

Bibliometrically disciplined peer review: on using indicators in research evaluation

Peter van den Besselaar* and Ulf Sandström**

* Vrije Universiteit Amsterdam & TMC bv, Amsterdam, The Netherlands

p.a.a.vanden.besselaar@vu.nl, orcid.org/0000-0002-8304-8565

** Royal Inst Technology-KTH, Dept Indek, Stockholm, Sweden

[orcid.org/0000-0003-1292-8239]

Abstract

Evaluation of research uses peer review and bibliometrics, and the debate about their balance in research evaluation continues. Both approaches have supporters, and both approaches are criticized. In this paper, we describe an interesting case in which the use of bibliometrics in a panel-based evaluation of a mid-sized university was systematically tried out. The case suggests a useful way in which bibliometric indicators can be used to inform and improve peer review and panel-based evaluation. We call this ‘disciplined peer review’, and disciplined is used here in a constructive way: Bibliometrically disciplined peer review is more likely to avoid the subjectivity that often influences the outcomes of the peer and panel review-based evaluation.

Keywords:

peer review, panel review, bibliometric indicators, research evaluation,

Introduction

With the increased use of bibliometric indicators for research evaluation, also the critique on their use has become louder. Especially after the Leiden Manifesto (Hicks et al., 2015), the publication of the Metric Tide report (Wilsdon et al., 2015), and the DORA declaration (2012), the use of bibliometrics in research evaluation has been discussed extensively. The main points of the criticism are as follows.

- i. Bibliometric indicators are only reflecting a part of research output – and in research fields like the qualitative social sciences, humanities, and engineering only a *small* part (Mongeon & Paul-Hus 2016).
- ii. Bibliometric indicators cover at best some dimensions of quality, but by far not all (Van den Besselaar & Sandström 2019).
- iii. Bibliometric indicators have perverse effects as researchers and research organizations will try to play them, using strategies like salami-slicing papers and excessive self-citation, and even go that far as misconduct and manipulation (Biagioli & Lippman 2020);
- iv. The bibliometric digital infrastructure transforms the research system by goal displacement and task reduction (de Rijcke et al. 2016; Krüger 2020).

One conclusion from the critique could be that bibliometric indicators have changed the nature of research and that assessing the value of scholarly work is not anymore part of the academic debate, but increasingly takes place at the managerial level (Biagioli & Lippman 2020), negatively influencing the entire scientific enterprise. However, the empirical basis for such claims is, in our view, thin. Just to give a few examples, criticism on bibliometric indicators often picks out one or two measures (especially the Journal Impact Factor and the H-index) and generalizes the (possibly correct) criticism on those to indicators *in general* (Biagioli & Lippman 2020; Gingras 2020). Also, for goal displacement and the related perverse effects not too much evidence exists. The often-cited study on Australia (Butler 2003) suggesting that productivity indicators lead to higher productivity but at the same time to lower quality is shown to be wrong (Van den Besselaar et al. 2017). And finally, work about indicator related misbehavior also seems to rest on rather thin evidence. For example, Oravec (2019) discusses the “emerging practices (of indicator manipulation) ... and their linkages with the norms and processes that support academic celebrity and stardom as well as the character of academic systems (Oravec, 2019, p859)”. Oravec borrows the empirical evidence from a study by Van Bevern et al. (2016) on incorrectly merging publications in Google scholar to increase one’s H-index. But this study does show the opposite: this kind of manipulation can only be done when highly similar titles are ‘merged’, which affects the h-index only to a rather limited extent. And as bibliometric databases like Google Scholar are openly accessible, such fraud is easy to detect and therefore unlikely to happen often. Furthermore, fraud and manipulation existed also before the ‘metrics revolution’, as did self-correcting mechanisms (Broad & Wade 1982).

The alternative conclusion of the critique on bibliometrics is that bibliometric indicators can at best have a role in supporting peer review (Hicks et al., 2015; Wilsdon et al. 2015). However, how that ‘supporting role’ should look like, remains unclear. These reports make peer review, or ‘qualitative review by experts’ the standard - not the gold standard, as problems with peer review are acknowledged, but still the standard: peer review is not perfect, but the best we have. This is a widespread opinion in the scientific community (Van Raan 2005; Moed 2007).

This strong trust in peer review is remarkable, as decades of research have shown the problems with peer review and expert panel review (Chubin & Hackett 1990; van Raan 2005; Moed 2005; Bornmann 2011; Lee et al. 2013). Peer review is biased, is hardly able to select the best applicants (Neufeld & von Ins 2011; Brezis & Birukou 2018, Van den Besselaar & Leydesdorff 2009), and has a low predictive validity (Bornmann & Daniel 2005; Van den Besselaar & Sandström 2015). When the same paper or proposal is reviewed twice by different reviewers, the correlation between the scores - the inter-reviewer reliability - is very low (Cole & Cole 1979; Cole et al. 1981; Cicchetti 1991). Finally, conservatism as well as risk-avoidance is all over the place (Heinze 2008; Garcia et al, 2020).

In parallel, a range of studies has been done comparing outcomes of peer review with bibliometric rankings, with diverging results (e.g. Narin 1976; Martin & Irvine 1983; Moed et al. 1985; and Rinia et al. 1998; Moed 2005, ch 18.4; Aksnes & Taxt 2004; Harzing 2018). In this type of studies, the bibliometric indicators are considered validated when the two approaches lead to the same outcome. But the other way around, it can also be interpreted as a validation of peer review, which may be more reliable than the critical studies suggest.

A comprehensive approach was suggested by Martin & Irvine (1983) in a study on radio astronomy. They combined different partial indicators and compared bibliometrics with peer evaluation, and had to some extent similar results for both methods. Somewhat later, CWTS did the same type of studies, e.g. Moed et al. (1985) and Rinia et al. (1998), the latter being a systematic test of the

agreement hypothesis. However, the Dutch findings were more diverse, as there was not a perfect correlation between peer assessments and bibliometric indicators. A difference between the UK study and the Dutch study was that the former deliberately included a large number of peers, in order to get rid of some of the statistical noise which comes with only a few peers per panel.

The agreement between bibliometric scores and peer review scores in the older studies may require re-examination, as they used indicators that would not have been chosen today. Rinia et al. (1998) used size independent indicators, whereas one now would probably prefer size dependent indicators, which may lead to very different results. Already back in 1997, Wennerås & Wold (1997) suggested that when using a size dependent indicator like total impact, the peer review scores were strongly influenced by applicant-reviewer relations (nepotism) and gender (sexism).

