WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Excellence, Relevance, Peer Review and Indicators – Scientific insights on how to evaluate science

June 27, 2018, Kyiv

Dr. Dagmar Simon

Agenda

- Trends in the evaluation landscape
- Assessement and/or learning organisations?
- Peer review and metrics: two sides of one coin?
- Unintended consequences of indicators icampact: the case of metrics
- Impact: the case of metrics
- Some conclusions

The Transition from Quality Control to Quality Monitoring in Science

(nach Hemlin, Rasmussen 2006)

Dimension	Quality Control (Product Orientation)	Quality Monitoring (Process Orientation)
Criteria	Scientific	Scientific and societal
Focus	Individual researchers	Organizations, networks
Goal	Valid, reliable knowledge	Socially robust knowledge, learning
Evaluator	Traditional peers	New peers, users. consultants, lay persons
Evaluation time	After production	Continuously
Science study perspective	First order: philosophy and sociology of knowledge	Second order: knowledge management, organizational learning


The Metric Tide
Report of the Independent Review of the Role of Metrics in Research Assessment and Management
July 2015
19
18
17
16
15
14
13
12
11
10
9
8
7

WEAKNESSES OF PEER REVIEW

- It is slow, inefficient and expensive, although most costs are hidden;
- Human judgment is subjective – which may however also be seen as a strength;
- It is almost by definition not transparent
- It is inconsistent, sometimes characterised as a lack of inter-rater reliability;
- It is a biased process (e.g. gender bias regarding career decisions, bias against negative studies in publication decisions, bias in favour of prestigious institutes, bias in favour of dominant paradigms);
- Its bias is strengthened by the Matthew effect
- The process can be abused (e.g. to block competitors, to plagiarise);
- It is not very good at identifying errors in data or even in detecting fraudulent research;
- It cannot process the complete research output of a nation and will therefore result in distorted rankings (since rankings are sensitive to the selection of submissions to the assessments);
- It cannot provide information about the productivity and efficiency of the research system;
- The selection of peer reviewers may create problems because of a variety of reasons (bias, lack of experts in emerging and interdisciplinary areas, lack of experts due to the speed of research areas, etc).

STRENGTHS OF PEER REVIEW

- Its foundation in specialised knowledge of the subject, methodology and literature relevant for specific decisions;
- Its social nature;
- The subjectivity of this approach could be seen as a strength (as well as a weakness);
- It can help assess elements of research which are challenging to quantify e.g. novelty;
- It can deliver more nuanced and detailed understandings of research in the context of research production.

Source: Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. HEFCE

Measuring scientific performance

Science performance cannot be measured. Indicators are able to display the level of performances, their quality, only approximately.

For a long time, the philosophy of science has given up on establishing criteria by which the quality of science performance can be measured. Instead they agreed upon a method: ‘’Reasonable betting among peers’’ (Toulmin).

Performance Indicators in Science

- Quantitative indicators reduce complexity
- Indicators depict science performance without explanation
- Differentiation of indicators ...
- …however (implicit) hierarchies: publications and third-party funding
- Establishing comparability of the incommensurable (Heintz 2010, Godin 2006)
- Indicators and numbers cannot display every potential science performance
- Anything that is not measured doesn't count (anymore)

Measurement of Inputs and Outputs

Impact of public research results on economic activities (Salter and Martin, 2001)

Open the „black box“ of the production of scientific knowledge (Carayol et al., 2006)

Using the standard human capital theory: accumulated stock of knowledge in human capital a priori being a critical production factor of further knowledge production

Measurement of Inputs and Outputs

„Economy of scale“: „size“ can be measured related to inputs as well as outputs. Scale is seen as relating to “productive capacity“ (Brinkmann and Leslie 1986)

What is input: individuals, teams, networks, departments, universities…?

What is output: single publication, quality-controlled publication, journal impact factor, citation ….and patents, spin-offs… (see SPRU 2003)

Publication indicators – an issue?

