Meaningful, responsible & effective research evaluation

Dr Elizabeth Gadd
Research Policy Manager
Loughborough University

Dr Baron Wolf
Asst. Vice President for Research Strategy & Data Analytics
University of Kentucky
Overview

- Who we are
- What we want to do
- Why we want to do it
- How we’re doing it
- Work-packages
- What next?
Who we are

Baldvin Zarioh
Hirofumi Seike
Dr Baron Wolf
Anne Albinak

Laura Beaupre
Dr Patricia Smit
Tanja Strøm

Justin Shearer
Laura Himanen

Aline Pacifico Rodrigues
Dr Elizabeth Gadd

Dipti Pandya

inorms
Research Evaluation Working Group
What we want to do

- Consider how best to ensure that research evaluation is meaningful, responsible and effective.
Why we want to do it

- Research is a global endeavour
- Research evaluation is also a global activity
- Many poor practices around the world (university rankings, inappropriate use of journal metrics, skewed individual researcher evaluation) are global problems and require global solutions
- Individual countries simply cannot address them alone
- The INORMS REWG is an attempt to provide a united global voice
How we’re doing it

- Ten colleagues allocated to 2 sub-groups
  - Rankings - led by Justin Shearer (ARMS)
  - Senior Manager Briefings - led by Laura Himanen (Fin-ARMA)
- Wider INORMS-RES-EVAL@jiscmail.ac.uk discussion list
Work package 1: Rating the rankers

- University rankings have global power
- Self-appointed
- Answer to no-one
- And yet:
  - Funders use them to decide who to fund
  - Credit-rating bodies use them to sway investments
  - Students use them to decide where to study
  - Academics use them to plan their careers
- Does this lead to meaningful, responsible & effective research evaluation?
Tackling the problem: Criteria for rating the rankers

- Available at: https://inorms.net/activities/research-evaluation-working-group/
- Five key themes:
  - General approach
  - Governance
  - Methodologies
  - Indicators
  - Usability
- Consultation open 22 May - 10 June 2019
- Outcomes will inform methodology for rating the rankers
Work package 2: Briefing Senior Managers

- Responsible research evaluation discussions are often held by middle-managers, practitioners or academics.
- Those with the most influence over responsible research evaluation practice are very senior leaders.
- Hard to access senior leaders; difficult to ‘train’.
- No readily available materials by which leaders can be briefed.
Briefing materials for senior managers

- Set of PowerPoint slides with notes
  - Short version (20 minutes)
  - Longer version (60 minutes)
- Can be adapted to different settings
- Translated into as many languages as we can
- May also be useful as basis for training others - e.g., research managers
- Better decision making through responsible research evaluation (VCOS model)
  - What do you Value?
  - Context-based evaluation
  - Options for evaluating (quantitative & qualitative)
  - Sense-check
  - Review
What next?

- Working Group set to run until June 2020
- Lots more work to do!
  - International Group on Journal Indicators - expressed interest
  - Plan S - Responsible Metrics element requires Research Manager input
  - Science of Science work - James Wilsdon recognises role for co-production with research managers and information specialists
- Possibility of extending Working Group for 1-2 years.
- Considering an INORMS International symposium on university rankings
Input welcome from SRAI Members

- Feedback on deliverables invited
  - Rating the rankers criteria
  - Content of Senior Manager briefings
- Guidance as to whether the Symposium idea is a good one and if so, how to make this happen?
- How to change perceptions of other rankings ‘users’?
  - Students
  - Funders
- How to get access to Senior Managers to provide the briefing?
Thanks for listening

Dr Elizabeth Gadd
Chair, INORMS Research Evaluation Working Group
Research Policy Manager, Loughborough University, UK
E.a.gadd@lboro.ac.uk
@lizziegadd
https://inorms.net/activities/research-evaluation-working-group/

Dr Baron Wolf
Member, INORMS Research Evaluation Working Group
Asst. Vice President for Research Strategy & Data Analytics, University of Kentucky, USA
baron.wolf@uky.edu
Meaningful, responsible & effective research evaluation

Dr Elizabeth Gadd
@lizziegadd
Overview

- World University Rankings
  - What’s the problem?
  - What are we doing about it?
- Senior University Leaders
  - What’s the problem?
  - What are we doing about it?
World University Rankings - what’s the problem?

