High Fluctuations of THES-Ranking Results in Lower Scoring Universities

Johannes Sorz[1], Martin Fieder[2], Bernard Wallner[2] and Horst Seidler[2]

johannes.sorz@univie.ac.at
[1]University of Vienna, Office of the Rectorate, Universitätsring 1, A-1010 Vienna (Austria)

martin.fieder@univie.ac.at, wallner@univie.ac.at, horst.seidler@univie.ac.at
[2]University of Vienna, Department for Anthropology, Althanstrasse 14, A-1090 Vienna (Austria)

Abstract
A regression analysis of results from the Times Higher Education World University Rankings (THES-Ranking) from 2010-2014 shows high fluctuations in the rank and score for lower scoring universities (below position 50) which lead to inconsistent "up and downs" in the total results. We conclude that these fluctuations do not correspond to actual university performance. They create the impression of the THES-Ranking as a "gamble" for universities below rank 50. We suggest that THE alters its ranking procedure insofar as universities below position 50 should be ranked summarized only in groups of 25 or 50. Additionally, we argue for introducing a standardization process for THES-Ranking data by using common suitable reference data to create calibration curves represented by non-linearity or linearity.

Conference Topic
University Policy and Institutional Rankings

Introduction
Global higher education rankings have received much attention recently and, as can be witnessed by the growing number of rankings being published every year, this attention is not likely to subside. Besides the arguable use of results from global rankings as an instrument for rational university management, they remain influential for stakeholders inside and outside academia. A plethora of regional and national rankings exist, and 10 global higher education rankings are currently attempting to rank academic institutions worldwide. Numerous studies have analyzed and criticized higher education rankings and their methodologies (van Raan, 2005; Buela-Casal et al., 2007; Ioannides et al., 2007; Hazelkorn, 2007; Aguillo et al., 2010; Benito and Romera, 2011; Hazelkorn, 2011; Rauhvargers, 2011; Tofallis, 2011; Saisana et al. 2011; Safon, 2013; Rauhvargers, 2013). This casts justified doubt on a sensible comparison of universities hailing from different higher education systems and varying in size, mission and endowment based on mono-dimensional rankings and league tables (Hazelkorn, 2014). Several studies have demonstrated that data used to calculate ranking scores can be inconsistent. Thus, bibliometric data from international databases (Web of Science, Scopus), used in most global rankings to calculate research output indicators, favor universities from English-speaking countries and institutions with a narrow focus on highly-cited fields, which are well covered in

these databases. This puts universities from non-English-speaking countries, with a focus on the arts, humanities and social sciences, at a disadvantage when being compared in global rankings (Calero-Medina et al., 2008; van Raan et al., 2011; Waltman et al., 2012). Data submitted by universities to ranking agencies (e.g. personnel data, student numbers) can be problematic to compare due to different standards. These incompatibilities are being amplified because university managers have become increasingly aware of global rankings and try to boost their performance by "tweaking" the data they submit to the ranking agencies (Spiegel Online, 2014). Beyond all the data issues, there is the effect that universities with lower positions in the rankings often encounter volatile ups and downs in their consecutive year-to-year ranks. This creates the sensation of contending in a "gamble" in which results are calculated at random by ranking agencies. Such effects make global university rankings in many cases an inappropriate tool for university managers: the ranking results simply do not reflect the universities' actual performance or their management strategies. Volatile jumps are also difficult to explain to the media, which often engage in sensationalism when covering rankings by interpreting subtle changes of scores, even within the margins of statistical deviations, as substantial shifts in performance. Bookstein et al. (2010) found unacceptably high year-to-year variances in the score of lower ranked universities caused by statistical noise in the Times Higher Education World University Ranking (THES), one of the currently most popular global rankings. We again observed puzzling variances in the THES-Ranking 2014-2015, published in October 2014. Accordingly, we here analyze the fluctuations in score and rank of the THES-Ranking by calculating a regression analysis for consecutive years for 2010-2014 to determine the random component of these fluctuations. The methodology of the THES-Ranking was revised several times in varying scale, before and after the split with Quacquarelli Symonds (QS) in 2010 and the new partnership with Thompson Reuters. Times Higher Education (THE) calculates 13 performance indicators, grouped into the five areas Teaching (30%), Research (30%), Citations (30%), Industry income (2.5%) and International outlook (7.5%). However, THE does not publish the scores of individual indicators, only those of all five areas combined. Since 2010, the research output indicators are calculated based on Web of Science data. Most of the weight in the overall score is made up by the normalized average citations per published paper (30%), and by the results of an academic reputation survey (33%) assessing teaching and research reputation and influencing the scores of both areas (Rauhvargers, 2013; THE, 2014). In the past, criticism has been levied against this survey. Academic peers can choose universities in their field from a preselected list of institutions and, although universities can be added to the list, those present on the original list are more likely to be nominated. This leads to a distribution skewed in favor of the institutions at the top of the rankings (Rauhvargers, 2011; Rauhvargers, 2013). THE allegedly addressed this issue by adding an exponential component to increase differentiation between institutions, yet no information is available on its mode of calculation (Baty, 2011; Baty, 2012).