Recently, Harzing shows in an interesting paper that she could carry out the UK evaluation system *REF on a rainy Sunday afternoon* (Harzing, 2018) using bibliometric indicators. She does not argue that bibliometric indicators are a better representation of the performance of research units than the peer review outcome, but only that the two approaches lead to more or less the same outcomes. That would be an argument to use bibliometrics as a much cheaper alternative, and have a panel doing an additional check by looking for possible unexpected and unlikely outcomes. However, it could also be the case that the outcomes of the two approaches are different, and then the crucial question is which approach provides the best reflection of research performance. For example, conservatism in peer review may lead to discarding innovative and promising new approaches. But, does not the same hold for bibliometric indicators? Low citation impact may not only refer to moderate important research work, but could also reflect not (yet) recognized innovations (Wouters 2020).

This is indeed a problem, which can be translated into the need for additional indicators. In fact, the main criticism on the bibliometrics community may be that the focus has been too long on indicators for productivity and impact, while neglecting other important quality dimensions such as newness, originality, or independence. Research groups or researchers with equal publication and impact scores can perform very different on other important quality dimensions, for which indicators are often missing. An example is independence. One of the main characteristics of excellent scientists is independence: being able to develop new lines of research and not remaining in the research lines of the early career that in most cases were defined by the PhD supervisor or by the unit where the researcher started his/her career. In a recent paper we developed an indicator to measure the level of independence of mid-career researchers, and showed that this can help to distinguish between researchers that seem to be equally productive and have equal impact (Van den Besselaar & Sandström 2019). Another example is newness, and several attempts have been made recently to develop indicators for newness, but this has not yet resulted in accepted indicators (e.g. Bornmann et al. 2019; Bornmann et al. 2020; Shibayama & Wang 2019; Wu et al. 2019).

Today's reality is that peer review and bibliometric assessment are not anymore two separate activities – in practice they have been merged: many peer reviewers and review panel members use bibliometric databases like WoS-Clarivate, Scopus, Dimensions, Microsoft Academic, Google Scholar, or even ResearchGate to obtain an impression of the applicants or research units they need to evaluate (Moed 2005, ch. 18.3; c.f. de Rijcke & Rushforth 2015). For professional bibliometricians it is obvious that this non-professional use of bibliometrics may lead to serious problems, as the latter often uses indicators like the journal impact factor and the H-index, which are considered flawed by the former. But among the 'amateur bibliometricians', there is much disagreement too. In a recent

research project¹, some 35-panel members were interviewed, to find out which criteria were deployed. Many reviewers indicated that they did use bibliometric data², but were complaining that colleague panel members were using the wrong indicators (e.g., citation counts versus the impact factor). This type of bibliometric checks can be expected to influence the assessment.

Summarizing, we can characterize the situation as follows.

- Peer review has big problems, as it suffers from bias, subjectivity and conservatism, and lacks predictive validity.
- Bibliometric indicators are not covering all research output and are not covering all quality dimensions.
- Bibliometric indicators are intensively used (and discussed) in the peer review practice by panel members and peer reviewers – but probably not always in a valid way.
- Advanced indicators are available, but need to be brought into the process explicitly.
- For several quality dimensions, indicators are lacking and need to be (further) developed.
- There is a general agreement that indicators should be included in the evaluation as an input, and that the panel members have the last word. This also reflects the position of the authors.

It seems as if the discussion on peer review and bibliometric indicators has not made much progress, probably because it has remained a debate on possible risks, without much empirical work on how the combination of the two strategies works out in practice. How ‘bibliometrically informed panel/peer review’ could and should look like, remains understudied. There are hardly examples of case studies to show how indicators function in the practice of valuing and evaluating, and with which consequences. Without this, it is difficult to develop models of how indicators could be used to improve evaluation. That such studies are lacking can be easily understood, as what happens inside panels is generally confidential and not accessible for research.

This paper contributes to filling that gap, by investigating how the use of bibliometric indicators works in practice. So we do not compare the panel scores with the bibliometric scores – a relation which is often done ex post (e.g., Rinia et al. 1998; Oppenheim 1996; Oppenheim 1997; Harzing 2018). Instead we describe the *process of informed peer review*. We analyze a case of panel review where a systematic inclusion of indicators was a core part of the process. The case is about a small to medium-sized Swedish university, where the entire research portfolio was evaluated by a committee of fourteen scholars who collectively covered all research fields present in the university. Our study is strongly helped by the fact that almost all information about the evaluation is publicly available, including a description and reflection of the panel activities and processes by the chair of the panel. The draft of that five-page report was circulated among the panel members who provided suggestions for revisions, and it can be considered as a consensus view of the panel. Finally, it helped much that the two authors of this article were involved in the evaluation process in different roles³, and therefore had access to the only thing that is not publicly available, but crucial for the evaluation of the effect of the bibliometric indicators: the initial scores and reports on the units. This enabled the current authors not only to analyze the panelists’ view of bibliometric indicators, and how these

¹ GendERC project 2014-2016 (FP7-IDEAS-ERC Grant agreement ID:610706).

² Generally, from the online version of WoS, implying that they didn’t realize the need for field normalization for reliable indicators also within the boundaries of ERC panels.

³ The first author was member of the evaluation panel, the second author had provided the university with an in-depth bibliometric assessment of the research.

indicators were discussed and used by the panel deliberations, but also to assess the effect of these bibliometric indicators on the final scores, without uncovering the confidential initial reports and scores.

The case

The case is a small to medium-sized Swedish university (Örebro University), with about 12.000 students and 300-400 researchers. Research is performed in many subfields of medicine, natural sciences, psychology, law, economics, 'soft' social sciences and humanities, and parts of computer science & engineering. There are three faculties at this state university: Faculty of Economic, Natural and Technical Sciences; Faculty of Humanities and Social Sciences; and Faculty of Medicine and Health. Since its foundation in the mid-1970s, Örebro University has been characterized by profession-oriented education.

Research at the university was evaluated five years earlier too (ÖRE2010). That evaluation was organized for 38 Units of Assessment (UoA). A few successful units were given strategic resources over the following five-years period. A policy was adopted to increase the scientific output of the university, as the aim was to change the regional hospital into an academic hospital with a medical school. In 2011, the university was granted the right to have a Medical faculty and exams in medicine. That was an important step in the development of the university which had started as a university college with university courses mainly in social work, social science, and humanities. Over the years the college built up a capacity for research and was it granted the status of university in 1999.