- Dominance of referred journals
- Impact-Factor: Modification of publication strategies
- Disciplinary features insufficiently addressed
- Little attention towards different kinds of research orientation
- Individualization of science performances

Different disciplinary perspectives

(The Metric Tide, 2015)

Area studies: Capturing metrics data for both outputs and impacts has proved very difficult in area studies;

Biological sciences: Citation metrics can be helpful as a last resort to inform borderline decisions but are not currently seen as widely useful;

Built environment: Some disciplines are more inclined to use quantitative data but they are in a minority. The use of metrics for assessment of architecture is flawed – most outputs are buildings, design projects, books, etc, which don't fit into metrics;

Computer science: There are significant problems relating to coverage of citations by providers, for instance, indexing conference proceedings. Other computer science outputs include software, which are poorly captured. Downloads might be one option but it is unclear what these say about the excellence of research;

Education: It was suggested that some quantitative measures in research assessment are appropriate, but there was a risk that reviewers might use metrics disproportionately within the peer review process;

Performing arts: There is no formalized process of outputs, so a metrics-based approach based on this assumption would be unsuitable. More discursive elements of assessment would be welcome in these disciplines;

Physics and **epidemiology**: Very large author groups can be an issue. Currently 'team science' and collaborative research is not well rewarded. It would be worth exploring whether metrics could address this. Current metrics and methods of assessment can create tensions in research practices for some disciplines;

Psychosocial studies: There is an important question about why papers are cited and how to interpret the meaning of high citation counts – for example, something written provocatively can be cited many times despite being a paper well known to be poor. There are also issues about use of metrics in people's individual references, when these are not necessarily comparable and produce certain kinds of gaming and individualistic culture.

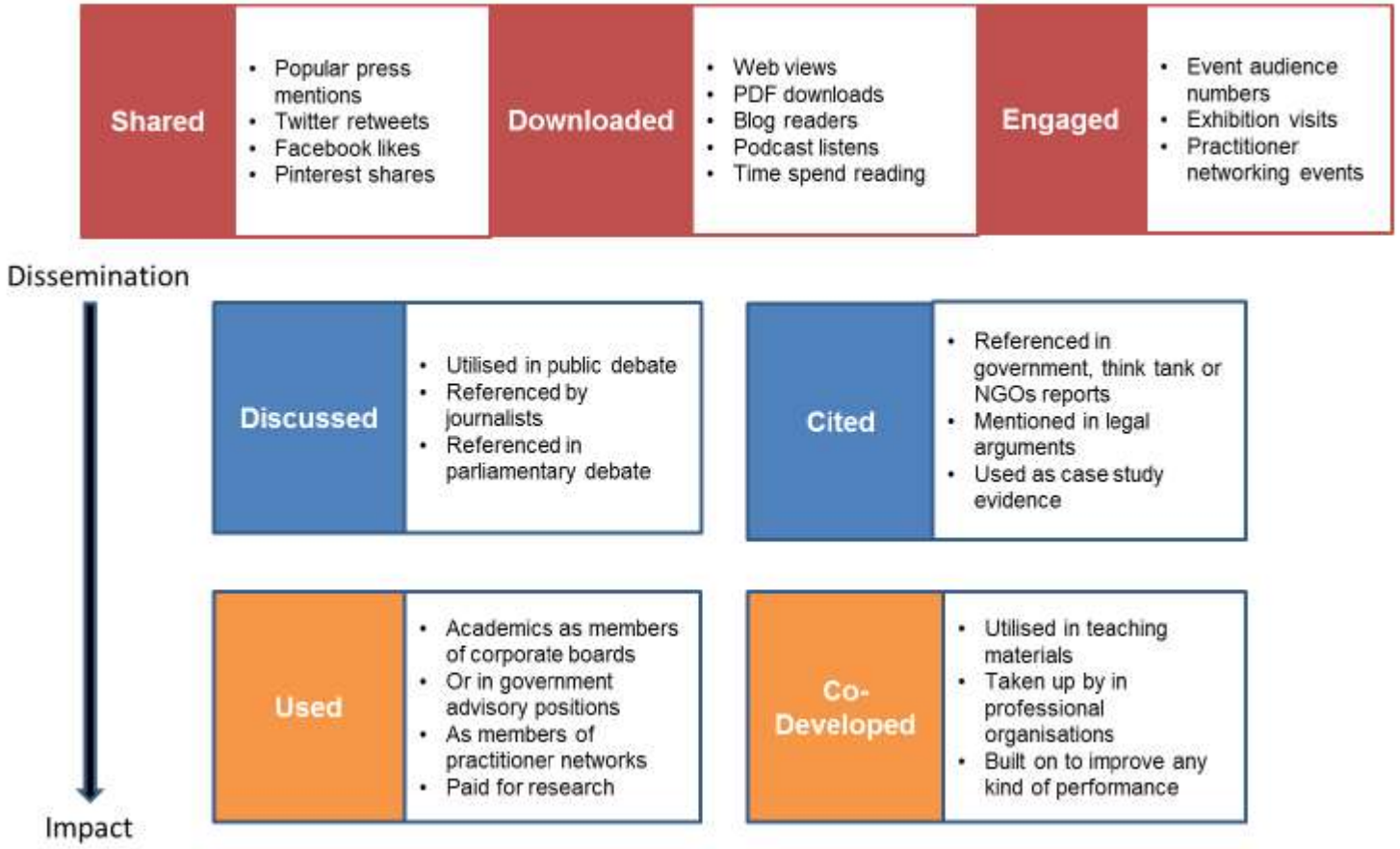
Source: Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. HEFCE