Methods

We used the publicly available data on scores and ranks from the THES-Ranking for the years 2010, 2011, 2012, 2013 and 2014, including only those universities ranked from 1 to 200. We performed the following analysis: i) we regressed the scores of the ranking of the year t-1 on the scores of the year t; ii) we regressed the ranks of the ranking of the year t-1 on the ranks of the year t; iii) we plotted the scores in descending order and iv) we determined the random component of the fluctuations in the ranks from year to year.

Results

Regression of the scores and ranks of two consecutive years

The regression of the scores—particularly of the ranking 2010-2011 regressing on the scores of the ranking of 2011-2012—shows a very high fluctuation/noise (Figure 1a), especially for the lower ranked universities. Moreover, the noise among the lower ranked universities seems to be higher compared to the already very noisy THES-Ranking performed by QS before 2010 (Bookstein et al., 2010, Figure 1). Note that in the rankings in the years following 2010-2011, the noise in the THES-Ranking did improve (Figure 1b-d).

Association between Scores and Ranks

Nonetheless, a general problem of the THES-Ranking remains: the difference in the scores among the 50 highest scoring universities is considerably higher compared to the difference among the lower scoring universities. This clearly suggests a non-linear relationship between scores and ranks (Figure 2 a-e). The consequence is that the ranks of the high scoring universities are much more robust to deviations in the scores from year to year. In the lower ranking universities, however, even very small, more or less random deviations (around 0.5%) lead to unexpected "high jumps" in the ranks from year to year (Figure 1e-h).

