Scholarly Assessment Reports

Developing a Method for Evaluating Global University Rankings

RESEARCH

ELIZABETH GADD
RICHARD HOLMES
JUSTIN SHEARER
Author affiliations can be found in the back matter of this article

Levy Library Press

ABSTRACT

Describes a method to provide an independent, community-sourced set of best practice criteria with which to assess global university rankings and to identify the extent to which a sample of six rankings, Academic Ranking of World Universities (ARWU), CWTS Leiden, QS World University Rankings (QS WUR), Times Higher Education World University Rankings (THE WUR), U-Multirank, and US News & World Report Best Global Universities, met those criteria. The criteria fell into four categories: good governance, transparency, measure what matters, and rigour. The relative strengths and weaknesses of each ranking were compared. Overall, the rankings assessed fell short of all criteria, with greatest strengths in the area of transparency and greatest weaknesses in the area of measuring what matters to the communities they were ranking. The ranking that most closely met the criteria was CWTS Leiden. Scoring poorly across all the criteria were the THE WUR and US News rankings. Suggestions for developing the ranker rating method are described.

CORRESPONDING AUTHOR:
Elizabeth Gadd
Loughborough University, GB
e.a.gadd@lboro.ac.uk

KEYWORDS:
University Rankings; Responsible Research Assessment; Evaluation Framework

TO CITE THIS ARTICLE:
Gadd, E., Holmes, R., & Shearer, J. (2021). Developing a Method for Evaluating Global University Rankings. *Scholarly Assessment Reports*, 3(1): 2, pp. 1–19. DOI: *https://doi.org/10.29024/sar.31*

1 INTRODUCTION

Global university rankings are now an established part of the global higher education landscape. Students use them to help select where to study, faculty use them to select where to work, universities use them to market themselves, funders use them to select who to fund, and governments use them to set their own ambitions. While the international research management community are not always the ones in their institutions that deal directly with the global university ranking agencies, they are one of the groups that feel their effect most strongly. This might be through their university's exclusion from accessing studentship funding sources based on its ranking position; through requests to collect, validate, and optimise the data submitted; or through calls to implement strategies that may lead to better ranking outcomes. At the same time as having to work within an environment influenced by university rankings, the research management community are acutely aware of, and concerned about, the perceived invalidity of the approaches they use.

For this reason, the International Network of Research Management Societies (INORMS) Research Evaluation Working Group (2021) decided to dedicate one of their work-packages to developing a tool by which the relative strengths and weaknesses of global university rankings might be surfaced, and used to influence behavioural change both by ranking agencies and those who rely upon them for decision-making.

2 LITERATURE REVIEW

The practice of ranking universities goes back to the early twentieth century when informal lists of US universities were occasionally published. The first formal and significant ranking, however, was the US News America's Best Colleges published from 1983 (Meredith, 2004), which was followed by national rankings in the UK, Canada and other countries. These national rankings have been subject to a great deal of discussion, including critical assessments of their methodologies by Dichev (2001), Bastedo and Bowman (2010) and Bowden (2000).

The first international, but not global, ranking was published by Asiaweek in 1999 and 2000 (Asiaweek, 2000) and this was followed by the first global ranking published in 2003 by Shanghai Jiao Tong University, the Academic Ranking of World Universities (ARWU). These developments are described in several sources including Usher and Savino (2007), van Raan (2007), Holmes (2010), and Pagell (2014). Since then, a large and varied literature on international rankings has emerged.

The year 2004 saw the appearance of the Ranking Web of Universities, or Webometrics rankings, which at first measured only web activity, and the World University Rankings published by Times Higher Education Supplement (THES) and the QS graduate recruitment firm. Critical accounts of the THES-QS world rankings and their successors can be found in Holmes (2006 & 2015). Other global, regional, subject, and specialist rankings have followed, a list of which was compiled by the Inventory of International Rankings published by the International Ranking Expert Group (IREG) (2021).

In recent years, media and academic interest has shifted to global rankings which have been severely criticised on several grounds within the international research community. General critiques are offered by Usher (2014; 2016), Holmes (2021), Pagell (2019), Bekhradnia (2016), Lee and Ong (2017), and Gadd (2020). Marginson and van der Wende (2007) have argued that they privilege the western model of research-intensive universities with an emphasis on the natural sciences, particularly in English speaking countries. Similarly, it has been noted that rankings favour universities that specialise in certain subjects (Bornmann, De Moya, Anegón & Mutz, 2013).

Other texts claim that rankings promote global elitism by encouraging a shift of investment towards highly ranked universities at the expense of others (Munch, 2014) and intensifying inequality within and between institutions (Cantwell & Taylor, 2013). According to Amsler and Bolsmann (2012) they are part of an exclusionary neoliberal global agenda.

Others have presented evidence that rankings encourage universities to forget their social role or to lose interest in local or regional problems. Lee, Vance, Stensaker, and Ghosh (2020), for example, report that world-class universities are likely to neglect the university third mission

and those that are unranked are more concerned with the local economy and its problems. Stack (2016) has written of the pressures that rankings place on universities in the struggle for resources in a competitive and mediatised global society.

A number of writers have discussed methodological and technical issues. Turner (2005) showed that there are problems with arbitrary weightings and the conflation of inputs and outputs. Others have focussed on the validity of reputational surveys. Safón (2013), for example, has argued that the main factor measured by global rankings is institutional reputation while a study by Van Dyke (2008) noted that academics were likely to rate their own institutions higher than others, a finding that has implications for the validity of the THE reputation survey. An analysis by Safón and Docampo (2020) indicates that reputational bias influences publication data in the Shanghai rankings while Ioannidis et al. (2007) have criticised those rankings and the THES-QS world rankings for measurement error and limited construct validity. Daraio, Bonaccorsi, and Simar (2015) have focussed on problems of monodimensionality, statistical robustness, institutional size, and the inclusion of measures of output and input. Florian (2007) has argued that the Shanghai Rankings cannot be reproduced exactly and therefore are methodologically deficient.

The impact of rankings on university policy has been discussed by Docampo, Egret and Cram (2015) who suggest that they have prompted structural changes, especially in France, that may not be entirely beneficial. Cremonini et al, (2014) claim that the use of rankings to promote world class university programmes may reduce the public benefits of higher education policies.

Aguillo, Bar-Ilan, Levene and Ortega (2010) have compared rankings with different methodologies and noted that while there was a significant similarity between the various rankings it was greater when European universities were considered. A study by Buela-Casal et al (2007) noted significant similarities in various rankings, although not on all indicators. Moed (2017) observed that of the five key university rankings, only 35 institutions appeared in the top 100 of all of them.

Piro and Svertsen (2016) have analysed the reasons why different rankings produce different results. Bookstein, Seidler, Fieder and Winckler (2010) have observed fluctuations in the indicator scores in the THE rankings while Vernon, Balas and Momani (2018) have cast doubt on the ability of rankings to measure and improve research. A longitudinal analysis of the university rankings by Selten et al (2020) suggests that the indicators used do not capture the concepts they claim they measure.

Although the critiques of global rankings are wide-ranging, they have not had much influence on higher education leaders. University administrators have sometimes expressed reservations about rankings but in general have been willing to participate, occasionally breaking ranks if their institutions fall too much (Hazelkorn 2008; 2011). Musselin (2018) explains this phenomenon in terms of university leaders utilising their ranking position as management and legitimisation devices in an increasingly competitive environment.

Some scholars believe that the defects of global rankings far outweigh any possible benefits. Adler and Harzing (2009) have gone so far as to propose a moratorium on ranking and recommend that scholars should "innovate and design more reliable and valid ways to assess scholarly contributions that truly promote the advancement of relevant 21st century knowledge, and likewise recognize those individuals and institutions that best fulfil the university's fundamental purpose."

It must be noted that the academic literature does include attempts to justify the role and methodology of rankings. Sowter (2008), Baty (2013) and Wu and Liu (2017), who are representatives of ranking agencies, have described the rationale behind the various methodologies.

There are others who find some merit in the rankings. Wildavsky (2010) sees the rankings as instrumental in the development of a new academic and scientific global marketplace leading to the spread of new knowledge. Boudard and Westerheijden (2017) describe how the shock of the first Shanghai rankings for continental European universities led to far-reaching structural changes. Rodionov, Fersman and Kushneva (2016) outline how in Russia the rankings are considered an important element in improving international visibility and status. In Taiwan, Shreeve (2020) has suggested that governmental ambitions encouraged by rankings may be beneficial for the institutions concerned by providing a focus for improvement. According to

a case study of the University of Maribor (Rozman & Marhl 2008) they can stimulate "healthy competition". Saisana, D'Hombres and Saltelli (2011) have found that rankings are fairly reliable for macro-regions although less so for comparisons between institutions or countries.

It must be said, however, that some of these studies also observe that the characterisation of a 'top' university as defined by the rankings, and the pursuit of a better ranking position based on developing such characteristics, might not always be locally relevant or ultimately beneficial.

