What about next-generation assessments?
The future of assessment: Bottom-up

Michael Ochsner, ETH Zurich & FORS, Lausanne

Conference: Beyond impact factor, h-index and rankings: Evaluate science in more meaningful ways, Bern, 21st November 2018

Presented work in collaboration with Sven E. Hug, Hans-Dieter Daniel, Mišo Dokmanović, Aldis Gedutis and Emanuel Kulczycki
Outline

- Validity or measuring implies Concepts
- Projects on quality, performance and potential in SSH
- 42 or Responsible Metrics?
- The Future: Bottom-Up Evaluation Procedures
42: Metrics and Concepts

- Novel/Radio play by Douglas Adams
  *Hitchhiker’s Guide to the Galaxy* (1979)
- the ultimate question of life, the universe and everything
  - 7.5 million years to compute and check
  - The answer was.... 42
- Deep Thought:
  answer is meaningless – because the question was stupid:
  - did not specify the form of answer
  - nor did they really know what they asked for
Validity

- The extent to which a measure (i.e., an indicator) actually measures what it purports to measure (i.e., a concept) (Borsboom et al., 2004, p. 1061)

- Data-driven: „measuring what can be measured“ endangers validity, mostly reducing it to correlation.

- Thunder correlates highly with lightning (and there is even a causal relationship). However, lightning cannot measure thunder.
Measurement in Scientometrics

- Bibliometrics and Scientometrics:
  - „[the assessments] often still make only a weak connection between theoretical definitions of quality and its measures“ (Brooks, 2005, p. 1-2)
  - „these metrics do not actually measure research quality. For example, research income is an input, rather than an output measure“ (Donovan, 2007, p. 586)
  - „It is [...] extremely difficult if not impossible to express what citations measure in one single theoretical concept [...] citations measure many aspects of scholarly activity at the same time.“ (Moed, 2005, p. 221)
  - Tahamtan & Bornmann (2017) review literature why authors cite. Quality of research is among a plethora of other reasons
Measurement Approach

Universe of quality criteria

Concepts

Indicators

What?

How?

Universe of indicators

Quality criterion A

Quality criterion i

Aspect A_1

Aspect A_2

Aspect i_1

Aspect i_j

1 2 3 4

y z

Projects on Quality, Performance and Career/Research Potential in the SSH
Time

Negatively connotated
‘traditional’ research

Positively connotated
‘modern’ research

Quality

Negatively connotated
‘traditional’ research

Positively connotated
‘modern’ research

autonomy

disciplinary

individual effort

Ground-breaking
innovation

interdisciplinary

international

public orientied

‘Small-step’
innovation

career oriented

economistic

internationalist

Simplifying

One sided, repetitive

Self-focused

Isolated

‘Ground-breaking’
innovation

autonomy

disciplinary

individual effort
Time

Traditional ‘traditional’ research

Negatively connoted

Autonomy

Disciplinary

Individual effort

'Ground-breaking' innovation

Economistic

Internationalist

Determined by others, predictable

Self-focused

Isolated

One sided, repetitive

Simplifying

Quality

Positively connoted

Public oriented

‘Small-step’ innovation

Career oriented

International

Uncritical, saving

We study nature... you go save it!
Criteria for Research Quality/Performance

- English Literature, German Literature and Art History
- Consensual Indicators (orange: all three; blue: in two disciplines)

1. Scholarly exchange
2. Innovation, originality
3. Productivity
4. Rigour
5. Fostering cultural memory
6. Recognition
7. Reflection, criticism
8. Continuity, continuation
9. Impact on research community
10. Relation to and impact on society
11. Variety of research
12. Connection to other research
13. Openness ideas and persons
14. Self-management, independence
15. Scholarship, erudition
16. Passion, enthusiasm
17. Vision of future research
18. Connection between research and teaching, scholarship of teaching
19. Relevance
Measuring Research Quality/Performance

- Which aspects can be measured by indicators?
  - 50% of the relevant aspects cannot be measured by indicators
- What do Indicators that are commonly used measure?
  - Are these the relevant criteria?