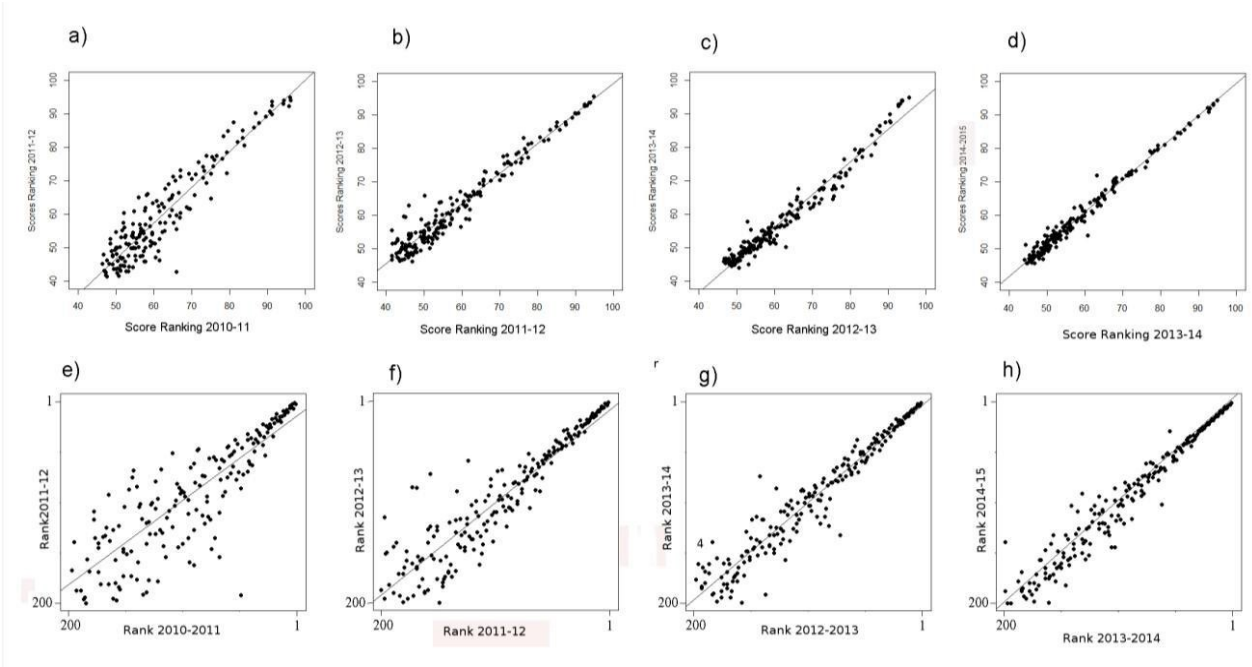

Figure 1a-1d) Scores of the year t-1 regressing on the score of the year t from the ranking 2010-11 on. Figure 1e-1h) Ranks of the year t-1 regressing on the ranks of the year t from the ranking 2010-11 on. Linear regression line indicates perfect association, e.g. no changes in ranks and scores between two consecutive rankings.