Measuring of impact

Impact is a contested term, with a variety of definitions and understandings of its implications ... In order for impact metrics should be developed, such information would need to be expressed in a consistent way, using standards units. However … the strength of the impact case studies is that they allow authors to select the appropriate data to evidence their impact“

(*The Metric Tide, 2015*)

Examples of the types of impact metrics tracking how research has been used

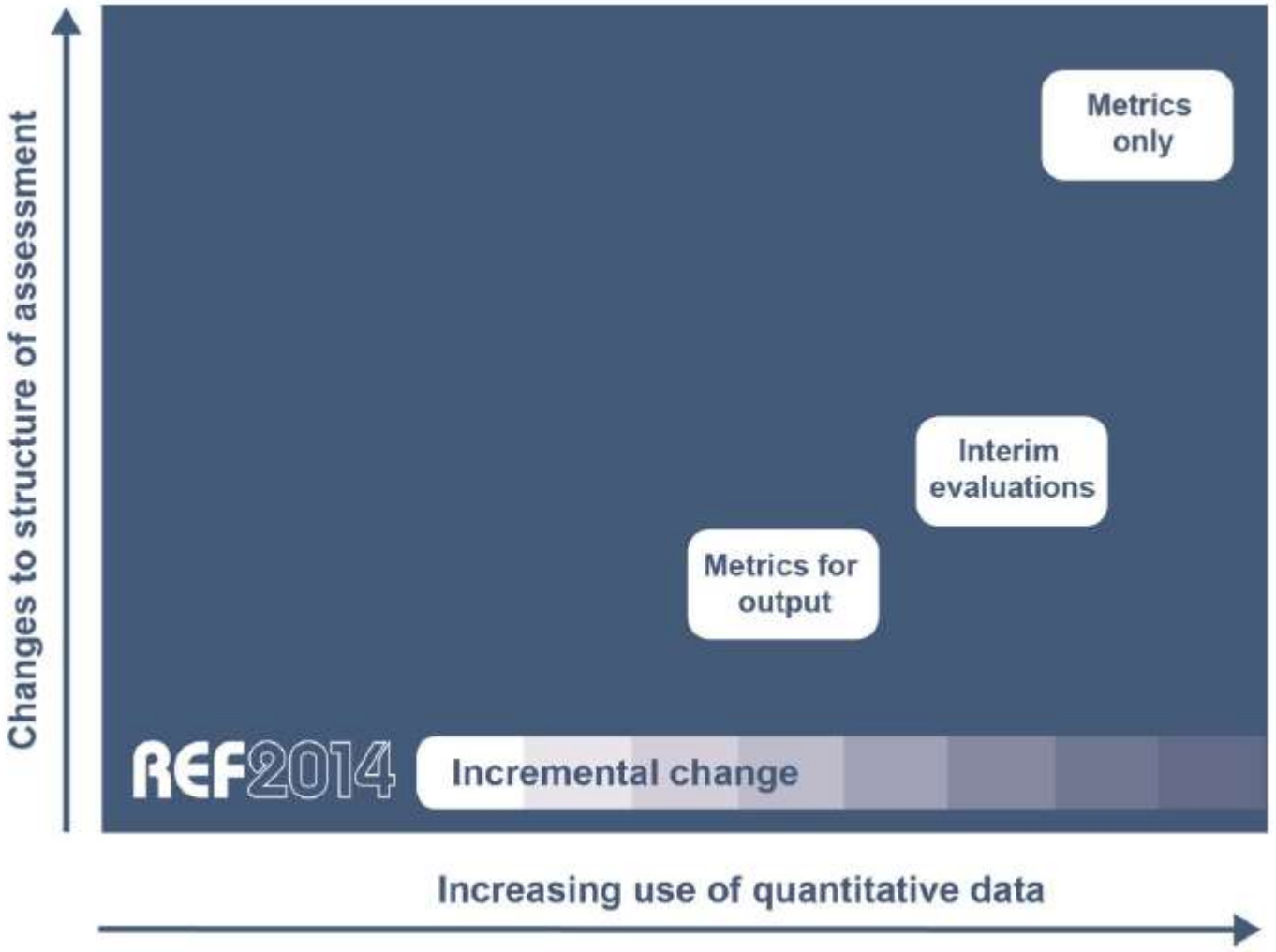


Source: Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. HEFCE

Conclusions and Recommentations of the Metric Tide Report (2015)

- Danger in rushing to over-interpret the available data
- Metrics should support, not supplant, expert judgement
- One size is unlikely to fit all
- Indicators can only meet their potential if there are underpinned by an open, transparent and coherent data infrastructure
- Inappropriate indicators create perverse incentives
- Correlation analysis has shown that individual metrics give significantly different outcomes from the REF peer review process
- Need of „science of science policy“

Options for the greater use of quantitative data in national assessments



Source: Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. HEFCE

Evaluation Criteria for Interdisciplinary Integration

Rating for individual criterion	Poor 1	Fair 2	Good 3	Very good 4	Excellent 5	Score
Integration Does the proposal address a holistic topic and present an integrated framework to approach to that topic?	The proposal makes no obvious attempt to create an integrated framework by combining different disciplinary knowledge and methods in the proposed study Or, a language of integration may be present but is very mechanistic or superficial at best	The proposal attempts to develop an integrated framework by drawing from different disciplinary knowledge bases and methods However, it does not integrate the elements of that framework in a generally coherent and effective way. In some instances, disciplinary concepts, theories, methods, etc. may be placed side by side; connections and analogies are made but no overall coherent integration is discernible	The proposal develops a framework by drawing from