There have been attempts to construct rankings that avoid the various problems and defects that have been identified. Waltman et al. (2012) describe the development of the Leiden Ranking which introduced innovations designed to meet some of the criticism that had been levelled, such as fractional counting and stability intervals. Since then, the Ranking has included data about open access publications and gender equity in publication. U-Multirank was developed with support from the European Commission to provide a user-driven, participatory and multi-dimensional ranking providing features neglected by the dominant rankings (Van Vught and Ziegele, 2012).

Another attempt to reform international rankings was the production of the Berlin Principles on Ranking of Higher Education Institutions in 2006 (IHEP, 2006). These were produced by the International Rankings Expert Group (IREG), which consisted of both rankers and academics, and was founded by the UNESCO European Centre for Higher Education (UNESCO-CEPES) and the Institute for Higher Education Policy. The principles covered the purposes of rankings, the design and weighting of indicators, the collecting and processing of data, and the presentation of results (IHEP 2006). The principles now underpin the criteria used to offer the 'IREG Seal of Approval' to rankings. However, Barron (2017) has questioned the extent to which these rankings meet the Berlin principles and expressed concerns that the principles seek to legitimise ranking practices by attempting to align them with academic values.

The Centre for Science and Technology Studies (CWTS) in Leiden, home of the Leiden Ranking and birthplace of the Leiden Manifesto on the responsible use of research metrics (Hicks, Wouters, Waltman, de Rijcke, and Rafols, 2015), also subsequently developed ten principles for the responsible design, interpretation, and use of university rankings (Waltman, Wouters and van Eck, 2017).

The existence of such principles is a welcome attempt to provide some best practice guidance for the design and use of university rankings. However, the fact that they were influenced and/or developed by university rankers themselves could be seen to affect their neutrality. It is also concerning that the only body currently providing any assessment of university rankers is one where rankers occupy five out of eleven seats on the Executive Committee (IREG, 2021).

It was against this background that the INORMS Research Evaluation Working Group sought to both provide an independent, community-sourced set of best practice criteria against which to assess the global university rankings and then to identify the extent to which a sample of rankings met those criteria.

3 METHODS

In parallel with the INORMS REWG's work to rate the global university rankings, the group also developed a framework for responsible research evaluation, called SCOPE (Himanen and Gadd, 2019). This is a five-stage process by which evaluations can be designed to consistently adhere to best practice in research assessment. In order to develop a responsible approach to evaluating the global university rankings, the SCOPE framework was adopted as follows.

3.1 START WITH WHAT YOU VALUE

The 'S' of SCOPE states that prior to any evaluation attempt there needs to be a clear articulation of what is valued about the entity under evaluation from the perspective of the evaluator and the evaluated. To this end the group undertook a literature search to develop a draft set of best practice criteria for fair and responsible university rankings. These were circulated to the international research evaluation community for comment via the INORMS REWG, LIS-Bibliometrics, and INORMS member organisation circulation lists such as that of the UK Association of Research Mangers and Administrators (ARMA) Research Evaluation

Special Interest Group as well as via various international research management conferences. Responses were received from a broad cross-section of the community: universities, academics, publishers, and ranking agencies, which were then worked into the final set of criteria (see section 4). The criteria were grouped into four common themes: good governance, transparency, measure what matters, and rigour.

It could be argued that some of the criteria are challenging to meet, especially for some of the commercial ranking agencies, for example, around conflicts of interest. However, it was felt to be important to remain true to the values of the community even where they were aspirational. As noted by Gadd and Holmes (2020) on the publication of the ratings, "just because something is difficult to achieve, doesn't mean we shouldn't aspire to it". The benefit of taking a value-led approach, such as that promoted by SCOPE, is that the evaluation is driven by what the community cares about, rather than by what might be possible or practical. Indeed, it could equally be argued that if it is not possible to rank organisations in accordance with the best practice principles developed by the communities being ranked, perhaps it is the rankings that should change, not the principles.

3.2 CONTEXT CONSIDERATIONS

The 'C' of the SCOPE framework states the importance of considering the evaluative context - why and who you are evaluating - prior to the evaluation. The purpose of this evaluation was to highlight the extent to which various ranking agencies adhered to the best practice expectations of the wider research evaluation community, to expose their relative strengths and weaknesses, with the ultimate purpose of incentivising them to address any deficiencies. By clarifying the context, the group moved away from early thoughts of 'ranking the rankings', recognising that this might only lead to self-promotion by the top-most ranked, rather than behaviour change.

As the work was undertaken by volunteers, it was not possible to assess all the global university rankings, so it was decided to test the model on six of the largest and most influential university rankings to provide a proof-of-concept. This group was selected by consulting with members of the INORMS REWG as to the most frequently used rankings in their region, and is not a reflection on their quality. The final list included *ARWU*, *CWTS Leiden*, *QS World University Rankings (QS WUR)*, Times Higher Education World University Rankings (*THE WUR*), *U-Multirank*, and *US News & World Report* Best Global Universities. Although many of these ranking agencies produce more than one ranking, it was decided to focus on their flagship 'overall' global ranking product for this evaluation, as these are the rankings most commonly used and cited.

3.3 OPTIONS FOR EVALUATING

Having established our values and context, the SCOPE framework's 'O' - options for evaluating - were then considered. To run the assessment, the criteria collected at the 'values' stage were translated into assessable indicators that were felt to be suitable proxies for the criteria being assessed. As the group were seeking to assess qualities rather than quantities, it was felt to be important to provide assessors with the opportunity to provide qualitative feedback in the form of free-text comments, as well as scores on a three-point scale according to whether the ranker fully met (2 marks), partially met (1 mark), or failed to meet the set criteria (0 marks).

To ensure transparency and mitigate against bias, twelve international experts were identified and invited by members of the INORMS REWG to provide a review of one ranking agency. Due to the pandemic only eight were able to provide a rating.

INORMS REWG members also undertook evaluations, and, in line with the SCOPE principle of 'evaluating with the evaluated,' each ranker was also invited to provide a self-assessment in line with the community criteria. Between one and four reviews were received for each ranking. Only one ranking agency, CWTS Leiden, accepted the offer to self-assess, providing free-text comments only.

The reviews were then forwarded to a senior expert reviewer, Richard Holmes, author of the University Ranking Watch blog (Holmes, 2021). He was able to combine the feedback from our international experts with his own detailed knowledge of the rankings supplemented by intelligence sourced from conferences and online communications, to enable a robust, expert

assessment. In cases where a question was interpreted differently by reviewers, he used his judgement to decide on the most appropriate interpretation and score.

3.4 PROBE DEEPLY

The 'P' of the SCOPE framework represents 'probe' and requires that any evaluative approach is examined for discriminatory effects, gaming potential and unintended consequences. We observed some criteria where rankings might be disadvantaged for good practice, for example where a ranking did not use surveys and so could not score. This led us to introduce a 'Not Applicable' category to ensure they would not be penalised.

It was also thought to be important that we did not replicate the rankings' practice of placing multi-faceted entities on a single scale labelled 'top'. Not only would this fail to express the relative strengths and weaknesses of each ranking, but it would give one ranking agency 'boasting rights' which would run counter to what we were trying to achieve.

3.5 EVALUATE

The 'E' of SCOPE invites assessors to both evaluate and evaluate their evaluation. The ranker assessment generated many learning points discussed in section 5 below which fed into recommendations for the revision of the ranking assessment tool.

4. FINDINGS

The full set of attributed ranking reviews and the final calibrated review have been made openly available (INORMS, 2020). Intra-class Correlation Coefficients were calculated for the each set of reviews (***Table 1***) which indicate moderate to good inter-rater reliability (Koo and Li, 2016). Some reflections as to how these might be improved, including clearer definitions, are provided in Section 5.

Table 1 Review volume and reliability.

UNIVERSITY RANKING	NUMBER OF REVIEWS (INCLUDING CALIBRATION)	INTRA-CLASS CORRELATION CO-EFFICIENT
ARWU	5	0.662
CWTS Leiden	4	0.862
QS	4	0.604
THE WUR	3	0.725
U-Multirank	2	0.663
US World News	2	0.725

4.1 GOOD GOVERNANCE

The five key expectations of rankers with regards to good governance were that they engaged with the ranked, were self-improving, declared conflicts of interest, were open to correction and dealt with gaming. The full criteria and indicators are listed in ***Table 2***.

Table 2 Criteria and indicators – Good Governance.

A	CRITERIA: GOOD GOVERNANCE
A1	Engage with the ranked. Has a clear mechanism for engaging with both the academic faculty at ranked institutions and their senior managers, for example, through an independent international academic advisory board, or other external audit mechanisms.
A1.1	Does the ranker have an independent international academic advisory board, that is transparent and representative?
A1.2	Does the ranker have other mechanisms by which they engage with the ranked, e.g, summits, non-transparent consultations, etc.
A2	Self-improving. Regularly applies measures of quality assurance to their ranking processes.
A2.1	Is there evidence that they are identifying problems with their own methodologies, and improving them?