The university has a relatively small research budget, which of course needs to be taken into account when evaluating research performance. The budget in the period of 2012-2014 was a bit above 1.2 billion SEK per year of which about two third is for teaching and one third for research. The university is ranked as one of the 401-500 leading universities in the world in the Times Higher Education ranking of the world's leading universities in 2019. The university is also ranked 74th out of the 150 best young (under 50 years) universities in the world. However, it is not ranked in the Leiden Ranking.

In 2015, the board, as well as the outgoing vice-chancellor, wanted a follow-up evaluation of all research in the university. The structure of the evaluation was as follows. A panel of 14 members was formed with some 25 % foreigners, covering all disciplines. The evaluation procedure was conceived as a meta-evaluation (ORU2015, 19 ff) based on extensive information about the research units, but without interviews with the research units and without reading publications authored by members of the research units. The evaluation had to cover the following aspects of research performance:

- A. The quality of research
- B. The research environment and infrastructure
- C. Scientific and social interaction
- D. Future potential.

The committee was asked to do the evaluation using several pieces of information: A self-evaluation report, written by the research units; a letter and a presentation of the deans of the faculties; a bibliometric report at the individual level and the unit level, as well as a summary of the bibliometric study.

Although the evaluation dimensions are similar, there are clear procedural differences with the way the research evaluation is done in other countries. Without being exhaustive, a few things should be mentioned here. First of all, in Sweden no national system for research evaluation exists, such as the Netherlands and the UK. In the latter countries, research evaluation is done at the discipline level, and not for the university as a whole. In the UK, it is done by national disciplinary committees, doing peer review of a selected set of researchers and core publications. In the Netherlands, all units in a field can be evaluated at the same moment by the same panel, but this is not necessarily the case. The Dutch evaluation is based on a self-evaluation report like in our Swedish case, which often includes bibliometric indicators of the performance of the unit(s). Although units mention five core publications, reading and evaluating the publications not part of the research assessment – similar to the Örebro approach. In contrast to the Örebro approach, the Dutch research assessment protocol includes a site visit where the committee talks with a variety of representatives of the unit under evaluation: PhD students, junior staff, senior staff, and the department and faculty management.

In our case the following information was provided:

- (i) The self-evaluation was the main piece of information, and it included a description of the research program and projects, grants and staff and sometimes the results, as well as organizational embedding of the research. These ‘self-evaluations’ were of different quality and content, as the units could decide themselves about the content and the format. The quality and level of detail of the self-reporting seemed to reflect the quality of the research. The self-evaluation consisted of the following topics, albeit in different details: 1) a self-assessment on the four performance dimensions mentioned above; 2) an overview of the research projects and/or teams within the unit; 3) an overview of the staff, budgets and grants.
- (ii) Apart from the self-evaluation reports, the university board had asked the second author of this paper to produce at the unit level and at the individual level a set of (Web of Science-based) indicators, including the numbers of (full and fractional) publications, the field normalized citation scores (with and without a time window), the share of top 10% (PP10%) most cited papers (and top 1%, 5%, 25% and 50%), as well as indicators for the average number of co-authors and the average number of international co-authors. In order to avoid as much as possible discussions about whether the WoS based indicators can be used at all in research evaluation, the bibliometric report also used publication counts in terms of the ‘Norwegian model’ (Sivertsen 2018). This means that publications were not restricted to WoS indexed journals, as this includes a much larger output coverage based on the universities’ publications repository DIVA. A similar valuing system was used as in Norway, where publications points were derived differently depending on the *quality class* they belong to. The innovation for this evaluation was that (1) reference values were developed for the DIVA scores based on disambiguated researchers from the main universities in Sweden (called DIVA PP REF) and (2) that the list of journals was even more expanded, based on Ulrich’s list of periodicals (c.f. Sandström 2014). This resulted in scores that can be seen as an alternative for the citation-based scores. Not all research units had data available to do this, especially for the research units in the university hospital – the main reason being that the hospital was not aligned with the university library at the time of the evaluation. Table 1 shows an example of the bibliometric results.

