplism

Using a single ranking unilaterally as the only indicator of university quality is highly inappropriate and presupposes an assessment of quality given by a single evaluative tool; as the adage says: "a single way of evaluating is tantamount to not evaluating." Evaluative multiplism, achieved using various approaches, instruments and distinctly mixed methods, emerges as the most advisable option. It is even less admissible to consider a ranking as the arbiter of academic excellence. Two basic consensuses would be necessary in this regard: which indicators to use and what weight to assign to each one; two facets of the Shanghai ranking that are highly problematic. We should remember here Berlin Principle 2 in this regard: "Indicators designed to meet a particular objective or to inform one target group may not be adequate for different purposes or target groups".

Expanding on the idea of evaluative multiplism, longitudinal evaluations should be made with data already available in the ARWU time series from 2003-2016, to achieve a more robust evaluative pattern over time in such a way that the available rankings are rationally used. Obviously, the paragraph above does not question ARWU, but it does recommend using its results carefully in a multiplist longitudinal logic in accordance with the Berlin recommendations (Berlin Principle 1: Ranking is one of a number of diverse approaches to the assessment of higher education inputs, processes, and outputs). But above all, multiplism involves specifying "the linguistic, cultural, economic, and historical contexts of the educational systems being ranked" (Berlin Principle 5).

Evaluation for improvement

Establish the function of evaluation for purposes of improvement before doing so for purposes of accountability. Escudero (2017) criticizes the obsession with rankings, which constitutes a risk for the global quality of university institutions because, although they

can verify possible accountability to a certain extent, the rankings hardly tend towards the improvement of all facets of university institutions.

Evaluation demands meta-evaluation as an exercise for improvement, giving advice to universities on good practices for promoting related internal evaluation of their services and policies. Stufflebeam (2001) states the meta-evaluation imperative to ensure that evaluations provide sound findings and conclusions; that evaluation practices continue to improve. But evaluation is involved in a constellation of values and beliefs about what constitutes "quality" [improvement] of tertiary institutions (Berlin Principle 5). ARWU does not give any contextual consideration, only cold classifications.

Evaluation is not to be undertaken lightly

Establish a consistent culture and tradition of evaluation through the provision of qualified staff, equipment and facilities taking in account that the less researchers concern themselves with scientific evaluation, the more they risk being corrected by the administration or the government.

Improve and increase each institution's allocation of funds for internal and external evaluations. The phrase of B. F. Skinner: "Choose the best, and give them the means" (1956) remains completely valid. In this sense, universities should obtain external funding that complements state and/or public allocations and even seek collaboration with private assessment corporations. Establishing higher fees for students is more questionable, although it is not risky to conjecture that there is an inversely proportional relationship between position in the ranking and the fees paid by students. At Harvard, the best positioned university in each successive edition of the Shanghai Ranking, student fees for a bi-semester course for a Master of Education are set at a clear $75,000 including living expenses (Harvard Graduate School of Education 2017). In the Spanish University of

Granada (2017) for a similar master's degree, the student would have to invest around $9,000 (eight times less).

The optimal extent of concentration for ranking research remains to be explored. Many of our proposals could best be fulfilled in universities. Self-assessment is a general principle of Quality Management Systems. For this reason, even smaller universities will have to afford their own specialized section for evaluation within a specific department, looking for a general involvement of every university member, as the Berlin Principle 14 proposes "... including advisory or even supervisory bodies, preferably with some international participation", and with a recommendable high degree of specialization for institutions, especially in applied research aimed at the registration and exploitation of patents.

Evaluation is an eminently human undertaking

Conducting an evaluation is a powerfully human enterprise, which involves human beings; it must, therefore, be carefully accomplished. But some suggestions could be given considering the manifold agents involved in any ranking, as Berlin Principle 3 points out, "Institutions that are being ranked and the experts …. should be consulted often."