- They seek to identify the ‘best’ without defining what that looks like
- They don’t measure what matters
- The indicators aren’t always valid
- The weightings are arbitrary
- The methods and data are not transparent
- They don’t compare like with like
- They use discriminatory data sources
- They embed the status quo
- Shall I go on?
Shanghai Ranking’s Academic Ranking of World Universities 2018 Press Release

Wednesday, August 15, 2018
Shanghai, People’s Republic of China

The 2018 Academic Ranking of World Universities (ARWU) is released today by Shanghai Ranking Consultancy. Since 2003, ARWU has been presenting the world top 500 universities annually based on transparent methodology.
A critical comparative analysis of five world university rankings

Henk F. Moed

Sapienza University of Rome, Italy

Version 15 Nov 2016, accepted for publication in Scientometrics

Table 2. Key results overlap analysis of top 100 lists in all 5 ranking systems

<table>
<thead>
<tr>
<th>Indicator</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of different institutions</td>
<td>194</td>
</tr>
<tr>
<td>Number of institutions appearing in the top 100 lists of all 5 systems</td>
<td>35</td>
</tr>
</tbody>
</table>
Rank these:
### Indicators and Weights for ARWU

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Indicator</th>
<th>Code</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of Education</td>
<td>Alumni of an institution winning Nobel Prizes and Fields Medals</td>
<td>Alumni</td>
<td>10%</td>
</tr>
<tr>
<td>Quality of Faculty</td>
<td>Staff of an Institution winning Nobel Prizes and Fields Medals</td>
<td>Award</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>Highly cited researchers in 21 broad subject categories</td>
<td>HiCi</td>
<td>20%</td>
</tr>
<tr>
<td>Research Output</td>
<td>Papers published in Nature and Science*</td>
<td>N&amp;S</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>Papers indexed in Science Citation Index-expanded and Social Science Citation Index</td>
<td>PUB</td>
<td>20%</td>
</tr>
<tr>
<td>Per Capita Performance</td>
<td>Per capita academic performance of an institution</td>
<td>PCP</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

* For institutions specialized in humanities and social sciences such as London School of Economics, N&S is not considered, their indicators.
“Never mind the quality, feel the width”

Faculty/Student Ratio (20%)

Teaching quality is typically cited by students as the metric of highest importance to them when comparing institutions using a ranking. It is notoriously difficult to measure, but we have determined that measuring teacher/student ratios is the most effective proxy metric for teaching quality. It assesses the extent to which institutions are able to provide students with meaningful access to lecturers and tutors, and recognizes that a high number of faculty members per student will reduce the teaching burden on each individual academic.

Faculty/student Ratio constitutes 20 percent of an institution’s final score.
World University Rankings - what’s the problem?

- They seek to identify the ‘best’ without defining what that looks like
- They don’t measure what matters
- The indicators aren’t always valid
- The weightings are arbitrary
- The methods and data are not transparent
- They don’t compare like with like
- They use discriminatory data sources
- They embed the status quo
- Shall I go on?
Do university rankings measure anything at all?

by the quality of research outputs. Indeed the extraordinarily high correlation between teaching and research reputation within the THES input data suggests that it is measuring neither of these things but some form of brand awareness, a measure that less privileged institutions will necessarily struggle with. There is also a massive difference for Chinese universities between the
How world university rankings are used

- By students - Where to study
- By employers - Who to employ
- By academics - Where to work
- By research funders - Who to fund
- By Universities - Who to recruit
- By Governments - Allocating resource
- By Credit rating bodies - Who to invest in
“research indicates that applicant behaviour is very much conditioned by rankings; slight changes can cause perceptible ebbs and flows in the number and quality of applicants, especially international students.”

“Of our graduate recruiter informants, 25% used league tables to shape their marketing strategies.”

University of Sussex. 2006. “Needs of Employers and Related Organisations for Information about Quality and Standards of Higher Education: Report to HEFCE by the University of Sussex School of Education.”
By universities

FX Coudert
@fxcoudert

Instead of judging each postdoc by evaluating its CV & quality of science, @univ_lille uses Shanghai ranking of their alma mater. 😞 Only candidates from 200 top univ are even allowed to apply.