UOA & SUB-UNIT		WEB OF SCIENCE 2008-12																DIVA 2008-14						
UOA	SUB-UNIT	Full P	Frac P	NCSj	NJCS	NCSf	TOP1	TOP5	TOP10	TOP25	TOP50	PNC	IntCO		FAP	Personnel	PROD	DIVA Level1	DIVA Level2	DIVA %Level1	DIVA %Level2	DIVA PP	DIVA REF	DIVA PP/REF
EDUC&SOCIAL		64	42,0	1,02	0,84	0,94	0,0%	4,8%	10,0%	27,1%	42,9%	27,8%	1,7	1,3	45,4	63	0,7	281	115	56,8%	43,2%	5,0	7,1	0,7
	CULT DIVERSITY	2	2,0	0,00	0,16	0,00	0,0%	0,0%	0,0%	0,0%	0,0%	100,0%	1,0	1,0	2,5	2	1,3	23	5	71,0%	29,0%	9,9	7,1	1,4
	EDUCATION	9	5,2	0,51	0,72	0,54	0,0%	0,0%	7,4%	10,4%	30,6%	54,8%	1,7	1,1	6,1	19	0,3	63	12	68,6%	31,4%	3,1	7,1	0,4
	GENDER	16	9,8	1,03	0,66	0,79	0,0%	0,0%	2,6%	28,7%	43,2%	25,6%	1,8	1,8	11,2	11	1,0	44	58	29,2%	70,8%	7,9	7,1	1,1
	GEOGRAPHY	3	1,0	2,34	0,67	1,81	0,0%	16,7%	16,7%	40,8%	66,7%	0,0%	3,0		1,1	5	0,2	8	2	65,1%	34,9%	0,9	7,1	0,1
	POLITICAL	10	7,8	0,57	1,07	0,93	0,0%	6,4%	12,8%	25,5%	42,2%	38,3%	1,7	1,2	8,8	10	0,9	58	17	53,7%	46,3%	6,1	7,1	0,8
	SOCIOLOGY	24	16,2	1,43	0,97	1,24	0,0%	8,4%	14,9%	34,7%	50,8%	8,2%	1,5	1,1	15,8	16	1,0	85	21	75,8%	24,2%	5,2	7,1	0,7
ENGINEERING		48	21,1	0,99	0,89	0,80	0,0%	0,0%	5,6%	24,9%	50,5%	27,6%	3,1	1,4	19,2	27	0,7	278	12	87,7%	12,3%	3,5	6,2	0,6
	COMPUTER	41	18,5	1,10	0,93	0,91	0,0%	0,0%	6,3%	28,4%	57,6%	21,2%	3,1	1,4	16,9	21	0,8	269	12	87,2%	12,8%	4,4	6,2	0,7
	MECHANICAL	7	2,6	0,20	0,54	0,05	0,0%	0,0%	0,0%	0,0%	0,0%	74,2%	2,7	1,2	2,2	6	0,4	9	0	100,0%	0,0%	0,6	6,2	0,1
HEALTH SCI		211	78,4	1,02	0,83	0,85	0,0%	2,8%	5,2%	22,5%	48,5%	20,8%	3,1	1,3	56,8	62	0,9	595	90	68,7%	31,3%	4,4	6,2	0,7
	CARING SCI	86	29,4	1,11	0,82	1,01	0,0%	5,7%	8,3%	23,2%	51,0%	13,6%	3,2	1,2	20,4	26	0,8	233	36	69,5%	30,5%	3,8	6,2	0,6
	OCCUP THERAPY	21	6,9	0,68	0,66	0,49	0,0%	1,8%	3,6%	5,4%	28,0%	31,1%	3,0	1,1	4,7	6	0,8	58	1	95,7%	4,3%	2,9	6,2	0,5
	PUBLIC HEALTH	28	11,6	1,11	0,68	0,50	0,0%	0,0%	0,9%	9,8%	32,9%	41,7%	3,1	1,4	9,2	9	1,0	104	2	94,4%	5,6%	4,0	6,2	0,6
	SIDR	38	13,3	0,85	0,89	0,72	0,0%	0,0%	2,9%	22,6%	43,6%	34,0%	3,1	1,2	10,7	9	1,2	104	9	82,9%	17,1%	3,9	6,2	0,6
	SPORT SCI	44	17,1	1,08	0,96	1,07	0,0%	2,4%	5,5%	36,9%	66,7%	4,5%	3,1	1,4	11,9	12	1,0	96	42	45,4%	54,6%	7,0	6,2	1,1
HUMANITIES		58	28,5	1,10	0,92	1,29	0,2%	2,4%	10,8%	20,5%	53,8%	25,2%	2,2	1,4	31,0	55	0,6	202	132	42,2%	57,8%	6,2	7,1	0,9
	CULINARY	29	5,5	0,84	1,10	0,87	0,0%	0,0%	5,8%	24,2%	56,0%	12,4%	5,3	2,7	4,2	9	0,5	53	59	16,1%	83,9%	9,7	7,1	1,4
	HISTORY														0,0	6	0,0	18	1	98,9%	1,1%	3,8	7,1	0,5
	HUM-ARTS	4	3,5	1,10	0,96	3,17	1,2%	14,3%	14,3%	28,6%	42,9%	42,9%	1,7	1,0	3,3	9	0,4	28	15	49,9%	50,1%	6,0	7,1	0,8
	LANG STUD														0,0	14	0,0	18	1	60,0%	40,0%	1,4	7,1	0,2
	MEDIA & COMM	25	19,5	1,17	0,86	1,08	0,1%	1,0%	11,6%	17,9%	55,2%	25,6%	1,4	1,1	23,5	14	1,7	65	55	34,7%	65,3%	9,7	7,1	1,4
	Rhetorics														0,0	3	0,0	20	1	97,7%	2,3%	7,3	7,1	1,0
JPS		139	55,1	1,07	1,14	1,33	2,6%	7,9%	12,7%	29,5%	56,7%	17,5%	3,3	1,4	41,8	65	0,6	500	138	67,7%	32,3%	6,6	7,1	0,9
	CHAMP	85	32,1	1,09	1,13	1,32	3,4%	6,7%	12,2%	30,4%	58,4%	16,6%	3,4	1,5	23,2	16	1,5	172	54	61,4%	38,6%	6,4	7,1	0,9
	CRIMINOLOGY	12	3,6	0,88	1,33	1,16	0,0%	2,2%	9,8%	39,1%	57,2%	13,8%	3,9	1,4	3,1	5	0,6	52	25	36,6%	63,4%	9,3	7,1	1,3
	LAW														0,0	22	0,0	138	28	78,9%	21,1%	8,8	7,1	1,2
	SOCIALWORK	8	3,4	0,40	0,93	0,44	0,0%	0,0%	3,7%	3,7%	21,4%	37,0%	3,3	1,0	3,6	12	0,3	47	11	64,8%	35,2%	2,6	7,1	0,4
	YOUTH STUDIES	44	16,1	1,21	1,16	1,58	2,2%	13,3%	16,3%	30,9%	60,5%	16,1%	3,0	1,2	11,8	10	1,2	91	22	67,7%	32,3%	5,3	7,1	0,7
MEDICAL SCI		577	130,8	1,10	1,02	1,16	2,1%	7,6%	12,0%	27,1%	53,1%	8,3%	5,1	1,5	90,5	83	1,1	747	195	66,3%	33,7%	3,4	6,2	0,6
	BIOMEDICINE	121	30,8	0,66	0,82	0,53	0,0%	0,6%	2,1%	8,1%	33,6%	9,6%	4,5	1,2	21,4	26	0,8	171	19	81,2%	18,8%	1,9	6,2	0,3
	MEDICINE	476	99,9	1,23	1,08	1,36	2,7%	9,8%	15,0%	33,0%	59,0%	7,9%	5,3	1,6	69,1	57	1,2	576	176	63,2%	36,8%	4,1	6,2	0,7
NATURAL		206	60,6	1,11	1,17	1,40	1,4%	12,5%	19,1%	37,8%	58,8%	12,3%	4,5	1,5	42,7	39	1,1	343	164	42,4%	57,6%	4,9	6,2	0,8
	BIOLOGY	57	13,4	1,13	1,06	1,05	0,0%	7,6%	11,1%	27,4%	61,1%	13,0%	5,1	1,6	9,3	10	0,9	95	36	52,2%	47,8%	3,6	6,2	0,6
	CHEMISTRY	120	30,5	1,35	1,28	1,86	2,4%	19,0%	29,7%	50,2%	66,9%	6,5%	5,2	1,6	20,6	15	1,4	222	101	41,2%	58,8%	6,9	6,2	1,1
	MATH-PHYS	41	16,7	0,65	1,08	0,82	0,6%	4,5%	6,1%	23,4%	42,2%	22,5%	2,5	1,3	12,8	14	0,9	26	27	37,5%	62,5%	3,5	6,2	0,6
SCH BUSINESS		102	63,3	0,73	0,99	0,72	1,0%	1,3%	3,0%	14,0%	43,5%	28,7%	1,9	1,1	67,1	53	1,3	324	69	64,0%	36,0%	5,0	7,1	0,7
	BUSINESS ADM	34	19,8	0,57	1,05	0,63	0,0%	0,4%	2,5%	9,6%	48,7%	26,9%	2,1	1,2	22,6	22	1,0	113	28	64,6%	35,4%	4,9	7,1	0,7
	ECONOMICS	51	34,4	0,88	0,98	0,82	1,9%	2,2%	4,1%	18,0%	42,4%	30,0%	1,6	1,1	35,0	18	1,9	95	18	65,1%	34,9%	4,3	7,1	0,6
	INFORMATICS	17	9,1	0,53	0,87	0,55	0,0%	0,0%	0,0%	8,7%	36,0%	27,4%	2,5	1,2	9,6	13	0,7	116	23	62,2%	37,8%	6,0	7,1	0,8