Create a tradition of evaluation that is not personalized around a *figure*, but preferably centered on young researchers who are well led by a manager. Senior university professors without a proven evaluation capacity should not remain at the head of evaluation and research institutions. Create the figure of an evaluation manager, who would be a senior expert with extensive experience. This could prove to be extremely significant even if it only achieved the two functions of invigorating the evaluation group and obtaining funds.

Enable the functioning of evaluation collectives with a corresponding professional accreditation structure and clearly defined objectives and interests, including personal ones. Avoid the dispersion of research groups, and groups built around a central figure or personality, promoting a trend towards the consolidation of broad-ranging stable evaluative groups, based on a stable structure supported by professional accreditation.

Give freedom of action to the group, avoiding political, union or corporate interference. Paradoxically, the most striking example of this could be the proposer of the Harvard Graduate School of Education and champion of socio-critical Pedagogy, philo-Marxist, Paulo Freire (1998a, b).

Recognize and enhance the evaluative *ethos* to avoid corruption such as parasitism, preferential treatment - especially nepotism - and patronage in its various manifestations (ideological, religious, political, economic or union-related).

Evaluation entails communication

Any ranking entails strengthening channels and sources of scientific-evaluative information (databases, journals, conferences and congresses) including the personal sources of that information (such as students, professors, employers and other stakeholders) as Berlin Principle 4 recommends. Therefore, evaluative indicators must be decidedly reconsidered by the manifold agents concerned.

Centralize expenditure on scientific-evaluative information (access to journals, payment of databases) and equipment, so that there is no expenditure on duplicate acquisitions. Agreement on evaluation agendas and on indicator weighting between the various stakeholders and affected groups would be desirable as a convergent exercise between the diverse audiences involved. The users of ranking should have some opportunity to make their own decisions (Berlin Principle 15).

Conclusions

As a final corollary, it must be said that we should not become obsessed with any ranking. There will always be universities that are better positioned than others, what should be investigated is the causes of the differences and the release of the ranking might prompt that investigation.

The assumption that a university will improve at the same time as its ranking position if a lot of money is invested is a fallacy: Consider the enormous amounts contributed to the universities of the Persian Gulf, for example, to the Saudi King Abdulazis University, whose website boasts that it has moved up 93-places (http://www.kau.edu.sa/home_english.aspx) on the Quacquarelli Symonds (QS) World University Ranking since 2014. In this sense, a university is conditioned by its economic context, which generates an underlying layer of critical mass upon which to advance; thus, an area with great agricultural development will be able to contribute to a university highly specialized in Agriculture.

Two principal evaluative determinations should be considered in university policies. It would be necessary and useful to differentiate institutions that focus on research from those that focus on professionalization, which place a greater emphasis on and dedication to teaching in accordance with Berlin Principle 3: "Recognize the diversity of institutions and take the different missions and goals of institutions into account". In consequence, the range of information sources for rankings and the messages each source generates could be very different depending the type of institution (oriented to research or centered on teaching for underserved communities).

This distinction should not constitute a bias that would justify the marginalization of the latter in their allocations and in the promotion of their teachers. External

evaluations, such as the Spanish medical MIR examination that gives access to Resident Intern Physician positions, could be informative about teaching potential (Vázquez, Murillo, Gómez, Martín, Chaves and Peinado 2008). On the other hand, the struggle for excellence requires that the most efficient institutions receive rewards, however how much this may bolster the Matthew effect.

A final inference is that any educational changes undertaken subsequent to the results of a ranking should be considered highly controversial and questionable because these systems lack the capacity to assess the complex issue of university quality and take into account that every change undertaken by universities is usually highly controversial and subject to criticism from the different stakeholders. The real issue is then whether the information that a ranking provides can be used, along with other information collected by interested parties, to inform sound evaluative decisions and not only to give satisfaction to egos of university authorities.