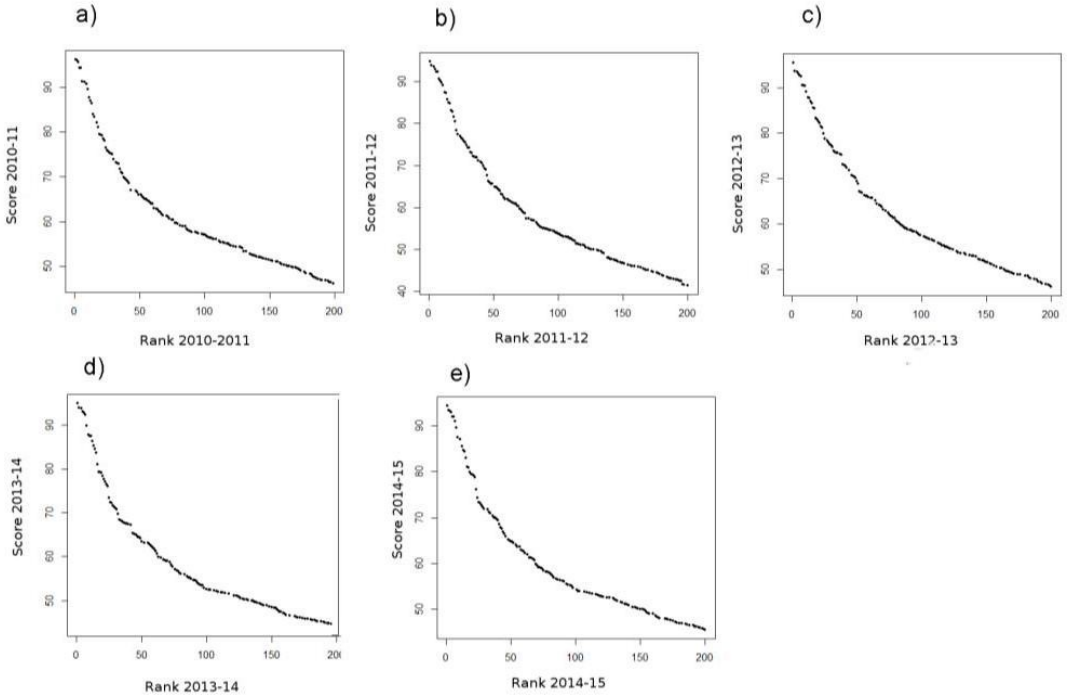

Figure 2 a-e). Ranks plotted against scores for the THES-Ranking a) 2010-11; b) 2011-12; c) 2012-13; d) 2013-2014; e) 2014-15

Discussion and Outlook

High ranking positions achieved by a small group of universities are often self-perpetuating, especially due to the intensive use of peer review indicators, which improve chances of maintaining a high position for universities already near the top (Bowman & Bastedo, 2011; Rauhvargers, 2011). This phenomenon also corresponds to the Matthew effect, which was coined by Merton (1968) to describe how eminent scientists will often get more credit than a comparatively unknown researcher, even if their work is similar: credit will usually be given to researchers who are already famous. The intensive and exaggerated discussion in the media of the "up and downs" of universities in the THES-Ranking is particularly misleading for the lower scoring universities (below approximately a score of 65% and a rank of 50; above scores of 65%, the relationship between ranks and scores is steeper, and it flattens for scores below 65%). This is because the ranking positions suggest substantial shifts in university performance despite only very subtle changes in score. In fact, merely random deviations must be assumed. One reason lies in the weighing of indicators by THE, with the emphasis on citations and peer review (totaling more than 65% of the total score). For lower ranked universities, a few highly cited publications, or the lack thereof, or few points asserted by peers in the reputation survey, probably make a significant difference in total score and position. In a follow up study that is currently under review we compared the results from THES with the results of the ARWU-Ranking (aka Shanghai-Ranking). Although the ARWU-Ranking seems to be more robust than the THES-Ranking (less year-to-year fluctuations probably due to the omittance of peer review indicators), we also found fluctuations below rank 50 and patterns of non-linearity between ranks and scores. Furthermore we found out that year-to-year results do not correspond in THES- and ARWU-Rankings for universities below that rank.

Ranking results have a major influence on the public image of universities and even impact their claim to resources (Espeland & Saunder, 2007; Hazelkorn, 2011). Accordingly, such fluctuations in the THES-Rankings can have serious implications for universities, especially when the media or stakeholders interpret them as direct results of more or less successful

university management. Our initial data in combination with the data from the literature strongly suggests that universities as well as policy makers and stakeholders should avoid to use rankings, especially league-tables, for management purposes or for strategic planning.

More specifically, the THES-Rankings in their current form have very limited value for the management of universities ranked below 50. This is because the described fluctuations in rank and score probably do not reflect actual performance, whereby the results cannot be used to assess the impact of long-term strategies. Thus, results from the THES (and to some extent also the ARWU) should be used only with great discretion. The low correlation between the ranks of the THES and the ARWU ranking, particularly for the universities ranked below 50 in both rankings, creates another serious doubt if rankings should be used for any management purposes at all. Maybe a "meta-analysis" of rankings could be reasonable to derivate consistent and reliable results from rankings. If done, such a meta-analysis should include as many rankings as possible to reduce random perturbations.

Multidimensional rankings, like the U-Multirank (http://www.u-multirank.eu), seem to offer a more versatile picture that reflects both the diversity of higher education institutions and the variety of dimensions of university excellence, allowing university managers to compare institutions on various levels. Although multidimensional rankings do get less public attention than league-tables and they can be prone for errors for the same reasons as monodimensional rankings (e.g., incompatibility of data provided by the universities), from the perspective of a university manager, they offer a more diverse toolset to gauge an institutions strength and weaknesses and to benchmark comparable universities.