Table 1: Excerpt from the bibliometric report covering all units of assessment (from ORU2015).

Table 1: legend*

UOA	= Unit
Subunit	= Research program (evaluation level)
Full P	= Number of full papers (articles, letters and reviews) during 2008-2012.
Frac P	= Number of papers, fractionally counted
NCSj	= Journal Normalized Citation Score (CPP normalized in relation to the unit journal set - average=1.00)
NJCS	= Normalized Journal Citation Score (cf. SNIP)
NCSf	= Field normalized Citation Score (CPP normalized in relation to the sub-field set - average=1.00).
TOP1	= Share of publications in top 1 % cited papers, field normalized (Percentage of papers above the 1 th citation percentile).
TOP5	= Share of publications in top 5 % cited papers, field normalized
TOP10	= Share of publications in top 10 % cited papers, field normalized
TOP25	= Share of publications in top 25 % cited papers, field normalized
TOP50	= Share of publications in top 50 % cited papers, field normalized
PNC	= Percentage Not-Cited papers
AUm	= Average number of authors
IntCOLLm	= Average number of international co-authors
FAP	= Field Adjusted Production (number of papers, field normalized)
Personell	= Number of staff members in the UoA/Subunit
PROD	= FAP per Research Personnel in the UoA/Subunit
DIVA Level1	= Fraction of Publications in level 1 (lower) media
DIVA Level2	= Fraction of Publications in level 2 (higher quality) media
DIVA %Level1	= Share of Publications in level 1 media
DIVA %Level2	= Share of Publications in level 2 media
DIVA PP	= Average Publication Points per Person in the UoA/Subunit
DIVA REF	= Reference value for Publication Points (field normalized to Swedish universities)
DIVA PP/REF	= Average Field Normalized Publication Points

* How each of the indicators are calculated, is described in the evaluation report ORU2015, pp. 303-312.

- (iii) However, the bibliometric report was very thick, and therefore not too user friendly. One of the panel members, the first author of this paper, was asked by the university board to produce a summary of the bibliometric report – a summary that would consist of an explanation of the meaning of the indicators, an explanation of how to interpret the data, and a short summary per research unit in those terms. Figure 1 shows an example of the text in the bibliometric summary. Annex 1 presents the details of how the scoring was done.
- (iv) In the weeks before the evaluation, each of the units was studied more in-depth by two panel members, who both prepared a draft evaluation text about the unit. That *pre-evaluation* described the strong and weak points and suggested a score.

<p><i>Research unit X:</i></p> <ul style="list-style-type: none"> - Medium sized (15 persons) productive (1.5 publications per person) unit, with excellent (1.32) citation impact and very good share (6.7%) in the top 5% cited papers. A high (1.14) vitality. - In DIVA, the unit scores at the Swedish average. <p><i>Research unit Y:</i></p> <ul style="list-style-type: none"> - A medium sized (15 persons) productive group (1.4 publication/person), with a very high citation impact – almost twice world average (1.86). The unit also has a high share (19%) in the top 5% cited papers. This is accompanied with a high (1.14) vitality of the research. - In the DIVA data, the unit is at Swedish average.

Figure 1: Example of a summary of the bibliometric report

Summarizing, the panel received for each unit (i) a self-evaluation report; (ii) a report with extensive bibliometric data; (iii) a short summary of the bibliometric data; and (iv) some additional information like letters of the dean(s), and an overview of teaching, and (v) two pre-evaluations by panel members. In some cases, the information pointed in the same direction, but often the scores were rather different. In the latter cases, the panel went in-depth into the arguments provided and discussed the differences extensively in order to get to a consolidated score. Overall, this resulted in consensus, as panelists were able and willing to change their views under the influence of the debate. In this consensus formation, the summary of the bibliometric report was heavily used, but it also had another advantage that we will discuss below.

Can WoS data be used in all fields?

The role and relevance of bibliometric information were discussed within the evaluation panel, using the common arguments (Brändström 2015). Naturally, the main issue was whether bibliometric data cover research output in a reasonable way. Also here, this issue came up for the social sciences and humanities and for computer science. Incomplete coverage, however, is only a problem if the WoS coverage is not representative for the total output of the units. From this perspective, it was a real advantage that apart from the WoS based indicators, the bibliometric report also included publication data from the national repository, including a much wider set of publications in two 'quality classes' (based on an enlarged version of the Norwegian system). The WoS indicators were calculated against the world average and the DIVA indicators against the *national* field averages using a number of researchers (50-100 per field) at Swedish universities. Table 2 summarizes the differences and commonalities between the two bibliometric indicator sets.