(Contd.)

A	CRITERIA: GOOD GOVERNANCE
A3	Declare any conflict of interests. Provides a declaration of potential conflicts of interest as well as how they actively manage those conflicts.
A3.1	Does the ranking make any reference to conflicts of interest on its web site?
A3.2	If yes, does the declaration outline how they actively manage those conflicts of interests.
A3.3	Does the ranking avoid selling services or data to support HEIs in improving their ranking position?
A3.4	If so, are the potential conflicts of interest here declared.
A4	Open to correction. Data and indicators should be made available in a way that errors and faults can be easily corrected. Any adjustments that are made to the original data and indicators should be clearly indicated.
A4.1	Do HEIs get a chance to check the data on themselves before being used for ranking purposes?
A4.2	Is there a clear line of communication through which ranked organisations can seek to correct any errors?
A4.3	Are corrected errors clearly indicated as such?
A5	Deal with gaming. Has a published statement about what constitutes inappropriate manipulation of data submitted for ranking and what measures will be taken to combat this.
A5.1	Does the ranking have a published statement about what constitutes inappropriate manipulation of data submitted for ranking?
A5.2	Does the statement outline what measures will be taken to combat this?
A5.3	Are those measures appropriate?

4.1.1 Engaged with the ranked

One of the SCOPE principles is to evaluate with the evaluated, and the community felt that having continued engagement with both the faculty and leadership of organisations that they ranked was an important activity. The rankings tended to score well here with most having advisory boards, and all engaging in some form of outreach activity.

4.1.2 Self-improving

One of the biggest concerns about the rankings is their methodological imperfections. This question sought to highlight that ongoing improvement was an essential activity for ranking agencies. Again, all rankers either fully or partially met this criterion.

4.1.3 Declare any conflicts of interests

There was a belief amongst the community that ranking agencies should remain independent in order to fairly rank universities. As such, where there were conflicts of interest, i.e., where rankers sold their data or provided consultancy services to institutions with the ability to pay for it, this should be declared. No ranker fully met these expectations, and all received at least one zero in this section.

4.1.4 Open to correction

The community felt that an important aspect of good governance was that any errors drawn to ranking agencies' attention should be corrected and clearly indicated as such. Where data was drawn entirely from third parties it was felt that this criterion was not applicable. In all other cases, HEIs were given some opportunity to check the data prior to the ranking being compiled. In most cases there was some line of communication by which HEIs could notify ranking agencies of errors, but only CWTS achieved full marks for clearly listing corrected errors.

4.1.5 Deal with gaming

The rewards associated with a high ranking position are such that 'gaming' is a regular feature (Calderon, 2020). The community were concerned that ranking agencies recognised this and took steps to address gaming where it was drawn to their attention. Where third-party data was used, again, this was not thought to be an applicable criterion. Other ranking agencies, all made some effort in this space, with full or partial compliance.

4.2 TRANSPARENCY

Transparency was very important to the community with many respondents making reference to the 'black box' nature of many rankings' approaches. The five expectations of rankers here were that they had transparent aims, methods, data sources, open data and financial transparency. The full criteria and indicators are listed in ***Table 3***.

Table 3 Criteria and indicators – Transparency.

B	CRITERIA: TRANSPARENCY
B1	Transparent aims. States clearly the purpose of the ranking, what it seeks to measure, and their target groups.
B1.1	Does the agency clearly state the ranking's purpose, what it seeks to measure, and its target groups?
B2	Transparent methods. Publishes full details of their ranking methodology, so that given the data a third party could replicate the results.
B2.1	Does the agency publish full details of their ranking methods? Including weightings, surveys, recruitment, etc.
B2.2	Does the ranking's website provide clear definitions of all the indicators used? E.g., what constitutes a publication.
B2.3	Does the ranking's website clearly state how universities are defined and whether they include off-shore campuses, and teaching hospitals in their definition. (You may wish to check a known University with off-shore campus and teaching hospital, e.g., University of Nottingham)
B2.4	Could a third party with access to the data replicate the results?
B3	Transparent data availability. Provides detailed descriptions of the data sources being used, inclusion and exclusion parameters, date data snapshots were taken, and so on.
B3.1	Has the agency described in detail the sources of the data used to calculate the rankings?
B3.2	Is the data described fully, i.e. inclusion and exclusion parameters, date data snapshots were taken, format of the data, the quantity of data (for example, the number of records and fields in each table), the identities of the fields, and any other surface features?
B3.3	Does the agency provide clear opportunities for errors to be corrected?
B4	Open data. Makes all data on which the ranking is based available in an open standard non-proprietary format and, where possible, use open standard definitions and classifications (e.g. for subjects, publication types, etc.) to aid interoperability and comparability, and so that those being evaluated can verify the data and analysis.
B4.1	Does the agency provide access to the data to all institutions/researchers being ranked?
B4.2	Does the agency provide access to the data to anybody?
B4.3	If so, is the data available in a non-proprietary format?
B5	Financially transparent. Publishes details of all sources of income from consultancy services, training, events, advertising, and so on including financial outgoings, e.g. sponsorships.
B5.1	Does the agency publish all details of income sources arising from consultancy services, training, events, advertising, etc. it may obtain from the institutions/researchers being ranked?

4.2.1 Transparent aims

All rankers were either fully or partially transparent about the aims of their ranking and its target groups. Of course, transparency about their aims is not the same as successfully meeting them.

4.2.2 Transparent methods

The requirement of transparent methods was particularly important to the research management community, as many are asked to reverse engineer their institution's ranking position and make predictions about future performance. Whilst most rankers fully met expectations around publishing their methods and indicators, in only one case (ARWU) was it thought to be possible for a third-party with access to the data to be able to replicate the results.

4.2.3 Transparent data availability

Questions around data availability required rankers to describe both their sources and their parameters in detail, with a specific question regarding the ability to correct data. Again, all rankers fully or partially met these criteria.

4.2.4 Open data

In addition to data being fully described, it was felt to be important that this was also openly available for the community to scrutinise and work with. Only ARWU received full marks on this, with other rankings making some data available.

4.2.5 Financially transparent

As with the declaration of conflicts of interest, the community were keen that ranking agencies were financially transparent, revealing sources of income. Only U-Multirank fully met this criterion, with four out of the remaining five failing to meet it.

4.3 MEASURE WHAT MATTERS

The five expectations of rankers here were that they drove good behaviour, measured against mission, measured one thing at a time (no composite indicators), tailored results to different audiences and gave no unfair advantage to universities with particular characteristics. The full criteria and indicators are listed in ***Table 4***.

Table 4 Criteria and indicators – Measure what matters.

C	CRITERIA: MEASURE WHAT MATTERS
C1	Drive good behaviour. Seeks to enhance the role of universities in society by measuring what matters, driving positive systemic effects and proactively seeking to limit any negative impacts such as over-reliance on rankings for decision-making.
C1.1	Does the ranking provide clear warnings on their website about the limitations of using rankings for decision-making?
C1.2	Does the ranking run any promotional campaigns around the limitations of rankings?
C1.3	Does the ranking measure a university's approach to equality, diversity, sustainability, open access or other society-focussed agendas?
C2	Measure against mission. Accepts that different universities have different characteristics – mission, age, size, wealth, subject mix, geographies, etc, and makes visible these differences, so that universities can be clustered and compared fairly.
C2.1	Does the ranker avoid offering one single over-arching ranking that claims to assess the 'top' universities?
C2.2	Does the ranking provide a facility by which institutions can be compared to others that share their mission?
C2.3	Does the ranking provide a facility by which institutions can be compared to others within the same subject area?
C2.4	Does the ranking provide contextual, qualitative and quantitative information on each of the institutions they rank?
C3	One thing at a time. Does not combine indicators to create a composite metric thus masking what is actually being measured.
C3.1	Does the ranking AVOID combining indicators to create a composite metric?
C4	Tailored to different audiences. The ranking provides different windows onto the data that may be relevant to different audiences. For example, by providing an opportunity to focus in on teaching elements for students.
C4.1	Does the ranking offer the audience the opportunity to weight the indicators in accordance with their preferences?
C4.2	Does the ranking offer different 'windows' onto the same data for different audiences?
C5	No unfair advantage. Makes every effort to ensure the approach taken does not discriminate against organisations by size, disciplinary mix, language, wealth, age and geography.
C5.1	Does the ranking only use data sources that offer equal global representation?
C5.2	Does the ranking only use indicators that offer institutions of all sizes, equal opportunity to succeed?
C5.3	Does the ranking only use indicators that offer institutions of all disciplinary mixes, equal opportunity to succeed?
C5.4	Does the ranking only use indicators that offer institutions where English is not the primary language, equal opportunity to succeed?