David Monniaux @MonniauxD

L’université de Lille propose un appel à financements post-doctoraux mais uniquement pour des jeunes chercheurs issus des 200 premières universités du classement de Shangaiï.
“High-quality future faculty are also more attracted to highly ranked institutions, making ‘recruitment easier because of their good reputation’ because ‘success breeds success’. [Also] good staff leave because they are disappointed by the rankings.”

By governments

We want the best universities in the world…. How many universities do we have? 83? We’re not going to divide the money by 83 (Nicolas Sarkozy, President, France, 2009).

Work [is underway] on establishing the country’s first ‘research-intensive’ university… universities which earned a place in the top 500 rankings… were entitled to financial support (Jurin Laksanavisit, Education Minister, Thailand, 2009).

By credit rating bodies

S&P Global Ratings

RatingsDirect

Research Update:
University of Nottingham Rat Affirmed At 'A+' On Solid Performance; Outlook Stable

Rating Action

Rationale
The rating on Nottingham University reflects its stand-alone credit profile (SACP), which we assess at 'a+'. In our view, Nottingham University operates in a low risk industry. In addition, a strengthening of its academic standing and reputation in the U.K., as well as in China and Malaysia, where it holds an interest in two campuses, supports a high level of student demand.

As a Russell Group member, Nottingham University is recognized for its commitment to maintaining a high level of research, teaching, and links with business and the public sector. This is also reflected in Nottingham University's rankings both in the U.K. and globally; the university climbed three places to No. 18 in the Complete University Guide 2018, and remains solidly in the top 100 of the QS World University Rankings 2018, at No. 84 (although falling from a joint No. 75). Nottingham University also has a good track record with regard to attracting students, with nearly seven applications for each of the 11,500 students enrolled in the year to July 2016. Entry standards remain comfortably in the top quartile, as does Nottingham University's reputation for research.
The research evaluation food chain

- Researchers
- Universities
- Funders
- Governments
- World rankings

Data vendors
What’s the solution?
Rating the rankers

- What if Rankers are no longer at the top of the food chain?
- Rate the rankers criteria: https://inorms.net/activities/research-evaluation-working-group/
- Key themes:
  - Responsibility
  - Transparency
  - Measuring what matters
  - Rigour
Responsibility

- Engage with the ranked.
- Declare any conflict of interests.
- Respect the ownership of underlying data
- Open to correction.
- Deal with gaming.
Transparency

- Transparent aims.
- Transparent methods.
- Transparent data availability.
- Open data.
- Financially transparent.
Measure what matters

- Drive good behaviour.
- Measure against mission.
- One thing at a time
- No monotonic indicators
- Tailored to different audiences.
- No unfair advantage.
Rigour

- Rigorous methods.
- No sloppy surveys.
- Defines “University”.
- Validity.
- Sensitivity.
- Honest about uncertainty.
Next steps

- Finalise criteria for rating
- Pilot criteria on one Ranking
- Qualitative and quantitative approach
- Revise criteria
- Invite five ranking bodies to self-assess alongside expert assessment
- Release scores and data in Summer 2020
Senior Managers - what’s the problem?

- Responsible research evaluation discussions are often held by middle-managers, practitioners or academics
- Those with the most influence over responsible research evaluation practice are very senior leaders
- Hard to access senior leaders; difficult to ‘train’
- No readily available materials by which leaders can be briefed
According to the tender documents in the link provided, candidates for the P.I. position have to pass an "Analysis of Absolute Worth" based on the criteria below. No comments, especially regarding #8.

César Pallares
@copdeb
Replying to @LizzieGadd @Protohedgehog and 2 ...

Well. Maybe a solution would be to offer a free training to university and research heads about responsible metrics. Also to those administrative staff in charge of research. They're the key to change uni policies.
"The ongoing transformation is not only about playing the competitive game, but also about shifting the role of university presidents into becoming competitors.... As they become competitors, they develop a new understanding of their role and become more interventionist. But members of their institution often contest this new role. In order to legitimize their actions, university leaders often point to the outcomes of competition to justify their decisions. The different competitive schemes in which their institutions partake become steering instruments. Competition leads them to be more strategic, but by the same token the results of the competition help them implement and legitimize their strategy."