different disciplinary knowledge bases and methods, but some opportunities to advance the proposed study with this framework may be overlooked or undeveloped	The proposal successfully develops a framework by drawing from different disciplinary knowledge bases and methods An integrated framework clearly brings disciplinary insights together in a coherent and effective way and takes advantage of the opportunities presented by the integration of disciplinary knowledge and methods to comprehensively address the proposed study	In addition to meeting the "very good" criteria, the integrated framework employs an imaginative, or well-articulated integrative device (e.g., a metaphor, a model, a complex causal explanation) and/or seems likely to yield novel or unexpected insights	

Evaluation Criteria for Interdisciplinary Integration

Rating for individual criterion	Poor 1	Fair 2	Good 3	Very good 4	Excellent 5	Score
Category 3: interdisciplinary integration of proposal						
Interdisciplinarity Does the proposal draw from different disciplinary literatures relevant to the proposed study?	The proposal is grounded in the literature of only one discipline	The proposal draws from the literature of two or more disciplines, but does not attempt to justify the inclusion of each or the connections between them Some of the included disciplines may not be relevant to the proposed study at all, and/or crucial disciplines may be missing	The proposal draws from the literature of two or more disciplines, but does not manage to justify or explicate the inclusion of each or the connections between them Some of the included disciplines may be only be tangentially related to the proposed study, and/or crucial disciplines may be missing	The proposal draws from the literature of two or more disciplines, and clearly articulates and justifies the inclusion of each and the connections between for the purposes of the study All of the included disciplines are relevant to the proposed study, and no crucial disciplines are missing	In addition to meeting the "very good" criteria, the proposal includes an original combination of disciplines that hold much promise for the proposed study The proposal applies an truly interdisciplinary knowledge structure to the proposed study	