4.3.1 Drive good behaviour

With widely acknowledged limitations of university rankings, the community felt it was important that ranking agencies themselves did their best to highlight this on their products. CWTS Leiden and U-Multirank clearly did so; ARWU and US News did not, and QS and THE made some reference to it which was felt to be undermined by their repeated reference to their rankings being 'trusted' or 'excellent' sources.

4.3.2 Measure against mission

Whilst universities largely seek to offer teaching and research in some form, their missions and other characteristics such as size and wealth, are hugely varied. The community felt it was important that rankers provided a facility by which institutions could be compared to others with similar characteristics rather than grouping all together on a single scale. Only U-Multirank and CWTS Leiden avoided offering one single over-arching ranking that sought to identify the 'top' universities, with only U-Multirank providing a facility by which rankers could be compared with those sharing their mission. All provided subject-based comparisons and most provided some qualitative data on the organisations being ranked.

4.3.3 One thing at a time

A related criterion to that specifying that rankers should avoid using a single scale of excellence, was that they avoided composite metrics that used pre-set weightings regardless as to whether institutions weighted their focus on the same way. Again, only CWTS Leiden and U-Multirank avoided composite metrics, thus achieving full marks.

4.3.4 Tailored to different audiences

Recognising that rankings are used by different audiences for different purposes, the community felt it important that the ranking data collected was delivered in different formats according to the interests of these different audiences. No ranking fully met expectations here (although through the avoidance of composite indicators, this was not thought to be an applicable question for CWTS Leiden). Most others scored poorly.

4.3.5 No unfair advantage

Whilst living in an 'unfair' world, it was still felt to be important to avoid offering an unfair advantage to universities with particular characteristics (size, discipline, geography and English language-use) as far as possible. While it was felt that all rankings made some effort in this space, none scored full marks.

4.4 RIGOUR

The five expectations of rankers in this section were around rigorous methods, no 'sloppy' surveys, validity, sensitivity and honesty about uncertainty. The full criteria and indicators are listed in ***Table 5***.

Table 5 Criteria and indicators – Rigour.

D	CRITERIA: RIGOUR
D1	Rigorous methods. Data collection and analysis methods should pass tests of scientific rigour, including sample size, representation, normalisation, handling of outliers, etc.
D1.1	Does the ranking transparently normalise indicators, in a robust way, for field?
D1.2	Does the ranking clearly state how it handles outliers, and is this fair?
D2	No sloppy surveys. Limit use of unverifiable survey information and ensures that where they are used that the methods are sound and unbiased, e.g. samples are large, representative and randomly selected; questions are reliability-tested and measure what they seek to measure.
D2.1	Does the ranking AVOID use opinion surveys to elicit reputational data?
D2.2	Where a ranking uses surveys are large samples used?
D2.3	Where a ranking uses surveys are random samples used?
D2.4	Where a ranking uses surveys are representative samples used?

(Contd.)

D	CRITERIA: RIGOUR
D2.5	Where a ranking uses surveys are questions reliability tested?
D2.6	Where a ranking uses surveys are the questions valid?
D3	Validity. Indicators have a clear relationship with the characteristic they claim to measure. For example, teaching quality should not solely be indicated by staff-student ratios.
D3.1	Do indicators have a clear relationship with the characteristic they claim to measure?
D4	Sensitivity. Indicators are sensitive to the nature of the characteristic they claim to measure.
D4.1	Does the ranking AVOID include monotonic indicators for which a good value will depend on the mission of the university, e.g., staff-student ratio; international-non-international staff ratio.
D4.2	Are ranking results relatively stable over time? E.g., are improvements in rank likely to reflect true improvements in University performance?
D5	Honest about uncertainty. The types of uncertainty inherent in the methodologies used, and of the data being presented should be described, and where possible, clearly indicated using error bars, confidence intervals or other techniques, without giving a false sense of precision.
D5.1	Does the ranking website provide any commentary on the limitations and uncertainties inherent within their methodologies?
D5.2	Does the ranking provide error bars or confidence intervals around the indicators provided?

4.4.1 Rigorous methods

A common complaint regarding the use of rankings by scientific organisations is that they use methods that those organisations would not consider valid in their own practices. Two questions around field normalisation and the handling of outliers yielded mixed results with some efforts around both but very few exemplars of best practice.

4.4.2 No 'sloppy' surveys

While the community were not against survey methodologies per se, there was a strong sense that the rankers use of surveys was problematic with questionable practices employed around samples and question choice. ARWU and CWTS Leiden avoided using surveys altogether thus achieving full marks on this criterion. Whilst sample sizes tended to be large, they were rarely random and not always thought to be representative. There was no evidence of reliability testing on questions, nor that the questions were entirely valid.

4.4.3 Validity

In any evaluation approach it is important that the indicators used are a valid enough proxy for the quality being measured. Only CWTS Leiden and U-Multirank were thought to fully meet this requirement, with the other rankers making some efforts but falling short of expectations.

4.4.4 Sensitivity

Gingras (2014) has noted the importance of avoiding monotonic indicators in evaluation approaches where a 'good' score will vary according to the mission or disciplinary mix of the organisation (e.g., staff:student ratios). He also highlighted the importance of evaluation outcomes varying only in accordance with real (and often slow-moving) changes within those organisations. Only CWTS Leiden scored full marks on these indicators, with others demonstrating less, or no, compliance.

4.4.5 Honest about uncertainty

In any evaluative data there are always going to be levels of uncertainty around the confidence in which the results can be relied on. The community were keen that confidence levels made visible the relatively small differences between those organisations at different ranking positions. Again, only CWTS Leiden clearly expressed the limitations around their methodologies and provided stability indicators for their rankings. Others made some efforts with regards to the former but failed to score on the latter.

4.5 SUMMARY

Figure 1 illustrates the relative strengths and weaknesses of each global university ranking by dividing their actual scores for each section into the total possible score they could have achieved, having removed all 'not applicable' criteria. It shows that in terms of good governance, the QS achieved the highest scores, closely followed by U-Multirank and CWTS Leiden. In terms of transparency, ARWU and CWTS Leiden performed the best, again with U-Multirank close behind. The strongest performer on 'measure what matters' was U-Multirank with CWTS Leiden closely following. Finally, in terms of rigour, CWTS Leiden was significantly stronger than all the other rankings. Scoring consistently poorly across all the community-developed criteria were the THE WUR and US News rankings. ***Figures 2-7*** provide a more granular picture for each of the rankings assessed by looking at their scores on each of the twenty criteria. Where a criterion was not applicable it was removed from the chart and indicated as such.

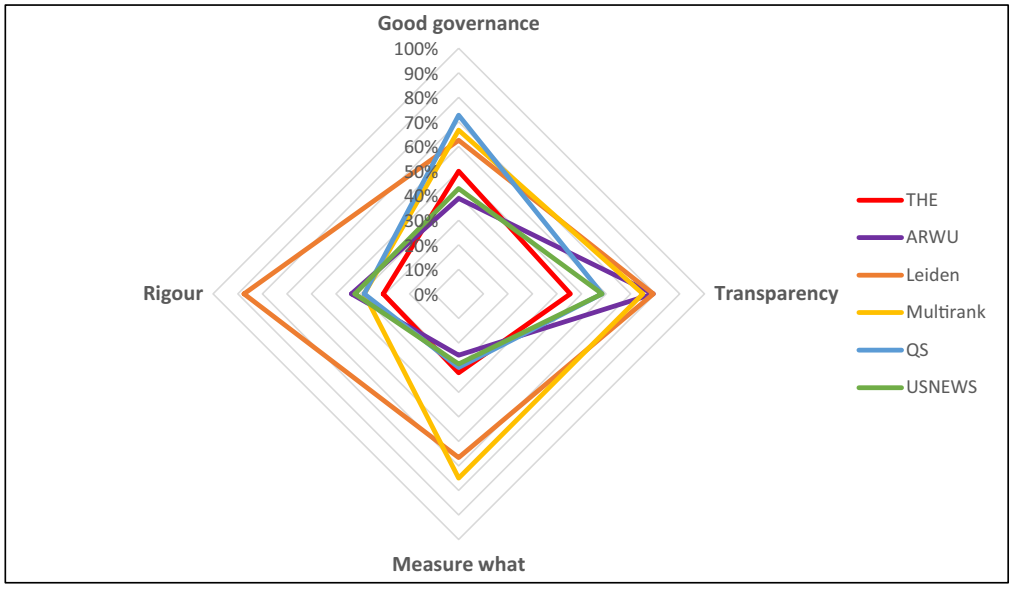

Figure 1 Spidergram illustrating the actual scores/ total possible score for each global ranking.