Introducing the INORMS SCOPE model

1. Start with your values
2. Context
3. Options
4. Probe
5. Evaluate
START with what you value
START with what you value

- Not with the data you have available
  - The Streetlight Effect
- Not what others value
- University autonomy: use it or lose it

“If my h-index is the answer, what is the question?”
The streetlight effect

- The streetlight effect is a phenomenon where people lose their wallets or other belongings in dimly lit areas and then find them after they are lit up by streetlights.

- The cartoon illustrates the concept with a police officer saying, "This is where you lost your wallet?" and the individual responding, "No, I lost it in the park, but this is where the light is."
START with what you value

- Not with the data you have available
  - The Streetlight Effect
- Not what others value
- University autonomy: use it or lose it

“If my h-index is the answer, what is the question?”
CONTEXT
Understand who & why you’re evaluating

Figure 1. Risks associated with metric use in various settings

- **Low risk**
- **Medium risk**
- **High risk**
Use of FWCI in measuring to understand International Comparative Performance of UK Research Base – 2016 report on 2011-2014 data

Use of FWCI to identify staff for redundancy...

So, these are the proposed criteria I (a historian, remember) would have to meet to avoid being among the ~140 of whom ~65 will lose our jobs.

The University will consult with the Trade Unions on a set of criteria to reduce the group of around 627 academic posts 'in scope' to a group of around 140 posts that will subsequently be 'at risk' of redundancy. The loss of 65 posts will come from this 'at risk' pool. The criteria that we are proposing to apply to identify the 'at risk' pool are defined below. However, please note that these criteria are subject to consultation with the Trade Unions and therefore may change.

If staff meet or exceed one or more of the proposed criteria below, they will not be at risk.

- Research and other income in the four-year period from 1 August 2012 to 31 August 2016 of £400k, £300k, £200k or more respectively for staff in Grades 9, 8, 7f, or
- Research awards from 1 August 2015 to 31 March 2017 of £225k, £150k, £75k or more respectively for staff in Grades 9, 8, 7f, or
- A sum of Field-Weighted Citation Impact greater than 1.5
- Staff on a core, permanent teaching only, teaching focused or teaching scholarship contract

Where staff have had a significant period of absence from work (three months or more) due to maternity leave or sickness absence for example, we will consider the data in these cases and seek to mitigate any adverse impact attributable to the period of absence, e.g. by using a time period that is more relevant to the individual circumstances. Equally we will give consideration as to how the criteria may need to be adjusted in relation to staff who have a disability.
Do we need to evaluate at all?

- Huge growth in incentivising behaviour through measurement
- Campbell’s Law: “The way you measure me is the way I’ll behave”
- Measuring is not always the best way to incentivise behaviour
Open Science and its role in universities:

research. In order for these goals to be achieved, universities should align their assessment, reward and evaluation systems with Open Science developments.⁹
The Hong Kong Manifesto for Assessing Researchers: Fostering Research Integrity

David Moher¹, Lex Bouter², Sabine Kleinert³, Paul Glasziou⁴, Mai Har Sham⁵

¹Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute; School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada; ²Department of Epidemiology and Biostatistics, Amsterdam University Medical Centers, location VUmc, and Department of Philosophy, Faculty of Humanities, Vrije Universiteit, Amsterdam, The Netherlands; ³The Lancet, London Wall Office, London, UK; ⁴Centre for Research in Evidence-Based Practice, Bond University, Gold Coast, Qld, Australia; and ⁵School of Biomedical Sciences, LKS Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
OPTIONS
Options

- Is your measure a suitable proxy for what you’re measuring?
- Quantitative measures are for quantifiable things...
  - Citations, publications, money, students
- Qualitative measures for qualifiable things...
  - Quality, diversity, excellence, value
- Beware using quantitative indicators as a proxy for qualitative things
  - Citations ≠ quality
  - Ranking position ≠ excellence
When is Peer Review the Gold Standard, and When is it Only Tin?