"Rankings are here to stay, and it is therefore worth the time and effort to get them right," warns Gilbert (2007). That is especially true for monodimensional rankings, like the THES, that spark a lot of media attention. What could be done to address the fluctuations in the THES-Rankings for universities below rank 50 and to avoid the impression of a "gamble" in which THE "rolls a dice" to determine scores and ranks? THE has already addressed fluctuations to some extent by ranking universities only down to position 200, followed by groups of 25 from 201-300 and groups of 50 from 300 to 400. Nonetheless, based on our data we believe that this is not going far enough and suggest that universities should be summarized in groups of 25 or 50 below the position of 50.

The analyzed curves of scores vs. ranking positions in Figure 2 do have analogous characteristics for example to semi-logarithmic curves produced in analytic biochemistry. The accuracy of such curves is limited to the steepest slope of the curve, whereas asymptote areas deliver higher fuzziness (Chan, 1992). Thus, a further suggestion to avoid the blurring dilemma is the methodological approach of introducing a standardization process for THES-Ranking data. This would involve using common suitable reference data to create calibration curves represented by non-linearity or linearity. However, more research in this area is necessary.

The results presented in this paper are only the starting point and we plan to do more in-depth analyses of the variations in the various indicators in the future. We already have extended our analysis to include the ARWU-Ranking (paper currently in review) and we plan to analyze and compare other major higher education rankings (e.g. the QS-Ranking) in future publications to assess their usability for university management purposes.

References

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

Benito, M. & Romera, N. (2011). Improving quality assessment of composite indicators in university rankings: a case study of French and German universities of excellence. *Scientometrics 89*, 153–176.

Baty, P. (2011, October 6) THE Global Rankings: Change for the better, *Times Higher Education*, http://www.timeshighereducation.co.uk/world-universityrankings/2011-12/world-ranking/methodology

Baty, P. (2012, October 4). The essential elements in our world-leading formula, *Times Higher Education*, http://www.timeshighereducation.co.uk/worlduniversity-rankings/2012-13/world-ranking/methodology

Bowman, A.M. & Bastedo, N.M. (2011). Anchoring effects in world university rankings: exploring biases in reputation scores. *Higher Education 61*, 431–444

Bookstein, F.L, Seidler, H., Fieder, M., & Winckler, G. (2010). Too much noise in the Times Higher Education rankings. *Scientometrics 85*, 295–299.

Buela-Casal, G., Gutierrez-Martinez, O., Bermudes-Sanchez, M., & Vadillo-Munoz, O. (2007) Comparative study of international academic rankings of universities. *Scientometrics 71*, 349-365.

Calero-Medina, C., López-Illescas, C., Visser, M.J and Moed, H.F. (2008). Important factors when interpreting bibliometric rankings of world universities: an example from oncology. *Research Evaluation, 17* (1), 71–81.

Chan, D.W. (ed) (1992) *Immunoassay Automation: a Practical Guide*. San Diego, CA: Academic Press.

Espeland, W.N & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds, *American Journal of Sociology, 113* (1), 1–40.

Gilbert, A. (2007). Academics strike back at spurious rankings. *Nature, 447*, 514–515.

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.

Hazelkorn, E. (2011) *Rankings and the Reshaping of Higher Education: the Battle for World Class Excellence*. Basingstoke: Palgrave-MacMillan.

Hazelkorn, E. (2014) Reflections on a decade of global rankings: what we've learned and outstanding issues, *European Journal of Education, 49* (1), 12–28.

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

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

Rauhvargers, A. (2011) EUA Report on Global Rankings and their Impact – Report I (European University Association). http://www.eua.be/pubs/Global_University_Rankings_and_Their_Impact.pdf

Rauhvargers, A. (2013). EUA Report on global rankings and their Impact – Report II (European University Association). http://www.eua.be/Libraries/Publications_homepage_list/EUA_Global_University_Rankings_and_Their_Impact_-_Report_II.sflb.ashx.