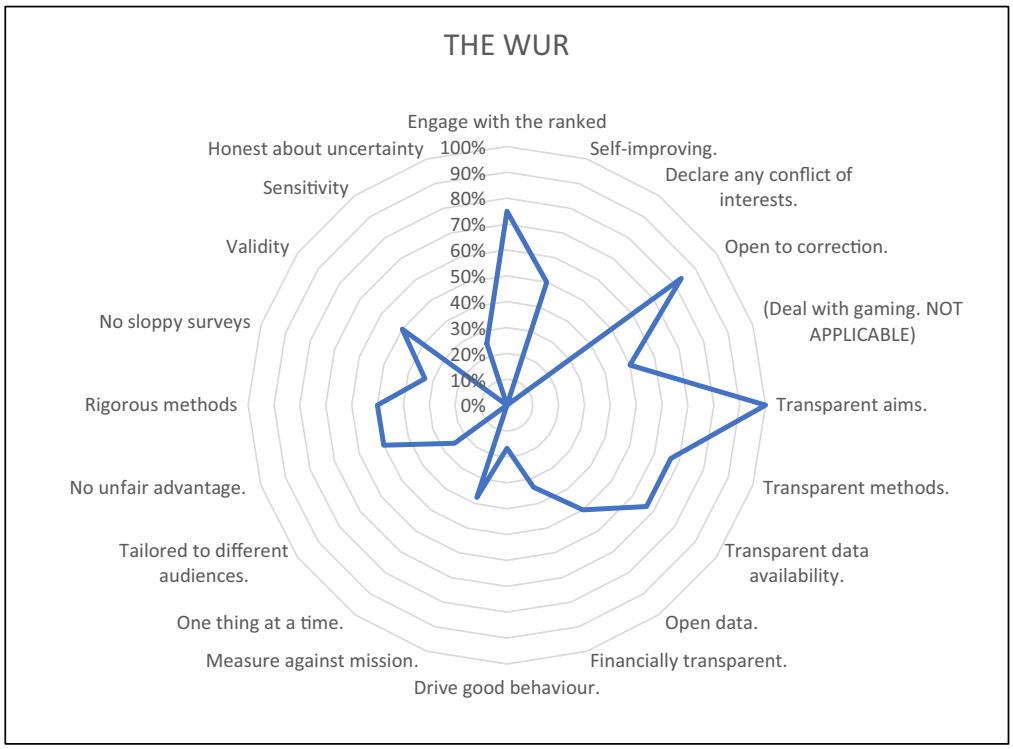

Figure 2 Spidergram illustrating THE WUR scores on all twenty criteria.

Whilst the rankings that score better on these indicators may feel pleased with their performance, it is important to note that the community expectations are set at 100% adherence to the criteria. The closest any of these ranking agencies came to that was CWTS Leiden which scored 100% on nine of the eighteen criteria deemed to be applicable to them. Indeed, when you

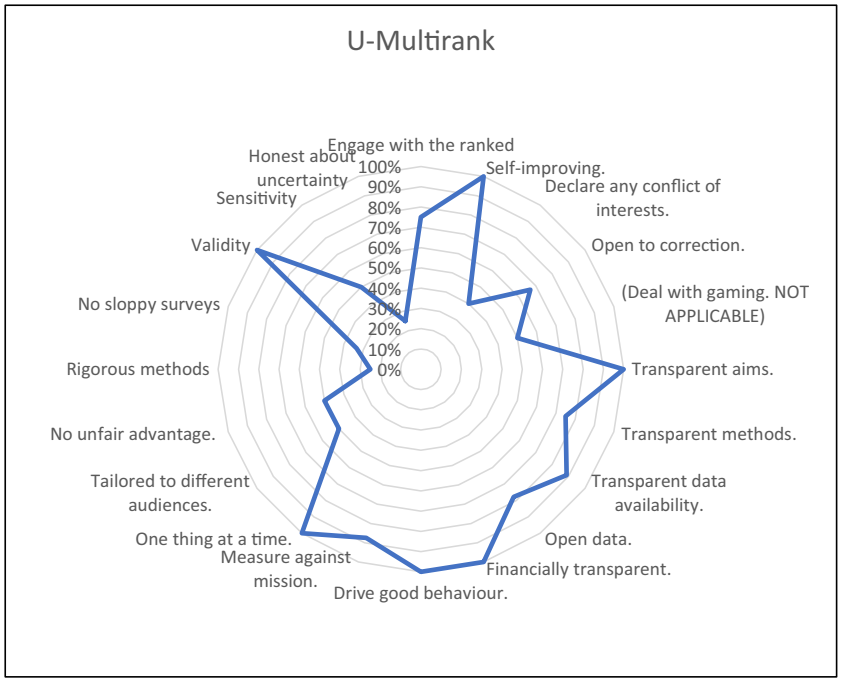

Figure 3 Spidergram illustrating U-Multirank scores on all twenty criteria.

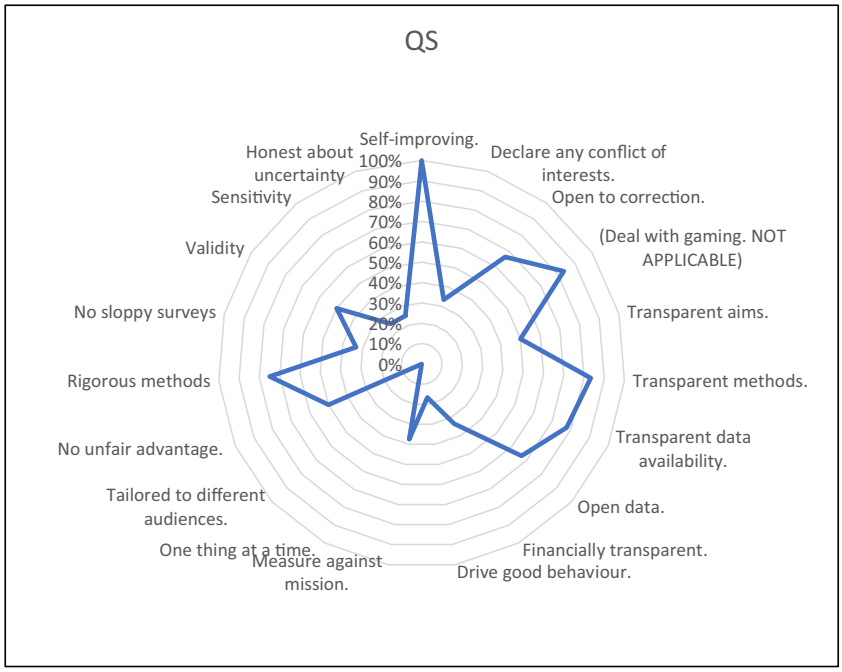

Figure 4 Spidergram illustrating QS scores on all twenty criteria.

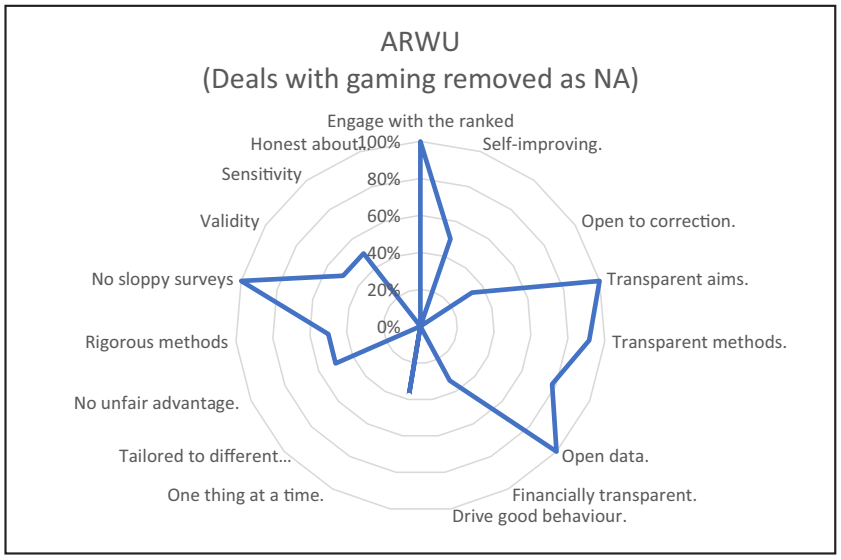

Figure 5 Spidergram illustrating ARWU scores on all nineteen applicable criteria.