Table 2: Unit scores in WoS and DIVA

Unit	Group*	Field	WoS	DIVA
1	1	Biomedicine (hospital)	excellent	n.a.
2		Surgery	very good	n.a.
3		Disability	good	n.a.
4		Medicine	good	n.a.
5		Nursing	moderate	n.a.
6		Occupational health	weak	n.a.
7	2	Medicine	excellent	weak
8		Biology	good	weak
9		Nursing (Caring)	good	weak
10	3	Chemistry	excellent	good
11		Youth studies	excellent	weak
12		Psychology (CHAMP)	very good	good
13		Political science	good	weak
14		Sociology	good	weak
15		Math/Physics	good	weak
16		Economics	good/moderate	weak
17	4	Musicology (Arts & hum.)	moderate	weak
18		Geography	moderate	weak
19		Computer science	moderate	weak
20	5	History	weak	weak
21		Language studies	weak	weak
22		Social work	weak	weak
23		Education studies	weak	weak
24		Business administration	weak	weak
25		Mechanical engineering	weak	weak
26		Occupational therapy	weak	weak
27		Public health	weak	weak
28		Disability	weak	weak
29		Biomedicine (university)	weak	weak
30	6	Media & communication	good	very good
31		Sport science	moderate	good
32		Informatics	weak	moderate
33	7	Criminology	moderate	very good
34		Culinary	moderate	very good
35		Cultural diversity	weak	very good
36		Gender studies	weak	good
37		Law	weak	good
38		Rhetoric	weak	good

* Group 1 consists of the university hospital; the others belong to the university

What do we observe? First of all, the medical research units in the university hospital only use Web of Science (units 1-6) and have no DIVA score. The next three units are university-based, but have very low DIVA scores, suggesting that they hardly registered their publications in the DIVA database. The same holds for the third group of seven units (10-16) which are also in fields that are highly dominated by international journal publications. The next 13 units in three groups have (almost) similar scores in WoS and DIVA. So for these 29 units, WoS seems to provide the relevant indicators for assessing the contribution to international science.

Only the last two groups of nine units show a somewhat or much higher score in DIVA than in WoS, and we will inspect those more carefully.

- *Media & communication (30), Sport science (31), and Informatics (32)* show a slightly higher score in DIVA than in WoS,
- *Criminology (33)* and *culinary arts and meal science (34)* score moderate in WoS and very good in DIVA – also in the highly ranked media (level 2). This suggests that the WoS score may underestimate the performance of these two units.
- In the *Gender studies (36)* unit, two-third of all publications were authored by one researcher only, a fixed-term visiting professor. All others score rather low in terms of productivity. The publications have a low impact in WoS. The high score in DIVA is based on many papers in the higher classified media.
- The *Law (37)* unit is large with its 22 members, and they have no international publications in WoS, which was unexpected as the theme of the unit is international law. The DIVA score (1.2) is good, but 85% of the papers are in the lower classified media.
- *Cultural diversity (35)* and *Rhetoric (38)* studies score very good and good in DIVA, but most papers (85% and 95% respectively) are in the lower ranked media. Rhetoric studies had no international WoS publications. But most importantly, both groups were too small to be meaningfully evaluated (two and three researchers respectively).

Summarizing, only two out of the 38 units would probably receive a too low score if one would apply only WoS scores (nr 33 and nr 34). One may add three other units (30, 31, 32) with similar scores in WoS and DIVA, but somewhat higher in DIVA.

This does not mean that we consider the local publications or the non-journal publications as meaningless. A more detailed look for those groups where DIVA and WoS strongly differ may show that some international publications are neglected in WoS (e.g., international books). For the rest, the local (language) publications may function as knowledge dissemination to stakeholders.⁴ But for the moment, the comparison suggests that the WoS-based output counts are representative for the larger DIVA set if one measures contributions to international science. This was also the shared view of the panel (Brändström 2015).

The role of bibliometrics in the evaluation process

The use of the bibliometric data in the assessment basically boiled down to the *reality check* of the *pre-evaluations* prepared by the panel members. In cases where the scores proposed by the panelists were substantially higher or lower than the bibliometric data suggested, a discussion emerged. Those that had proposed those scores were asked to explain the difference. Often this was easy, as panelists indicated to have been somewhat too harsh, or (mostly) too friendly, and they also could easily explain why. For example, units may have been very good in the past and may have built up a strong reputation in the past that still influenced the evaluation. But as the evaluation covered recent (5 years) research only, the scoring based on reputation was far too high for the more recent research and was then lowered.

A second example is that panel members had been doing some bibliometrics, in order to inform themselves about the work of the unit. This is, by the way, a very general phenomenon that we also have seen in other studies. Using bibliometrics is not something that is ‘alien’ to science, but many researchers do accept that bibliometric indicators say something about performance. However, this is also risky, as we observed panel members bringing in the wrong data – thus forgetting that field

⁴ However, this often does not go through publications but via other channels (De Jong et al. 2012)

normalization should be done (which they generally cannot do themselves), and also forgetting to take the correct time frame for the publication counts. By having the 'correct' bibliometric indicators at hand, the discussions could generally be easily closed in a consensual way.

And in again other cases, the WoS scores were regarded as too low and the good DIVA scores were included to reach the final assessment score. In the previous section, we mentioned cases in which this happened.

Convergence of scores?

The deliberations were used to underline the scores units had received in the initial evaluation, or to change these scores. In which directions did the scores changed after the deliberation? Did the panel score converge to the WoS-based score? The question of convergence is relevant for 30 units.⁵ As table 3 shows, in 27 of these (90 %), the final panel score was closer to the bibliometric score than the initial score given by the reviewers, whereas in the remaining three (10 %) cases the final score was more different from the WoS score than the initial score was. Overall, there were good reasons for the divergence too. As emphasized in the previous section, in some cases the WoS indicators were felt to underestimate the quality of the units, which did have many papers in (the higher quality category of) the DIVA system. This shows that the panel assessment remains important.

Table 3: The changes between the initial and the final scores during the process

Convergence?	Total	Final score vs initial score	Total
Yes	27	Higher	5
		Equal	0
		Lower	22
No	3	Higher	2
		Equal	1
		Lower	0
WoS equal to the initial score	2	Higher	0
		Equal	2
		Lower	0
Doubt about the use of WoS *	4	Higher	2
		Equal	1
		Lower	1
No grade**	2	Higher	n.a.
		Equal	n.a.
		Lower	n.a.
Total	38	Higher	9
		Equal	3
		Lower	23

* The inspection of the bibliometric report suggested already that WoS may have been underestimating quality (see the previous section).

**No grade as unit was too small.

⁵ In two cases no evaluation was done as the units were too small, in four the WoS data were considered as not useable, and in two cases the WoS and DIVA scores were equal.