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

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

Spiegel Online (2014). Deutsche Unis im "THE"-Ranking: Das Wunder von Tübingen. 02.10.2014. http://www.spiegel.de/unispiegel/studium/uni-ranking-hochschulenim- the-ranking-a-994684.html

Times Higher Education. (2014). World University Rankings 2014-2015 methodology http://www.timeshighereducation.co.uk/world-university-rankings/2014-15/worldranking/methodology

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

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

van Raan, T., Leeuwen, T., & Visser, M. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics, 88*, 495–498.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, N.C., Tijssen, R.J., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. CWTS Working Paper Series. http://arxiv.org/abs/1202.3941

The Vicious Circle of Evaluation Transparency – An Ignition Paper

Miloš Jovanović[1]

[1]*milos.jovanovic@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis, Appelsgarten 2, 53879 Euskirchen (Germany)

Introduction

The present paper introduces a model, which describes different phases that typically occur in situations, in which a researching subject (e. g. an author, an institution, a country etc.) needs to be evaluated and in which some kind of reward (e. g. monetary in the form of a bonus or funding) is based on this evaluation. This model, the present author calls it the "vicious circle of evaluation transparency", will be underlined by giving examples for each of its phases. In order to be able to observe a process that is described by this model, there first needs to be something that is to be evaluated, for example a research group at a university. Such a need normally comes up, when money is to be divided among different groups or focused on one. The problem of evaluation and rewarding is at the core of the model (see Figure 1).

Figure 1. The "vicious circle of evaluation transparency"-model.

Phase I – Evaluation and rewarding by subjective and intransparent criteria

The first question that might come up in such a situation is the question of how to evaluate a research group. In hierarchically organized universities the leader of a department will decide whether or not and how this group is evaluated. Very often, this person is also the one that conducts the evaluation and, based on this, determines the type and amount of a reward or funding (or some kind of penalty, if the evaluation is negative). In today's world of vast amounts of digital data, it might be hard for only one person to do such an evaluation. Naturally, having one person alone evaluate a group's performance and decide on rewards will lead to a number of persons feeling unfairly evaluated, because the evaluator might not know about their achievements or their work in detail. This criticism might be alleviated in part by expanding the number of evaluators, for example by having a board of evaluators. Another possibility is to improve the transparency of the evaluation by documenting and publishing certain evaluation criteria by which the evaluated subjects can read about the evaluations and try to strive to get a better evaluation. These evaluation criteria are a first step towards phase II of the model.

Phase II – Introduction of "objective" and transparent criteria

These evaluation criteria might be subjective. For example "Quality of work" can be a criterion that is evaluated differently by different people. In order to make evaluation criteria comparable and independent of the evaluating person, "objective" criteria are often introduced. The reason why the word is put into quotation marks is due to the fact that very often these "objective" criteria are not objective at all. The introduction of "objective" and transparent criteria is a simplification of reality, an attempt to put parts of reality into some kind of a score in order to compare them with each other. Bibliometric indicators are one example of such a simplification. In many countries, different kinds of "objective" and subjective evaluation criteria have been introduced, for example in Italy (Abbott, 2009). Normally, these "objective" evaluation criteria (often in the form of different kinds of indicators) are communicated transparently. And while transparency is an important factor for these evaluations, it also leads to one problem in this phase: the fact that the evaluated subjects, in our example researchers at universities, react to the evaluation by starting to change their behavior, in order to maximize their scores in the evaluation. Of course, one reason behind evaluation is to positively influence the behavior of the evaluated researchers. But in Germany, for example, this has led to authors aiming to publish more in internationally known journals that have a US publisher and which are more general in their scope (Michels & Schmoch, 2013). This underlines the fact that authors do not base the decision in which journal they wish to publish in on scientific reasons

alone and constitutes a negative change of behavior. Also, some of the evaluated subjects might complain that the evaluation criteria do not reflect their work adequately and need to be refined. This leads to the next phase.

