

Bibliometric differences – a case study in bibliometric evaluation across SSH and STEM

Poul Meier Melchiorson

Aalborg Universitetsbibliotek, Aalborg, Denmark

Received 5 July 2018
Revised 30 September 2018
Accepted 30 September 2018

Abstract

Purpose – The purpose of this paper is to acknowledge that there are bibliometric differences between Social Sciences and Humanities (SSH) vs Science, Technology, Engineering and Mathematics (STEM). It is not so that either SSH or STEM has the right way of doing research or working as a scholarly community. Accordingly, research evaluation is not done properly in one framework based on either a method from SSH or STEM. However, performing research evaluation in two separate frameworks also has disadvantages. One way of scholarly practice may be favored unintentionally in evaluations and in research profiling, which is necessary for job and grant applications.

Design/methodology/approach – In the case study, the authors propose a tool where it may be possible, on one hand, to evaluate across disciplines and on the other hand to keep the multifaceted perspective on the disciplines. Case data describe professors at an SSH and a STEM department at Aalborg University. Ten partial indicators are compiled to build a performance web – a multidimensional description – and a one-dimensional ranking of professors at the two departments. The partial indicators are selected in a way that they should cover a broad variety of scholarly practice and differences in data availability.

Findings – A tool which can be used both for a one-dimensional ranking of researchers and for a multidimensional description is described in the paper.

Research limitations/implications – Limitations of the study are that panel-based evaluation is left out and that the number of partial indicators is set to 10.

Originality/value – The paper describes a new tool that may be an inspiration for practitioners in research analytics.

Keywords Benchmarking, Scholarly communications, Research impact, Societal impact, Bibliometric model, Knowledge frameworks

Paper type Case study

Introduction

You may find at least two distinct “research areas” in the world of research: the area of Science, Technology, Engineering and Mathematics (STEM) and the area of Social Sciences and Humanities (SSH). With a catching title, Olmos-Peñuela *et al.* (2014) describe issues in evaluating and analyzing research from both areas: “Are ‘STEM from Mars and SSH from Venus?’” They question the widespread assumption that STEM research is more useful than SSH research, and they find that STEM and SSH research are useful in different ways.

Is it possible to make comparisons across research areas? Is it possible to evaluate research at a full-fledged university, where using a certain metric for evaluation may favor either the STEM or the SSH environment (Melchiorson and Thidemann, 2016)?

Several solutions have been proposed in the search for a bibliometric[1] tool, which may give a proper representation of research across disciplines. The use of citations in bibliometrics grew out of Eugene Garfield’s work on the Science Citation Index (SCI) in the 1960s and the 1970s (De Bellis, 2009, p. xx). The SCI became the Web of Science (WoS);



however, the limitations built into the database stuck. SSH differs from STEM research in ways that make it more difficult to cover with a citation database. Publications in national languages, in the form of books, at a slower pace and with a low degree of collaboration are some of the key characteristics of SSH compared to STEM (Waltman, 2016). Would a solution be to widen the coverage of the database? Google Scholar (GS) and Scopus are the important alternatives (Waltman, 2016) with a broader coverage. Studies show correlations between the three databases (e.g. Harzing and Alakangas, 2016), but can the slower pace, lesser degree of collaboration and the wide-ranging audience characterizing SSH research be compensated for by a broader coverage. Probably not (De Bellis, 2009, p. 286).

Another solution may be “to move beyond coverage” (Hammarfelt, 2016). Björn Hammarfelt finds that bibliometric research on the humanities is maturing. Hammarfelt (2016) reviews a range of methods developed “for the humanities” (p. 121). That the methods are developed for the humanities may be both an advantage and a disadvantage. Advantageous because the nature of the humanities is considered; but disadvantageous because we end up with methods developed for STEM and methods developed for the humanities (or for the SSH). One method for both sides may be the better solution. Hammarfelt mentions GS as an alternative to WoS and Scopus because it is not only different in terms of coverage of sources, but it also reaches more audiences and contains more publication types. Using GS is a step beyond coverage. Other steps – among even more mentioned by Hammarfelt – are counting publications and altmetric approaches.