Table 3 shows also that in two-third of the cases, the final score was lower than the initial score and in a quarter of the cases higher than the initial score, showing that the panel deliberations and the bibliometric information did have an impact more often in lowering the scores.

Despite the convergence, the panel scores remained somewhat higher on average than the bibliometric scores, as Table 4 indicates. The panel hardly gave the score 1 (weak): only five times, whereas the bibliometric score 1 was given 16 times. And the DIVA score 1 even occurred 21 times – so the average level according to the DIVA data is lower than according to the WoS data. So, despite the fact that overall the panel adapted its initial score to the bibliometric score; it remained more moderate in its assessment.

Table 4: Comparison of the WoS, DIVA and panel scores

Score	WoS	Diva	Panel final score
5	4		2
4	2	4	8
3	8	6	12
2	8	1	9
1	16	21	5
average	2.21	2.01 ⁶	2.81

Conclusions

Overall, panel members considered the inclusion of the bibliometric data and the summary of the report as useful⁷. The summary was intensively used to reach the panel final consensus scores. As showed above, in many cases the WoS scores were sufficient bibliometric input to reach a final judgement – also in quite some fields within the social sciences and humanities. However, in several cases, the DIVA information functioned as a useful *supplement*. The bibliometric data helped to correct subjective views during the panel discussions, where all available information was used – which is exactly what should take place in a panel. As a consequence, the group dynamics played a less important role. Despite the common criticism on bibliometrics, which did also exist in the panel, the overall appreciation of the bibliometric indicators was rather positive (ORU2015, pp.19ff.).

Some may see the influence of the indicators on the outcome as a reduction of the freedom of peers to give their assessment of the different research units, and as an example of how bibliometrics is “exercising power”, and “forcing science” in specific directions, for example towards topics that are preferred by international journals. From another perspective, this “reduced freedom” could be seen as a positive instance in the evaluation system. Available information should limit the range of appreciation of the performance. That is why the title *bibliometrically disciplined peer review* was chosen for the current paper. In order to bring peer review to a level of disinterestedness and fairness (Merton 1973), and to avoid many of the problems of subjectivity and bias that research on peer review has reported, it would be a challenge for the bibliometric community to produce a larger set of valid indicators covering the more quality dimensions that are important when

⁶ The DIVA scores were on a 4-point scale, with an average of 1.78. After rescaling to a 5-point scale, the average becomes 2.01.

⁷ Expressed verbally and in emails to the panel members. This is also reflected in the panel chair’s report (Brändström 2015, pp.19ff.).

evaluating research, including quality indicators for applied research and societal impact. The current dominance of impact and productivity indicators is too narrow.

Some limitations need to be mentioned here. The participatory approach has resulted in detailed insights into an assessment process, but at the same time, in light of its confidential nature, it affects and limits to some extent what can be communicated in the current paper. For example, panel deliberations remain a social process in which opinions, interests and practical issues like time pressure may play some role (Van Arensbergen et al. 2014). This remains confidential, as are some data like the pre-evaluations. But the bibliometric study, the final assessment reports of all the units, and the full evaluation report are openly available. Together with our participation and observation of the process, this provides a rich and reliable picture. Furthermore, this paper comes with the limitations of a single case study. It would therefore be useful to have some more of these experiments to get a broader insight in how bibliometric data can be used to improve research assessment. Finally, the case presented in this paper is about evaluating universities and research units. In other contexts, such as evaluating individuals for hiring or promotion, or evaluating societal impact of research, it may work different. More work covering a variety of contexts may be useful.

Funding Information

The authors received no funding for this work.

Competing interests

The authors have no competing interests to declare.

References

- Aksnes DW & Taxt RE (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. *Research Evaluation* 13 (1): 33–41,
- Biagioli M & Lippman A (2020). *Gaming the Metrics: New Ecologies of Academic Misconduct* (MIT Press, 2020)
- Bornmann L (2011). Scientific peer review. *Annual Review of Information Science and Technology* 45 199-245.
- Bornmann L, Daniel HD (2005). Selection of research fellowship recipients by committee peer review: Analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, 63: 297–320.
- Bornmann L, Tekles A, Zhang HH, Ye FY (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. CoRR abs/1910.03233.
- Bornmann, L Devarakonda S, Tekles A, Chacko G (2020). Disruptive papers published in Scientometrics: meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019). *Scientometrics* 123:1149–1155 <https://doi.org/10.1007/s11192-020-03406-8>

- Brändström D (2015). "The Research Evaluation ORU2015 – The Panel's Perspective", In *ORU 2015*, pp. 19-23.
- Brezis E & Birukou A (2020). Arbitrariness in the peer review process. *Scientometrics* 123 (1): 393-411.
- Broad W & Wade N (1982). *Betrayers of the truth: Fraud and Deceit in Science*. Oxford: Oxford Univ Press.
- Butler L (2003). Explaining Australia's increased share of ISI publications: the effects of a funding formula based on publication counts. *Research Policy* 2003:143-155.
- Chubin DE & Hackett EJ (1990). *Peerless science: Peer review and U.S. science policy*. Albany, NY: State University of New York Press.
- Cicchetti DV (1991). The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. *Behavioral and Brain Sciences* 14: 119–135.
- Cole J & Cole S (1979). Which researcher will get the grant? *Nature* 279 (June 14):575-576.
- Cole S, Cole J & Simon G (1981). Chance and Consensus in Peer Review. *Science* 214:881-886.
- de Jong S, Barker K, Cox D, Sveinsdottir T, Van den Besselaar P (2014). Understanding societal impact through productive interactions: ICT research as a case. *Research Evaluation* 23 2: 89-102
- De Rijcke S & Rushforth A (2015). To intervene or not to intervene; is that the question? On the role of scientometrics in research evaluation. *Journal of the Association for Information Science and Technology* 66: 1954–1958.
- De Rijcke S, Wouters PF, Rushforth AD, Hammarfelt B (2016). Evaluation practices and effects of indicator use – a literature review. *Research Evaluation* 25(2):161-169
- Dora Declaration (2012). [available at <https://sfdora.org/read/>].
- Garcia JA, Rodriguez-Sánchez R & Fdez-Valdivia R (2020). Confirmatory bias in peer review. *Scientometrics* 123 (1):517–533.
- Gingras, Y (2020). "The transformation of the scientific paper: From knowledge to accounting unit", in Biagioli M & Lippman A (2020), pp. 43-56.
- Harzing AW (2018). Running the REF on a rainy Sunday afternoon: Can we exchange peer review for metrics? STI 2018 Conference Proceedings, pp 339-345.
- Heinze T (2008). How to sponsor ground-breaking research: a comparison of funding schemes. *Science and Public Policy* 35, 802–818.
- Hicks D, Wouters P, Waltman L, de Rijcke S & Rafols I (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, April 22, 520: 429-431.
- Krüger AK (2020). Quantification 2.0? Bibliometric Infrastructures in Academic Evaluation. *Politics and Governance* 8 (2): 58–67. DOI: 10.17645/pag.v8i2.2575
- Lee CJ, Sugimoto CR, Zhang G, Cronin B (2013). Bias in Peer Review. *Journal of the American Society for Information Science and Technology* 64 (1): 2–17.
- Martin B & Irvine J (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy* 12 (2): 61-90.