In summary, we offer some methodological warnings to anyone who would like to use ARWU inappropriately outside its originally intended use. Notwithstanding, and unfortunately, for the last fifteen years none of the researchers involved in this area have managed to establish a contrasting university ranking of recognized usefulness as effective as ARWU.

References

Aguillo, I. F., Bar-Ilan, J., Levene, M. & Ortega, J. L. (2010). Comparing university rankings. *Scientometrics*, *85*(1), 243-256.

Bagozzi, R. P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, *8*(4), 244-254.

Barron, G. R. S. (2017). The Berlin principles on ranking higher education institutions: Limitations, legitimacy, and value conflict. *Higher Educa*tion, *73*(2), 317-333.

Basu, A., Banshal, S. K., Singhal, K. & Singh, V. K. (2016). Designing a composite index for research performance evaluation at the national or regional level: Ranking central universities in India. *Scientometrics*, *107*(3), 1171-1193.

Billaut, J. C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shanghai ranking? *Scientometrics*, *84*(1), 237-263.

Bouchard, J. (2017). Academic media ranking and the configurations of values in higher education: A sociotechnical history of a co-production in France between the media, state and higher education (1976-1989). *Higher Education*, *73*(6), 947-962.

Bougnol, M-L., & Dulá, J. H. (2015). Technical pitfalls in university rankings. *Higher Education*, *69*(5), 859-866.

Bowman, N. A., & Bastedo, M. N. (2011). An anchoring effect on assessments of institutional reputation. *Higher Education*, *61*(4), 431-444.

Cakur, M. P., Acarturk, C., Alasehir, O. & Cilingir, C. (2015). A comparative analysis of global and national university ranking systems. *Scientometrics*, *103*(3), 813-848.

Cao, C. (2004). Chinese science and the 'Nobel Prize complex'. *Minerva*, *42*(2), 151-172.

Cao, C. (2014). The universal values of science and China's Nobel Prize pursuit. *Minerva*, *52*(2), 141-160.

Chanowitz, B., & Langer, E. J. (1981). Premature cognitive commitment. *Journal of Personality and Social Psychology*, *41*(6), 1051-1063.

CWTS-Centre for Science and Technology Studies. (2017). *CWTS Leiden ranking*. Accessed on line from: www.leidenranking.com/ranking/2017/list

CEPES-Institute for Higher Education Policy (2006). *Berlin principles on ranking of higher education institutions*. Accessed on line from: https://www.che.de/downloads/Berlin_Principles_IREG_534.pdf

Comins, J. A. (2015). Data-mining the technological importance of government-funded patents in the private sector. *Scientometrics*, *104*(2), 425-435.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. *European Journal of Operational Research*, *244*(3), 918-930.

Davidov, E. (2009). Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspective. *Political Analysis*, *17*(1), 64-82.

Dehon, C., McCathie, A., & Verardi, V. (2010). Uncovering excellence in academic rankings: A closer look at the Shanghai ranking. *Scientometrics*, *83*(2), 515-524.

Ding, J., & Qiu, J. (2011). An approach to improve the indicator weights of scientific and technological competitiveness evaluation of Chinese universities. *Scientometrics*, *86*(2), 285–297.

Docampo, D. (2011). On using the Shanghai ranking to assess the research performance of university systems. *Scientometrics*, *86*(1), 77-92.

Docampo, D. (2013). Reproducibility of the Shanghai academic ranking of world universities results. *Scientometrics*, *94*(2), 567-587.

Docampo, D. & Cram, L. (2014). On the internal dynamics of the Shanghai ranking. *Scientometrics*, *98*(2), 1347–1366.

Docampo, D. & Cram, L. (2017). Academic performance and institutional resources: a cross-country analysis of research universities. *Scientometrics*, *110*(2), 739-764.

Docampo, D., Egret, D., & Cram, L. (2015). The effect of university mergers on the Shanghai ranking. *Scientometrics*, *104*(1), 175–191.