Published on October 22, 2016 by Tony Waters

A new gold standard of peer review is needed

September 29, 2017 * Author: Tony Waters
metrics are rubbish
but ...(far)
people are worse

Alan Dix
University of Birmingham and Talis
http://alandix.com/ref2014/
PROBE
Probe for potential negative impacts

1. Who does this discriminate against?
2. How could this be gamed?
3. What might the perverse incentives and consequences be?
4. Do the benefits of measuring outweigh the cost of measuring?
5. Is evaluating research actually going to make it any better?
REF 2014 cost almost £250 million

Accountability review finds cost of assessment equates to 2.4 per cent of funding bodies' expected spend over next six years

July 13, 2015

By Holly Else
Twitter: @HollyElse
Does the cost outweigh the benefit?

1 citation tool + 1 research metrics post = £90,000 p.a.

1 “FAR” tool + 1 Research Impact Librarian = $175,000 pa
Probe for potential negative impacts

1. Who does this discriminate against?
2. How could this be gamed?
3. What might the perverse incentives and consequences be?
4. Do the benefits of measuring outweigh the cost of measuring?
5. Is evaluating research actually going to make it any better?
You don’t fatten a pig by weighing it
EVALUATE your evaluation...
Next steps

- Framing exercise with Senior Leaders of UK and Finnish universities
- Webinar to pilot the SCOPE methodology with Research Managers
- Ultimately, outputs available to all - CC-BY
- Set of powerpoint slides with notes
  - Short version (20 minutes)
  - Longer version (60 minutes)
- Can be adapted to different settings
- Translated into as many languages as we can
Better to light a candle than curse the darkness.
Get involved!

- “Light a candle” in your own setting
- Join the conversation:
  - INORMS-RES-EVAL@jiscmail.ac.uk discussion list
  - Lis-Bibliometrics@jiscmail.ac.uk discussion list
Thanks for listening

- Dr Elizabeth Gadd
- Chair, INORMS Research Evaluation Working Group
- Research Policy Manager, Loughborough University, UK
- E.a.gadd@lboro.ac.uk
- @lizziegadd
- https://inorms.net/activities/research-evaluation-working-group/
Ranking media hype

Oxford Tops Times Higher Education World University Rankings For Third Straight Year

02:00PM Sep 26, 2018
Let me fix that for you

Runner who starts race ahead of others finishes first
Responsibility

- **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.

- **Self-improving.** Regularly applies measures of quality assurance to their ranking processes.

- **Declare any conflict of interests.** Provides a declaration of potential conflicts of interest as well as how they actively manage those conflicts.

- **Respect the ownership of underlying data.** Those submitting data for ranking purposes do not lose ownership of that data and are given options as to what secondary uses might be made of it.

- **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.

- **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.
Transparency

- **Transparent aims.** States clearly the purpose of the ranking, what it seeks to measure, and their target groups. (BP)

- **Transparent methods.** Publishes full details of their ranking methodology, so that given the data a third party could replicate the results. (CWTS, DORA, BP)

- **Transparent data availability.** Provides detailed descriptions of the data sources being used, inclusion and exclusion parameters, date data snapshots were taken, and so on.

- **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. (LM, DORA, BP)

- **Financially transparent.** Publishes details of all sources of income from training, events, advertising, and so on.
Measure what matters

- **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.

- **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. (LM, BP, Blank, Shen)

- **One thing at a time.** Does not combine indicators to create a composite metric thus masking what is actually being measured.

- **No monotonic indicators** for which a good value will depend on the mission of a university.

- **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.

- **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. For bibliographic databases that do not have global representation.
Rigour

- **Rigorous methods.** Data collection and analysis methods should pass tests of scientific rigour, including sample size, representation, normalisation, handling of outliers, etc.

- **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.

- **Defines “University”.** When using multiple data sources to take measurements, uses a consistent definition of university across the different data sources. (E.g., universities with multiple campuses - including off-shore, or those with teaching hospitals)

- **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.

- **Sensitivity.** Indicators are sensitive to the nature of the characteristic they claim to measure.

- **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 intervals or other techniques, without giving a false sense of precision.