Counting and weighting publications is exemplified in the well-known “Norwegian Model” (Hammarfelt, 2016). Counting publications is also a way to extricate the discussion from the coverage issues of citation databases. Traditionally, peer review and citation analysis were the tools for evaluating research (De Bellis, 2009, pp. 182-183; Bloch and Schneider, 2016).

The altmetric approach builds upon data from the social web. Altmetrics is not only trying to solve problems with traditional methods of analysis, but also to measure impact beyond the scholarly audiences (Hammarfelt, 2016). Besides metrics based on citations and publications are “outcome measures,” whereas altmetrics are “process indicators” (Moed, 2016). This means we should take care when using altmetrics in evaluation of research together with citation- and publication-based metrics. The object of measurement may be different. However, if the process of research is important for the outcome, we may include measurement of process in the evaluation.

Research is multidimensional and therefore a range of indicators is required to evaluate research, as it may be argued (Martin, 1996). Bibliometrics is, or ought to be, multidimensional as well (De Bellis, 2009, p. 209). Nicola De Bellis (2009) writes in a description of Ben Martin and John Irvine’s work: “Whatever the indicator, obviously no absolute appraisal is allowed, but only relative or comparative. Most important, none of the measures ensures conclusive evidence on the relative contribution to scientific progress of a research unit; rather, each serves the purpose of building, at best a partial indicator of scientific output or performance” (p. 211).

The aim is to use a “multidimensional research assessment matrix” (Moed and Halevi, 2015) in a specific layout to evaluate across research areas. This may be a third attempt to solving the issue of research evaluation across disciplines. First attempt is widening the coverage of the citation databases. Second, altmetrics and publication-based metrics are introduced. The combination of a range of metrics – citation- and publication-based metrics plus altmetrics – moves beyond coverage, considers the diversity of research practices and publication patterns across disciplines and, finally, we can combine methods developed for both sides in one evaluative scheme. In the multidimensional matrix, we use citation-based metrics for STEM areas as well as metrics developed for the SSH. It is important to stress that citation-based metrics and altmetrics not necessarily evaluate the same aspects of research. There is no evidence yet for what altmetric scores denote – research quality or societal impact? This means we must be careful not to use altmetrics in research evaluation without considering the quality

of the research evaluated. As Bornmann and Haunschild (2018) point out, we might end up appraising bad research, just because it received much attention in the social media.

In this paper, we ask if evaluating a STEM against an SSH researcher is possible and defensible. In a case study, we compare a STEM and an SSH researcher in two evaluative frames: keeping the multidimensionality; and collapsing the multidimensionality into one score. For a clearer answer, we will imagine a researcher and a research management perspective on the two frames. What would a researcher and a research manager gain from either of the two frames?

Imagine two spaces: one space is an equal-sided cube (10 cm × 10 cm × 10 cm) and another space is long and thin (1 cm × 1 cm × 1,000 cm). The two spaces fill out the same volume: 1,000 cm³. Can we apply this analogy to research assessment? Can we find a dimension that the two units of assessment share? Alternatively, are we better off not collapsing the diversity into one? The question Olmos-Peñuela *et al.* ask in these terms: is an equal-sided cube more useful than a long and thin tube? Our question in these terms: how do we perform the best comparison between the two spaces – by keeping the three dimensions or by collapsing the three length values into a measure for volume? At best, the results will be exploratory, but, hopefully, we can find a guideline for when to use the expanded description and when to use the collapsed one.

Method

Why make this bibliometric comparison between researchers? P. Zunde (1971) put forward three possible purposes of bibliometrics. More purposes can be found (Nicolaisen, 2007), and we have at least four: evaluation of research, modeling of the historical development of science, information search and retrieval and knowledge organization based on bibliographic coupling and co-citation analysis. We propose a fifth, visibility: visibility in the research databases (Aksnes and Rip, 2009; Wildgaard, 2015) and visibility as a prerequisite for accountability. Transparency and accountability as required in contemporary management and maybe reasons why bibliometrics is playing a role in the academic community (Mingers and Leydesdorff, 2015; Moed and Halevi, 2015).