Table 1: Frequently used indicators and criteria they can potentially measure

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations</td>
<td>Recognition; impact on research community; relevance</td>
</tr>
<tr>
<td>Prizes</td>
<td>Recognition; impact on research community; relevance</td>
</tr>
<tr>
<td>Third party funding</td>
<td>Recognition; impact on research community; relevance; relation to and impact on society</td>
</tr>
<tr>
<td>Collaborations</td>
<td>Scholarly exchange; recognition</td>
</tr>
<tr>
<td>Transfers to society and economy</td>
<td>Relation to and impact on society</td>
</tr>
<tr>
<td>Publications</td>
<td>Scholarly exchange; productivity</td>
</tr>
<tr>
<td>Board memberships</td>
<td>Scholarly exchange; recognition; impact on research community</td>
</tr>
<tr>
<td>Recruitment</td>
<td>Continuity, continuation</td>
</tr>
</tbody>
</table>
Criteria for Research Quality/Performance

- Measured by commonly used indicators (**bold and italic**)

1. **Scholarly exchange**
2. Innovation, originality
3. **Productivity**
4. Rigour
5. Fostering cultural memory
6. **Recognition**
7. Reflection, criticism
8. **Continuity, continuation**
9. **Impact on research community**
10. **Relation to and impact on society**
11. Variety of research
12. Connection to other research
13. Openness ideas and persons
14. Self-management, independence
15. Scholarship, erudition
16. Passion, enthusiasm
17. Vision of future research
18. Connection between research and teaching, scholarship of teaching
19. **Relevance**
Validity Check for Commonly used Metrics

- Valid measures for research quality?
  orange: three disc.; blue: two disc.; bold and italic: commonly used

1. Scholarly exchange
2. Innovation, originality
3. Productivity
4. Rigour
5. Fostering cultural memory
6. Recognition
7. Reflection, criticism
8. Continuity, continuation
9. Impact on research community
10. Relation to and impact on society
11. Variety of research
12. Connection to other research
13. Openness ideas and persons
14. Self-management, independence
15. Scholarship, erudition
16. Passion, enthusiasm
17. Vision of future research
18. Connection between research and teaching, scholarship of teaching
19. Relevance
Other Applications

- Questions
  - How can the criteria be used in specific evaluation situations?
  - How do the criteria travel across more disciplines?

- Applications
  - Criteria for Grant Evaluation for Early Career Investigators
  - Social Sciences (Ochsner & Dokmanović, 2017)
  - Law (Lienhard et al., 2016), Theology (Mertens et al.)
European Network for Research Evaluation in Social Sciences and Humanities

- COST-Action ENRESSH
- Set out to find responsible research evaluation procedures
- Including a metrics part but not limited to it
- www.enressh.eu
National Evaluation Procedure
National Career Promotion
Metric, push for English

No db, not SSH- or discipline specific

Pb-funding, metric

Pb-funding, non-metric

Non-metric, SSH- or discipline-specific
What kind of Next-Generation Assessment Procedures?
Summary

- Most criteria travel well across disciplines and situations
  - SS & H have been investigated
  - STEM disciplines: most criteria seem to be valid as well
- Disciplinary differences exist
  - Most often only in the weighting (importance) of the criterion
- Societal Impact is not equal to Quality
- Science Europe: Best way to provide value to society is fostering **quality of research** (Science Europe, 2017)
- Metrics are not measuring quality comprehensively (<50%)
Conclusion: We know the answer anyway – 42

- Deep Thought created a new solution including beings that will resolve the question: Planet Earth, directed by white lab mice
  - Calculating time: 10 million years.
  - Earth destroyed before the result was ready by Psychiatrists who feared loss of their careers
Next-Generation Evaluation Procedures

- Importance of Criteria
  - Criteria lead to more robust and fair assessments (Thorngate et al., 2009)
  - Bottom-Up procedure leads to legitimacy (Hug, Ochsner & Daniel, 2014; Derrick & Samuel, 2017)
  - Metrics are never responsible, the users should be responsible

- Specificities of Bottom-Up Procedures
  - Respect disciplinary differences and levels of assessment
  - Based on research practices and knowledge production
  - Assure diversity and think of incentives metrics/criteria introduce
  - Transparency with decisions taken
Open Access Edited Volume on Bottom-Up Assessment Procedures in the Humanities


- With contributions by i.a.: Wiljan van den Akker, Alfred Hornung, Wilhelm Krull, Michèle Lamont, Gerhard Lauer, Christian Mair, Ingo Plag, Björn Hammarfelt, Ingrid Gogolin, Gunnar Sivertsen, Elea Giménez-Toledo, Thomas König, Remigius Bunia
Literature on the projects (logical order)


References

Contact

- Michael Ochsner, ETH Zurich & FORS, Lausanne, Switzerland
  ochsner@gess.ethz.ch
  http://forscenter.ch/de/about-us-3/staff/michael-ochsner/

- COST- Action CA-15137:
  European Network for Research Evaluation in the Social Sciences
  and Humanities ENRESSH
  www.enressh.eu

- EvalHum:
  European Association for Research Evaluation in the Social Sciences
  and Humanities
  www.evalhum.eu