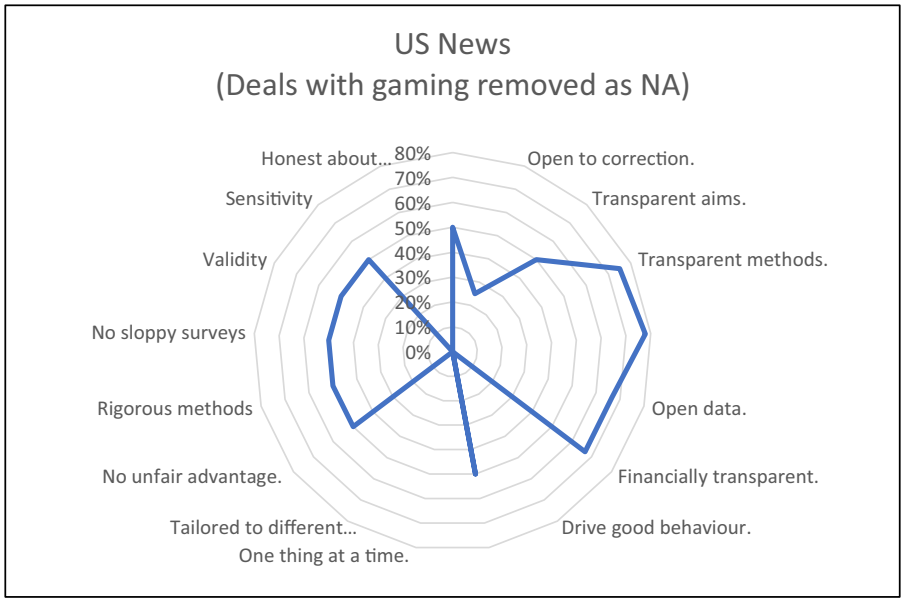

Figure 6 Spidergram illustrating US News scores on all nineteen applicable criteria.

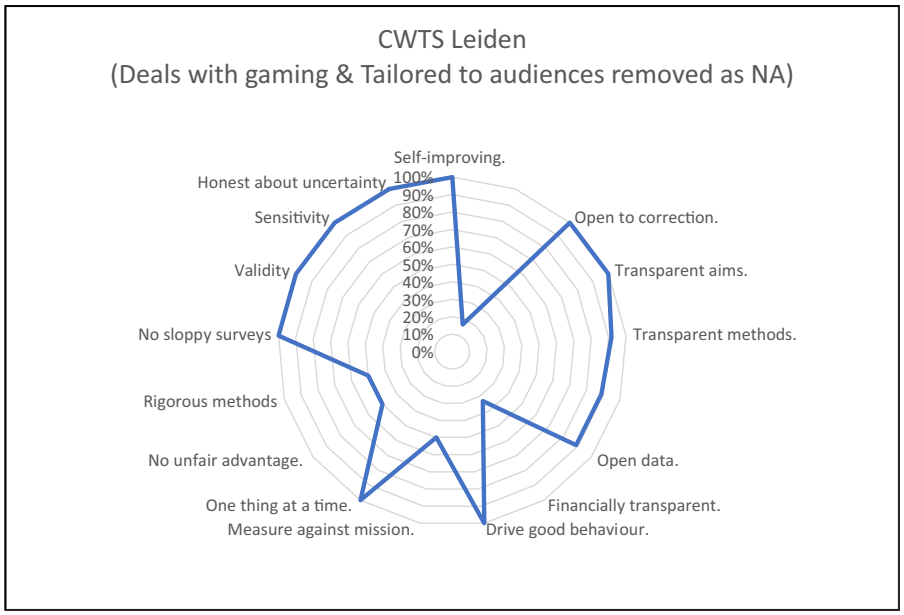

Figure 7 Spidergram illustrating CWTS Leiden scores on all 18 applicable criteria.

look at the average scores for all six rankers across the four criteria (***Figure 8***) you can see that overall, the ranking sector falls considerably short of all criteria, with the greatest strengths in terms of transparency and the greatest weaknesses in terms of measuring what matters to the communities they are ranking.

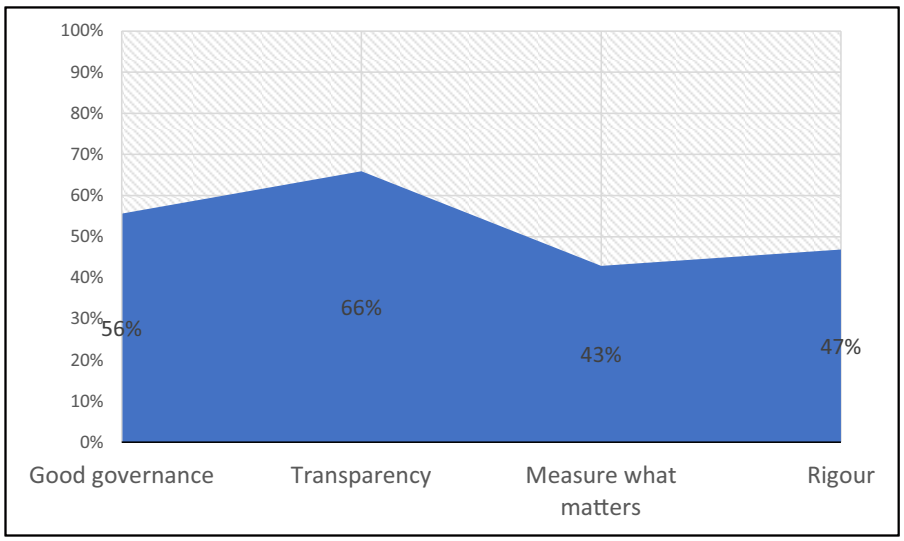

Figure 8 Average scores of all six ranking agencies on the four criteria.

5 REVISED RANKING ASSESSMENT TOOL

The process of piloting the ranker rating tool surfaced many helpful learning points, including feedback from one the ranking agencies under assessment, CWTS Leiden, which it would be useful to incorporate into any future iteration. These are outlined below.

5.1 START WITH WHAT YOU VALUE

By putting out a 'straw person' list of draft criteria for fair and responsible university rankings and inviting free-form feedback, useful input was received. However, all the criteria were given equal weight in the resulting assessment tool. This may not reflect community expectations, with some criteria being of paramount importance and other criteria holding less importance. Future iterations of the assessment tool may wish to revisit both the chosen criteria and the resulting weightings via some kind of survey instrument. A survey may have the benefit of reaching a wider audience through the relative ease of completion and may enable a more nuanced assessment of ranking agencies.

5.2 CONTEXT CONSIDERATIONS

The selection of ranking agencies and the focus on their flagship ranking for this pilot was a pragmatic choice due to time constraints and the need to recruit reviewers. However, to provide a more complete assessment of a much wider range of the increasing number of global rankings, in an ideal world this would be extended both to additional rankers and additional rankings (e.g., subject rankings).

Due to the impact of COVID-19 on workloads and the resulting availability of expert reviewers, some rankers only received one expert review and one senior expert reviewer calibration in this exercise. In an ideal world each ranking would receive a minimum of two expert reviews plus calibration. Even better, in line with the principle of evaluating with the evaluated, each ranking agency would submit a self-assessment to fill in any gaps not publicly available, or not known to the expert reviewers. If such assessments grow in popularity and visibility, it may be that more ranking agencies become willing to provide a self-assessment to make the case for their activities.

5.3 OPTIONS FOR EVALUATING

It was noted by reviewers that multi-part questions such as D1.2 "Does the ranking clearly state how it handles outliers and is this fair?" were difficult to assess. In future, such questions should be split into two. The other over-arching recommendation is that a more granular scoring system, perhaps across a five-point scale, would allow for fairer assessment. In the current exercise the use of 'partially meets' covered a whole range of engagement with the stated criterion, from slightly short of perfection to a little better than fail.

There were also some issues with particular questions as outlined below.

A1 Engage with the ranked. *Rankers should score less well if their engagement activity was simply marketing and promotion.*

B1 Transparent aims. *Rankers should score less well if they make claims to identify the 'best' or 'top' institutions.*

B2 Transparent methods. *Rankers should score less well if their methods were not transparent about their normalisation mechanisms.*

B3 Transparent data availability. *Remove question B3.3 (Does the agency provide clear opportunities for errors to be corrected?) because it is very similar to A4.3 (Are corrected errors clearly indicated as such?).*

B4 Open data. *Reward ranking agencies for the use of open standards for making data openly available.*

C5 No unfair advantage. *It was felt that such requirements were impossible to meet even by the best-intentioned of rankers due to the inequalities inherent in society. It was therefore proposed to retain the heading 'no unfair advantage' as an important principle, but to amend the sub-criteria to reward those that seek to reduce disadvantage along these lines.*

D1 Rigorous methods. *On criterion D1.1, CWTS Leiden have argued that normalization may not always be appropriate and therefore where a ranking can justify their normalization decisions this should give them full marks. For example, indicators of gender balance should not be normalized by field, but by representation in the global population.*

On criterion D5.2 (Does the ranking provide error bars or confidence intervals around the indicators provided?) it was pointed out that error bars may in fact introduce a false sense of certainty about the level of uncertainty due to the challenges of properly quantifying error.

Other questions that might be useful to include would related to the user-friendliness of the ranking web page, perhaps under C4 'Tailored for different audiences', and the number of universities included in the ranking, perhaps under C5 'No unfair advantage'.