Evaluation Criteria for Interdisciplinary Integration

(The Snowbird **Charrette: Integrative Interdisciplinary Collaboration in Environmental Research Design, S. 438)**

Rating for individual criterion	Poor 1	Fair 2	Good 3	Very good 4	Excellent 5	Score
Synthesis Is there a sense of balance in the overall composition of the proposal with regard to how the disciplines are brought together?	The proposal shows an imbalance in the way particular disciplinary perspectives are presented in light of the proposed study (e.g., particular disciplinary perspectives are given disproportionate weight for no obvious reason)	The proposal attempts to balance perspectives but this is built on artificial or algorithmic grounds rather than substantive ones (e.g., giving equal weight to each discipline studied irrespective of its substantive relevance to the problem at hand)	Disciplinary contributions to the proposal are generally balanced on substantive grounds in light of the purpose of the work. However, one or more aspects of the argument may be weakly addressed	Disciplinary contributions to the proposal are delicately balanced to maximize the effectiveness of the proposal in light of the purpose of the work	In addition to meeting the "very good" criteria, the presentation is elegant and coherent and there are no distractions in the building of the argument	
Category 4: overall summary						
Rigor						
Originality						
Breadth						
Depth						
Comprehensiveness						

This rubric has been adapted from the original rubric created by and currently under testing by Veronica Boix Mansilla, Liz Dawes, Carolyn Haynes & Chris Wolfe at the Harvard Interdisciplinary Studies Project. While HISP seeks to apply their original version of this rubric to high school and undergraduate writing assignments, they have agreed to "loan" it to us for modification and use for the assessment of graduate student research proposals

Markers of Success

(V.B. Mansilla, M. Lamont; K. Sato: Shared Cognitive-Emotional-Interactional Platforms: Markers and Conditions for Successful Interdisciplinary Collaborations, . 18)

Groups	Dimensions[b]	Primarily Cognitive[a]					Primarily Emotional		Primarily Interactive	
		Cross-disciplinary exchange C, (I)	Generativity beyond program C	Shared intellectual tools C, (I)	Excellent and relevant expertise C	Knowledge advancement C	Collective excitement E, C, I	Joy in collaboration E, I	Group deliberation and learning competency I, C	Meaningful relationships I, (C), (E)
A	$n = 11$	5	0	2	0	6	3	0	7	6
	%	46	0	18	0	55	27	0	64	55
B	$n = 9$	5	3	3	1	3	3	2	7	2
	%	56	33	33	11	33	33	22	78	22
C	$n = 7$	6	5	3	4	3	4	4	4	3
	%	86	71	43	57	43	57	57	57	43
D	$n = 6$	5	5	3	6	2	4	0	4	2
	%	83	83	50	100	33	67	0	67	33
E	$n = 5$	5	5	4	2	3	4	1	4	2
	%	100	100	80	40	60	80	20	80	40
F	$n = 7$	7	1	3	5	0	5	4	2	2
	%	100	14	43	71	0	71	57	29	29
G	$n = 7$	2	4	2	1	0	3	2	0	0
	%	29	57	29	14	0	43	29	0	0
H	$n = 5$	3	2	3	2	3	3	3	2	1
	%	60	40	60	40	60	60	60	40	20
Total	$N = 57$	38	25	23	21	20	29	16	30	18
	%	67	44	40	37	35	51	28	53	32

Note: Most relevant dimensions are listed first; secondary dimensions are listed in parentheses.
[a]Each marker is heuristically categorized for its most relevant dimension.
[b]Dimensions are cognitive (C), emotional (E), and interactive (I).

Some conclusions

Peer review is the best we have, but we can observe side effects

Also indicators are useful for evaluating science, but we have to prove continuously if they are adequate to the subject of evaluation

Disciplines and also inter- and disciplinary research matters

Try to combine internal und external evaluation procedures

You are on a good way!

WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Thank you very much for your attention!

Contact:
Dr. Dagmar Simon
dagmar.simon@wzb.eu
WZB Berlin Social Science Center
Science Policy Studies` Research Group
Reichpietschufer 50, 10785 Berlin
Phone: + 49-172-3151063