- Merton RK (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. (Ed. N Storer). Chicago and London: Univ Chicago Press.
- Moed HF, Burger WJM, Frankfort JG, van Raan AFJ (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy* 14 (3): 131–149.
- Moed HF (2005). *Citation analysis in research evaluation*. Dordrecht: Springer Verlag.
- Moed HF (2007). The Future of Research Evaluation Rests with an Intelligent Combination of Advanced Metrics and Transparent Peer Review. *Science and Public Policy*, 34 (8): 575–83.
- Mongeon P & Paul-Hus A (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics* 106 (1): 213–228,
- Narin F (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Computer Horizons Inc [456 pp].
- Neufeld J & von Ins M (2011). Informed peer review and uninformed bibliometrics? *Research Evaluation* 20 (1): 31–46.
- Oppenheim C (1996). Do Citations Count? Citation Indexing and the Research Assessment Exercise (RAE). *Serials* 9(2): 155–161. DOI: <http://doi.org/10.1629/09155>
- Oppenheim C (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation* 53 (5): 477–487.
- Oravec JA (2019). The "Dark Side" of Academics? Emerging Issues in the Gaming and Manipulation of Metrics in Higher Education. *Review of Higher Education* 42 (3): 859–877.
- ÖRE2010: Örebro Research Evaluation. Utvärdering av forskningen vid Örebro universitet. Örebro: Örebro universitet.
[available from https://www.oru.se/contentassets/b5a749e536864a2297ceb33d93e038ed/rapporten_ore2010_101213.pdf]
- ORU2015: *Örebro University Research Evaluation 2015: Evaluation Report*. (Eds. Masterton & Sjödin). Örebro: Örebro University.
[available from <https://www.oru.se/globalassets/oru-sv/om-universitetet/oru2015.pdf>]
- Piro FN, Rørstad K & Aksnes DW (2016). How does prolific professors influence on the citation impact of their university departments? *Scientometrics* 107:941–961.
- Rinia EJ, Van Leeuwen TN, Van Vuren HG, Van Raan AFJ (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: evaluation of condensed matter physics in the Netherlands. *Research policy* 27 (1), 95–107.
- Sandström U (2014). *Bibliometric Evaluation of SEPA-funded large research programs 2003–2013*. Report 6636, December 2014. Stockholm: Swedish Environmental Protection Agency.
[Available from <https://www.forskningspolitik.se/files/dokument/sepa-report-dec2014.pdf>]
- Shibayama S & Wang J (2020). Measuring originality in science. *Scientometrics* (2020) 122:409–427.
- Sivertsen G (2018). The Norwegian Model in Norway. *Journal of Data and Information Science* 3 (4):3–19.

- Van Arensbergen P, van der Weijden I, van den Besselaar P (2014). The selection of talent as a group process; a literature review on the dynamics of decision-making in grant panels. *Research Evaluation* 23 (4): 298–311.
- Van Bevern R, Komusiewicz C, Niedermeier R, Sorge M & Walsh T (2016). H-index manipulation by merging articles: Models, theory, and experiments. *Artificial Intelligence* 240: 19–35
- Van den Besselaar P & Leydesdorff L (2009). Past performance, peer review and project selection: a case study in the social and behavioral sciences. *Research Evaluation* 18 (4): 273–288.
- Van den Besselaar P & Sandström U (2015). Early career grants, performance and careers; a study of predictive validity in grant decisions. *Journal of Informetrics* 9: 826–838
- Van den Besselaar P & Sandström U (2019) Measuring researcher independence using bibliometric data: A proposal for a new performance indicator. *PLoS ONE* 14(3): e0202712.
- Van den Besselaar P, Heyman U & Sandström U (2017). Perverse Effects of Output-based Research Funding? Butler’s Australian Case Revisited. *Journal of Informetrics* 11: 905–918.
- van Raan, AFJ (2005). Measurement of Central Aspects of Scientific Research: Performance, Interdisciplinarity, Structure. *Measurement* 3 (1), 1–19.
- Wennerås C & Wold A (1997). Nepotism and sexism in peer-review. *Nature* 387, 341–343 (22 May 1997) | doi:10.1038/387341a0
- Wilsdon J et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. DOI: 10.13140/RG.2.1.4929.1363
- Wouters PF (2020). “The mismeasurement of quality and impact”, in Biagioli M & Lippman A (2020), pp. 67–76.
- Wu L, Wang D, & Evans JA (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566, 378–382. <https://doi.org/10.1038/s41586-019-0941-9>.

Annex1: The scores

A short report was made to translate the bibliometric data in an evaluation score. The main issue is impact: the field normalized citation score and the share of top papers in the oeuvre of a unit (or person).⁸

- The citation scores are as follows
 - 1.4 and higher: excellent
 - 1.2 and higher: very good
 - 1.0 and higher: good (1.0 is international average)
 - 0.8 and higher: moderate
 - Lower than 0.8: weakHowever, this is moderated by the share of top papers (top 5% cited papers). A lower average impact (NCSf) can still be good, if the top 5% score is high (above 5%), and a high citation impact can be downgraded if there are no top-cited papers.
- The DIVA scores are:
 - 1.3 and higher: very good
 - 1.0 and higher: good (1.0 is national average)
 - 0.8 and higher: moderate
 - Lower than 0.8: weak

⁸ In the bibliometric report, also an experimental percentile-based indicator (P-model) was used. We do not include it here, and it did not have a specific impact on the assessment procedures described here.