6 CONCLUSIONS

Global ranking agencies have a significant influence on the strategic and operational activities of universities worldwide, and yet they are unappointed and unaccountable. As a research management community we believe that there is a strong argument for providing an open and transparent assessment of the relative strengths and weaknesses of the global university rankings to make them more accountable to the higher education communities being assessed. We believe that the approach described in this report, as refined, offers a fair and transparent tool for running such assessments.

The findings of this exercise highlight that those rankings that are closest to the universities being assessed, the CWTS Leiden Ranking run by a university research group, and U-Multirank run by a consortium of European Universities, tended to better meet the community's expectations of fairness and responsibility. Unfortunately, those rankings that are more highly relied upon by decision-makers, such as the Times Higher Education World University Ranking and the US News and World Report ranking, tended to score less well. However, all rankings fell short in some way and this work highlights where they might focus their attention.

One of the challenges of short-term project-based work such as this is long-term sustainability and influence, options for which are now being explored by the INORMS REWG. As well as drawing this work to the attention of ranking agencies, it needs to also reach those relying on rankings data for decision-making. This is also one of the next steps for the group. Overall, this work has been warmly welcomed by the HE community, and by some of the rankers assessed. We hope that the next iteration of this tool, revised in line with our recommendations, will play a formative role in improving the design of university rankings and limiting their unhelpful impacts on the HE community.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **INORMS REWG Ranker Ratings Data.** Qualitative and quantitative ranker ratings.
  https://doi.org/10.29024/sar.31.s1

ACKNOWLEDGEMENTS

The authors should like to acknowledge the ranker ratings provided by the INORMS Research Evaluation Working Group: Laura Beaupre (University of Guelph), Lone Bredahl Jensen (Southern Denmark University), Laura Himanen (Tampere University), Aline Rodrigues (Sociedade Beneficente Israelita Brasileira Hospital Albert Einstein), Hirofumi Seike (Tohoku University), Justin Shearer (University of Melbourne), Tanja Strom (Oslo Metropolitan University) Baron Wolf (University of Kentucky), and Baldvin Zarioh (University of Iceland), and the Expert Reviewers Group: Stephen Curry (Imperial College, London), Markku Javanainen (University of Helsinki), Kristján Kristjánsson (Rekyavik University), Jacques Marcovitch (Universidade de São Paulo (USP)), Maura McGinn (University College Dublin), Cameron Neylon (Curtin University), Nils Pharo (Oslo Metropolitan University), and Claus Rosendal (Southern Denmark University).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Elizabeth Gadd *orcid.org/0000-0003-4509-7785*
Loughborough University, GB

Richard Holmes *orcid.org/0000-0002-3235-8029*
University Ranking Watch, MY

Justin Shearer *orcid.org/0000-0002-5653-850X*
University of Melbourne, AU

REFERENCES

Amsler, S., & **Bolsmann, C.** (2012). University ranking as social exclusion. *British Journal of Sociology of Education, 33*(2), 283–301. DOI: *https://doi.org/10.1080/01425692.2011.649835*

Adler, N. J., & **Harzing, A.** (2009). When knowledge wins: Transcending the sense and nonsense of academic rankings. *Academy of Management Learning and Education 8*(1), 72–95. DOI: *https://doi.org/10.5465/amle.2009.37012181*

Aguillo, I. F., Bar-Ilan, J., Levene, M., & **Ortega, J. L.** (2010). Comparing university rankings. *Scientometrics, 85*(1), 243–256. DOI: *https://doi.org/10.1007/s11192-010-0190-z*

Asiaweek. (2000). Asia's Best Universities 2000. Retrieved from *Asiaweek.com* | Asia's Best Universities 2000 | Overall Ranking (*cnn.com*).

Barron, G. R. S. (2017). The Berlin principles on ranking higher education institutions: limitations, legitimacy, and value conflict. *Higher Education, 73*(2), 317–333. DOI: *https://doi.org/10.1007/s10734-016-0022-z*

Bastedo, M. N., & **Bowman, N. A.** (2010). The U.S. News and World Report college rankings: Modeling institutional effects on organizational reputation. *American Journal of Education, 116*(2), 163–183. DOI: *https://doi.org/10.1086/649437*

Baty, P. (2013). The Times Higher Education World University Rankings, 2004–2012. *Ethics in Science and Environmental Politics, 33*(2), 125–130. DOI: *https://doi.org/10.3354/esep00145*

Bekhradnia, B. (2016). International university rankings: For good or ill? Oxford: HEPI. Retrieved 09 January 2021 *www.hepi.ac.uk/2016/12/15/3734/*

Bookstein, F. L., Seidler, H., Fieder, M., & **Winckler, G.** (2010). Too much noise in the Times Higher Education rankings. *Scientometrics, 85*(1), 295–299. DOI: *https://doi.org/10.1007/s11192-010-0189-5*

Bornmann, L., De Moya Anegón, F., & **Mutz, R.** (2013). Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? *Journal of the American Society for Information Science and Technology, 64*(11), 2310–2316. DOI: *https://doi.org/10.1002/asi.22923*

Boudard, E., & **Westerheijden, D.** (2017). France: Initiatives for excellence. In *Policy analysis of structural reforms in higher education*. London: Palgrave Macmillan. DOI: *https://doi.org/10.1007/978-3-319-42237-4_8*

Bowden, R. (2000). Fantasy higher education: University and College league tables. *Quality in Higher Education, 6*(1), 41–60. DOI: *https://doi.org/10.1080/13538320050001063*

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M., & **Vadillo-Muñoz, O.** (2007). Comparative study of international academic rankings of universities. *Scientometrics, 71*(3), 349–365. DOI: *https://doi.org/10.1007/s11192-007-1653-8*

Calderon, A. (2020). 'New rankings results show how some are gaming the system'. *University World News*, 12 June 2020. *https://www.universityworldnews.com/post.php?story=20200612104427336*

Cantwell, B., & **Taylor, B.** (2013). Global Status, Intra-Institutional Stratification and Organizational Segmentation: A Time-Dynamic Tobit Analysis of ARWU Position Among U.S. Universities. *Minerva, 51*(2), 195–223. DOI: *https://doi.org/10.1007/s11024-013-9228-8*

Cremonini, L., Westerheijden, D., Benneworth, P., & **Dauncey, H.** (2014). In the shadow of celebrity? World-class university policies and public value in higher education. *Higher Education Policy, 27*, 341–361. DOI: *https://doi.org/10.1057/hep.2013.33*

Daraio, C., Bonaccorsi, A., & **Simar, L.** (2015). Rankings and university performance: A conditional multidimensional approach. *European Journal of Operational Research, 244*(3), 918–930. DOI: *https://doi.org/10.1016/j.ejor.2015.02.005*

Dichev, I. (2001). News or Noise? *Research in Higher Education, 42*, 237–266. DOI: *https://doi.org/10.1023/A:1018810005576*

Docampo, D., Egret, D., & **Cram, L.** (2015). The effect of university mergers on the Shanghai rankings. *Scientometrics, 104*, 175–191. DOI: *https://doi.org/10.1007/s11192-015-1587-5*

Florian, R. (2007). Irreproducibility of the results of the Shanghai academic ranking of world universities. *Scientometrics, 72*(1), 25–32. DOI: *https://doi.org/10.1007/s11192-007-1712-1*

Gadd, E. (2020). University rankings need a rethink. *Nature, 587*(7835), 523. DOI: *https://doi.org/10.1038/d41586-020-03312-2*

Gadd, E., & **Holmes, R.** (2020). Rethinking the rankings. *ARMA News*. Retrieved 11 January 2021. *https://arma.ac.uk/rethinking-the-rankings/*

Gingras, Y. (2014). Bibliometrics and research evaluation: uses and abuses. Cambridge, Mass.: MIT Press.

Hazelkorn, E. (2008). Learning to live with league tables and ranking: the experience of institutional leaders. *Higher Education Policy, 21*(2), 193–215. DOI: *https://doi.org/10.1057/hep.2008.1*

Hazelkorn, E. (2011). Rankings and the reshaping of higher education: The battle for world-class excellence. London: Palgrave Macmillan. DOI: *https://doi.org/10.1057/9780230306394*

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & **Rafols, I.** (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature, 520*, 429–431. DOI: *https://doi.org/10.1038/520429a*

Himanen, L., & **Gadd, E.** (2019). Introducing SCOPE – a process for evaluating responsibly. *The Bibliomagician Blog*. 11 December 2019. URL: *https://thebibliomagician.wordpress.com/2019/12/11/introducing-scope-aprocess-for-evaluating-responsibly/* Retrieved 14 January 2021.