Elken, M., Hovdhaugen, E., & Stensaker, B. (2016). Global rankings in the Nordic region: Challenging the identity of research-intensive universities? *Higher Educa*tion, *72*(6), 781-795.

Escudero, T. (2017). La fiebre con los rankings. Un riesgo para la calidad global de las instituciones universitarias [Ranking fever. A risk to the overall quality of the universities]. *El País Digital*, June 26. Accessed on line from: http://elpais.com/elpais/2017/06/23/opinion/1498226306_209367.html

Fernández-Cano, A. (1995). *Métodos para evaluar la investigación en Psicopedagogía* [Methods for evaluating psychopedagogical research]. Madrid: Síntesis.

Ferreira, C., & Vidal, J. (2017). El impacto de los rankings sobre la actividad de las universidades [The impact of rankings on universities activity]. In AIDIPE (Eds.), *Actas XVIII Congreso Internacional de Investigación Educativa. Interdisciplinariedad y transferencia* [Proceedings of the 18th International Congress on Educational Research. Interdisciplinarity and transfer] (pp. 691-698). Salamanca: AIDIPE.

Freire, P. (1998a). The adult literacy process as cultural action for freedom. *Harvard Educational Review*, *68*(4), 480-498. (Reprinted from *Harvard Educational Review*, *40*, 1970).

Freire, P. (1998b). Cultural action and conscientization. *Harvard Educational Review*, *68*(4), 499-521. (Reprinted from *Harvard Educational Review*, *40*, 1970).

Flórez, J. M., López, M. V., & López A. M. (2014). El gobierno corporativo de las Universidades: Estudio de las 100 primeras Universidades del ranking de Shanghái [Corporate governance: Analysis of the top 100 universities in the Shanghai ranking]. *Revista de Educación*, *364*, 170-196.

Florian, R. (2007). Irreproducibility of the results of the Shanghai academic ranking of world universities. *Scientometrics*, *72*(1), 25-32.

Freyer, L. (2014). Robust rankings: Review of multivariate assessments illustrated by the Shanghai rankings. *Scientometrics*, *100*(2), 391-406.

Fundación BBVA-IVIE (2017). *U-ranking de las universidades españolas* [U-ranking of Spanish universities]. Accessed on line from: http://www.u-ranking.es/analisis.php#

Gupta, B. M. & Karisiddappa, C. R. (2000). Modelling the growth of literature in the area of theoretical population genetics. *Scientometrics*, *49*(2), 321-355.

Harvard Graduate School of Education. (2017). *Tuition & costs*. Accessed on line from: https://www.gse.harvard.edu/financialaid/tuition

Huang, F. (2015). Building the world-class research universities: A case study of China. *Higher Education*, *70*(2), 203–215.

Grimes, D. A. (1993). Technology follies. The uncritical acceptance of medical innovation. *JAMA*, *269*(23): 3030-303

Jabnoun, N. (2015). The influence of wealth, transparency, and democracy on the number of top ranked universities. *Quality Assurance in Education*, *23*(2), 108-122.

Jeremic, V., Bulajic, M., Martic, M., & Radojicic, Z. (2011). A fresh approach to evaluating the academic ranking of world universities. *Scientometrics*, *87*(3), 587–596.

Jovanovic, M., Jeremic, V., Savic, G. Bulajic, M. & Martic, M. (2012). How does the normalization of data affect the ARWU ranking? *Scientometrics*, *93*(2), 319-327.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127-137.

López-Martín, E., Alexis Moreno-Pulido, A. & Expósito-Casas, E. (2018). Validez predictiva del u-ranking en las titulaciones universitarias de ciencias de la salud [The U-ranking's predictive validity in university health studies]. *Bordón. Revista de Pedagogía*, *70*(1), 57-72.

Macri, J. & Sinha, D. (2006). Rankings methodology for international comparisons of institutions and individuals: An application to economics in Australia and New Zealand. *Journal of Economic Surveys*, *20*(1), 111-156.