Phase III – Adaptation and enrichment of "objective" criteria

The need to fairly represent and evaluate researchers' work in the evaluation criteria and to adapt these in order to not allure unwanted change of behaviour leads to reforms in the evaluation system, e.g. new or a mix of indicators are proposed. The current discussion on alternative metrics is an example for phase III (e.g. in Haustein et al., 2014). The problem here is, that phase III is actually reintroducing parts of the simplification of reality, which was conducted in phase II. The evaluation criteria become more complicated again. A country example for this phase is the Czech Republic, which introduced performance-based research funding (phase II). A study by Vanacek (2014) found that the number of publications increased very quickly. He shows that in comparison to the quickly growing number of publications the quality seems to have stagnated and recommends reworking the procedure of evaluation and performance-based funding in order to increase not only the number of publications but also their quality (phase III). But for some research communities, the adaptation and enrichment of the "objective" criteria is no option. Instead, these criteria are rejected. For example, there is an ongoing discussion in the mathematical community. Authors note that bibliometric data lose "crucial information that is essential for the assessment of research". It is pointed out that bibliometric indicators can be manipulated and lead to undesirable publishing practices (Adler, Ewing, & Taylor, 2009). The authors also dismiss reputation, as determined by surveys as a possible way of measuring the quality of a journal. The evaluation of journal editorial processes is not seen as a good way of ranking journals either. Instead, the authors recommend an "honest, careful rating of journals based on the judgment of expert mathematicians", which is the point, where phase IV starts.

Phase IV – Removal of "objective" criteria and return to phase I

Concretely, the IMU recommends that a rating committee of 16-24 experienced and respected mathematicians should be appointed. Without going into too much detail, this committee (via various panels) is then supposed to rate the different journals and assign them to tiers (ranging from tier 1 = high quality journal to tier 4 = low-class journal) (Journal Working Group, 2011). This system is similar to the peer review process. Introducing evaluation by a committee of experts, either by rejecting "objective" evaluation criteria or because the evaluation system has become too complicated, brings the model full circle. The evaluation has reached phase I again. One should note that in phase II of this new cycle, the criteria probably will not be the same as in the first cycle. Newly developed and more sophisticated criteria will take their place.

Conclusion

It is this author's personal opinion that the above described model of evaluation transparency not only describes a typical process in which bibliometric indicators are involved but rather evaluation processes in general. If this is the case, one may discuss possibilities to change this, since a cycle like this is not an optimal solution. An option might be the introduction of diametrically opposed evaluation criteria so that an evaluated subject could not be good in all criteria. Another idea that might serve to fan the discussion on this topic would be the introduction of a changing system of criteria, akin to the disciplines at Olympic Games. The criteria could be published a year before the evaluation takes place and would change each year. This would be a transparent system, while the evaluated researchers would not need to change their behavior in a negative way because the next year the criteria would be different. Whatever changes might be introduced, it is this author's opinion that the vicious circle has to be stopped and replaced by a different system that leads to the desired goal: a fair evaluation of research.

References

Abbott, A. (2009). Italy introduces performance-related funding. *Nature, 460*, 559.

Adler, R., Ewing, J. & Taylor, P. (2009). Citation Statistics: A Report from the IMU in Cooperation with the ICIAM and the IMS. *Statistical Science,* 1-14.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., &Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics, 101*, 1145-1163.

Journal Working Group (2011). *Report of the IMU/ICIAM Working Group on Journal Ranking.* Retrieved March 10, 2015 from: http://www.mathunion.org/fileadmin/IMU/Report/WG_JRP_Report_01.pdf

Michels, C. & Schmoch, U. (2013). Impact of bibliometric studies on the publication behaviour of authors. *Scientometrics, 98*, 369-385.

Vanacek, J. (2014). The effect of performance-based research funding on output of R&D results in the Czech Republic. *Scientometrics, 98*, 657-681.