Holmes, R. (2006). The THES university rankings: Are they really world class? *Asian Journal of University Education, 2*(1), 1–15. 07 January 2021 retrieved from *ir.uitm.edu.my/id/eprint/296/*

Holmes, R. (2010). The THES-QS World University Rankings, 2004–2009. *Asian Journal of University Education, 6*(1), 91–113. 07 January 2021 retrieved from *https://core.ac.uk/download/pdf/322375308.pdf*

Holmes, R. (2015). Searching for the Gold Standard: The Times Higher Education World University Rankings, 2010–2014. *Asian Journal of University Education, 11*(2), 1–30. 07 January 2021 retrieved from *https://education.uitm.edu.my/ajue/wp-content/uploads/2017/02/Asian-Journal-Of-University-Education-AJUE-Vol.-11-No.2-December-2015.pdf*

Holmes, R. (2021). University Rankings Watch Blog. URL: *http://rankingwatch.blogspot.com/*

IHEP. (2006). Berlin principles of ranking of higher education institutions. Washington, DC: IHEP.

INORMS Research Evaluation Working Group. (2020). Ranker Ratings Data. URL: *https://osf.io/mykud/* (Retrieved 14 January 2021).

INORMS Research Evaluation Working Group. (2021). URL: *https://inorms.net/activities/research-evaluation-working-group/*

Ioannidis, J. P. A., Patsopoulos, N. A., Kavvourai, F. K., Tatsioni, A., Evangelou, E., Kouri, I., Contopoulos-Ioannidis, D., & **Liberopoulos, G.** (2007). International ranking systems for universities and institutions: A critical appraisal. *BMC Medicine, 5*(30), 1–9. DOI: *https://doi.org/10.1186/1741-7015-5-30*

IREG. (2021). Retrieved 03 January 2021. *IREG Inventory on International Rankings*. *https://ireg-observatory.org/en/initiatives/ireg-inventory-of-international-rankings/*

Lee, J. J., Vance, H., Stensaker, B., & **Ghosh, S.** (2020). Global rankings at a local cost? The strategic pursuit of status and the third mission. *Comparative education, 56*(2), 236–25. DOI: *https://doi.org/10.1080/03050068.2020.1741195*

Lee, Z. S., & **Ong, K. M.** (2017). An unhealthy obsession with global university rankings. Kuala Lumpur: Penang Institute. Retrieved 09 January 2021 *https://penanginstitute.org/resources/research-papers/*

Marginson, S., & **van der Wende, M.** (2007). To rank or to be ranked. The Global rankings impact of global rankings in higher education. *Journal of Studies in Higher Education, 11*(3–4), 306–329. DOI: *https://doi.org/10.1177/1028315307303544*

Meredith, M. (2004). Why do universities compete in the ratings game? An empirical analysis of the effects of the US news and World Report college rankings. *Research in Higher Education, 45*(5), 443–461. DOI: *https://doi.org/10.1023/B:RIHE.0000032324.46716.f4*

Moed, H. (2017). 'A Critical Comparative Analysis of Five World University Rankings'. *Scientometrics, 110*(2), 967–90. DOI: *https://doi.org/10.1007/s11192-016-2212-y*

Munch, R. (2014). Academic capitalism: Universities in the global struggle for excellence. New York: Routledge. DOI: *https://doi.org/10.4324/9780203768761*

Musselin, C. (2018). New forms of competition in higher education. *Socio-Economic Review, 16*(3), 57–683. DOI: *https://doi.org/10.1093/ser/mwy033*

Pagell, R. (2014). Ruth's rankings 2: A Brief History of Rankings and Higher Education Policy. *librarylearningspace* 08 August 2014. Retrieved 09 January 2021 *https://librarylearningspace.com/ruths-rankings-2-brief-history-rankings-higher-education-policy/*

Pagell, R. (2019). Ruth's rankings 40: Deconstructing QS subjects and surveys. *librarylearningspace* 12 April 2019. Retrieved 09 January 2021 *https://librarylearningspace.com/ruths-rankings-40-deconstructing-qs-subjects-surveys/*

Piro, F. N., & **Sivertsen, G.** (2016). How can differences in international university rankings be explained? *Scientometrics, 109*(3), 2263–2278. DOI: *https://doi.org/10.1007/s11192-016-2056-5*

Rodionov, D. G., Fersman, N. G., & **Kushneva, O. A.** (2016). Russian universities: Towards ambitious goals. *International Journal of Environmental and Science Education, 11*(8), 2207–2222. DOI: *https://doi.org/10.12973/ijese.2016.591a*

Rozman, I., & **Marhl, M.** (2008). Improving the quality of universities by world-university-ranking: A case study of the University of Maribor. *Higher Education in Europe, 33*(2–3), 317–329. DOI: *https://doi.org/10.1080/03797720802254221*

Safón, V. (2013). What do global university rankings really measure? The search for the X factor and the X entity. *Scientometrics, 97*(2), 223–244. DOI: *https://doi.org/10.1007/s11192-013-0986-8*

Safón, V., & **Docampo, D.** (2020). Analyzing the impact of reputational bias on global university rankings based on objective research performance data: the case of the Shanghai Ranking (ARWU). *Scientometrics, 125*(3), 2199–2227. DOI: *https://doi.org/10.1007/s11192-020-03722-z*

Saisana, M., D'Hombres, B., & **Saltelli, A.** (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research Policy, 40*, 165–177. DOI: *https://doi.org/10.1016/j.respol.2010.09.003*

Shreeve, R. L. (2020). Globalisation or westernisation? The influence of global universities in the context of the Republic of China (Taiwan). *Compare, 50*(6), 92–927. DOI: *https://doi.org/10.1080/03057925.2020.1736403*

Sowter, B. (2008). The Times Higher Education Supplement and Quacquarelli Symonds (THES - QS) World University Rankings: New Developments in Ranking Methodology. *Higher Education in Europe, 33*(2–3), 345–347. DOI: *https://doi.org/10.1080/03797720802254247*

Stack, M. (2016). Global university rankings and the mediatization of higher education. London: Palgrave Macmillan. DOI: *https://doi.org/10.1057/9781137475954*

Turner, D. (2005). Benchmarking in universities: League tables revisited. *Oxford Review of Education, 31*(3), 353–371. DOI: *https://doi.org/10.1080/03054980500221975*

Usher, A. (2014). When the Times Higher Education rankings fail the fall-down-laughing test. *Higher Education Strategy Associates* 06 February 2014. Retrieved 09 January 2021. *https://higheredstrategy.com/when-the-times-higher-education-rankings-fail-the-fall-down-laughing-test/*

Usher, A. (2016). Unreliable data, unreliable rankings. *Inside Higher Ed* 18 May 2016. Retrieved 09 January 2021 *www.insidehighered.com/blogs/world-view/data-and-rankings-healthy-debate*

Usher, A., & **Savino, M.** (2007). A global survey of university ranking and league tables. *Higher Education in Europe, 32*(1), 5–15. DOI: *https://doi.org/10.1080/03797720701618831*

Van Dyke, N. (2008). Self- and peer-assessment disparities in university ranking schemes. *Higher education in Europe, 33*(2–3), 285–293. DOI: *https://doi.org/10.1080/03797720802254114*

van Raan, A. (2007). Challenges in the ranking of universities. In J. Sadlak & N. C. Liu (Eds.), *The World-Class University and Ranking: Aiming Beyond Status* (pp. 87–122). Shanghai/Cluj: UNESCO-CEPES.

Van Vught, F. A., & **Ziegele, F.** (Eds.) (2012). Multidimensional ranking: The design and development of U-Multirank (Vol. 37). Dordrecht: Springer Science & Business Media. DOI: *https://doi.org/10.1007/978-94-007-3005-2*

Vernon, M. M., Andrew Balas, E., & **Momani, S.** (2018). *PLoS ONE, 13*(3), Article number e0193762. DOI: *https://doi.org/10.1371/journal.pone.0193762*

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J. & **Wouters, P.** (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419–2432. DOI: *https://doi.org/10.1002/asi.22708*

Waltman, L., Wouters, P., & **van Eck, N. J.** (2017). Ten rules for university rankings. *Research Europe*. URL: *https://www.researchresearch.com/news/article/?articleId=1368350*

Wildavsky, B. (2010). The Great brain race: How global universities are reshaping the world. Princeton and Oxford: Princeton University Press.

Wu, Y., & **Liu, N. C.** (2017). Academic ranking of world universities (ARWU): Methodologies and trends. In F. J. Cantú-Ortiz (Eds.), *Research analytics: Boosting University Productivity and Competitiveness through Scientometrics*. New York: Auerbach. DOI: *https://doi.org/10.1201/9781315155890-6*

TO CITE THIS ARTICLE:
Gadd, E., Holmes, R., & Shearer, J. (2021). Developing a Method for Evaluating Global University Rankings. *Scholarly Assessment Reports*, 3(1): 2, pp. 1–19. DOI: *https://doi.org/10.29024/sar.31*

Submitted: 26 January 2021
Accepted: 12 April 2021
Published: 28 April 2021

Scholarly Assessment Reports is a peer-reviewed open access journal published by Levy Library Press.