Margison, S. (2014). University rankings and social science. *European Journal of Education*, *49*(1), 45-59.

Martínez-Rizo, F. (2011). Los rankings de universidades: Una visión crítica. [University rankings: A critical view]. *Revista de la Educación Superior*, *40*(157), 2-21.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56-63.

Moed, H. F. (2017). A critical comparative analysis of five world university rankings. *Scientometrics*, *110*(2), 967-990.

Moksony, F., Hegedus, R. & Csaszar, M. (2014). Rankings, research styles, and publication cultures: A study of American sociology departments. *Scientometrics*, *101*(3), 1715-1729.

O'Connell, C. (2013). Research discourses surrounding global university rankings: Exploring the relationship with policy and practice recommendations. *Higher Education, 65*(6), 709-723.

Piro, F. N. & Sivertsen, G. (2016). How can differences in international university rankings be explained? *Scientometrics*, *109*(3), 2263-2278.

Quacquarelli Symonds –QS- (2017). *QS World University Rankings 2016-2017*. Accessed on line from: https://www.topuniversities.com/university-rankings/world-university-rankings/2016

Sadlak, J., & Liu, N-C. (Eds.). (2009). *The world-class university and ranking: Aiming beyond status*. Bucharest: UNESCO-CEPES.

Scriven, M. (2009). Meta-evaluation revisited. *Journal of MultiDisciplinary Evaluation*, *6*(11), iii-viii.

Safon, V. (2013). What do global university rankings really measure? The search for the X factor and the X entity. *Scientometrics*, *97*(2), 223-244.

Shanghai Ranking Consultancy (2017). *Academic Ranking of World Universities 2017*. Accessed on line from: http://www.shanghairanking.com

Shehatta, I. & Mahmood, K. (2016). Correlation among top 100 universities in the major six global rankings: policy implications. *Scientometrics*, *109*(2), 1231-1254.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist*, *11*(5), 221-223,

Sorz, J., Wallner, B., Seidler, H., & Fieder, M. (2015). Inconsistent year-to-year fluctuations limit the conclusiveness of global higher education rankings for university management. *PEERJ*, *3*, e1217.

Stufflebeam, D. (2001). The metaevaluation imperative. *American Journal of Evaluation*, *22*(2), 183–209.

THE-Times Higher Education World University Ranking. (2017). *Times Higher Education World University Rankings 2016-2017*. Accessed on line from: https://www.timeshighereducation.com/world-university-rankings/2017/world-ranking

Tijssen, R. J. W., Yegros-Yegros, A. & Winnink, J. J. (2016). University-industry R&D linkage metrics: validity and applicability in world university rankings. *Scientometrics*, *109*(2), 677-696.

Tofallis, C. (2012). A different approach to university rankings. *Higher Education*, *63*(1), 1-18.

Universidad de Granada. (2017). *Másteres oficiales de la UGR* [Oficial masters in the UGR]. Accessed on line from: https://masteres.ugr.es/pages/masteres

URAP- Informatics Institute of Middle East Technical University (2017). University Ranking by Academic Performance. Accessed on line from: http://www.urapcenter.org/2017/

van Raan, A. F. C. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133-143,

Vázquez, G., Murillo, F. Cabezas, J., Gómez, J., Martín, C., Chaves, J., & Peinado, J. L. (2008). El examen MIR, su cambio como una opción estratégica [The MIR

examination, the change as a strategic option]. *Educación Médica*, *11*(4), 203-206.

Williams, R., & de Rassenfosse, G. (2016). Pitfalls in aggregating performance measures in higher education. *Studies in Higher Education*, *41*(1), 51-62.

Virk, H. S. (2016). Shanghai rankings 2016: Poor performance of Indian universities. *Current Science*, *111*(4), 601-601.

Zeller, R. A. (1997). Validity. In J. P. Keeves (Ed.), *Educational research, methodology and measurement: An international handbook* (pp. 822-829). New York: Pergamon.