Three models for research evaluation can be pinned out (Bloch and Schneider, 2016): publication-based, citation-based and panel-based. Panel-based evaluation has the great advantage – it is quantitative – but it also has disadvantages – for instance, biases in some cases (De Bellis 209, p. 182). The greatest disadvantage in our context, however, is that it is resource consuming.

When it comes down to doing the job, we are stretched out between two real conditions: the quest for multidimensionality and quality, and, on the other hand, resources. Because resources are limited, we must be careful as the scope of our analysis is apprehended and accepted. Another reason is of course that the results may be misinterpreted, if stipulations are not laid out from the beginning. However, in this work, we will have to make delimitations regarding multidimensionality and quality. Using protocols for data collection, both bibliographic (GS) and bibliometric databases and not too simple indicators we try to rise above “desktop bibliometrics” (Moed *et al.*, 2010; Moed and Halevi, 2015).

An important, leading principle in choosing indicators is multidimensionality. We want to cover a range of performance categories. Citation- and publication-based measures traditionally aim at covering scholarly impact – impact among peers (Moed and Halevi, 2015; Moed, 2016), but as Moed and Halevi (2015) point out, more types of impact are included in these years’ assessment exercises. Societal or cultural impact requires other indicators. It may be altmetrics, considering non-publications (data files, videos, presentation, event, performance, interviews, etc.), or considering publications that do not appear in citation databases (scholarly monographs, textbooks, commissioned research reports, newspaper articles, etc.). Despite warnings about descending into a “new age of numerology,” (Bawden and Robinson, 2016) uses are made of collapsed scores, where several indicators are combined in one number because it is useful[2].

Depending on whether we regard the research process as an important part of evaluation of research performance, we should include process measures in our multidimensional picture (Moed, 2016). Numbers of publications and citations can be considered output measures, while mass media coverage, social media activities and usage statistics (downloads, etc.) describe some aspects of the research process.

We choose a radar chart as the expression of the multidimensional picture and a table that collapses into a single number to express the direct comparable. Other approaches are described in the literature. Bornmann *et al.* (2012) visualize percentile scores in a very efficient way by using “the box and whisker chart” (p. 253). Bornmann and Marx (2014) also suggest beam plots to illustrate distributions of percentiles for publication impact of a single researcher. Bornmann and Haunschild (2018) have a very interesting example of using beam and scatter plots when comparing paper impact percentiles with journal impact percentiles. We investigate the use of a multidimensional picture vs the collapse of diversity in one number. To bring matters to a head, we avoid the more detailed visualizations and use quartile numbers instead. Although it is clear that a single number must lead to loss of information. So why investigate the use of single numbers – because it is so widely used among researchers and in research management (Bornmann and Haunschild, 2018).

In the ever-growing list of indicators, many analyses address unintended consequences of their uses and some address ethical issues in evaluating research and researchers (Hicks *et al.*, 2015). Yves Gingras investigates the actual indicators themselves and stresses the need for a clearer understanding of the indicators: what are sound indicators? Gingras (2014) lists three criteria for valid indicators:

- (1) Fit for purpose: do citations measure “quality” and “impact,” or maybe they indicate “visibility”? If we want to keep the concept of “quality,” we need to establish a connection between citations and an independent concept of “quality.” Gingras finds that “great science” is always highly cited. Not that a specific number can be assigned, but great science is always way beyond average, when citations are counted. De Bellis (2009) questions the assumption that citations can be a proxy for quality (p. 264).
- (2) Sensitivity to intrinsic inertia of the object measured: the period for evaluation must fit the speed with which the measured object is changing – intrinsic inertia. A university may change over a decade while the number of publications for a researcher may change every year. A dashboard with publication production per day would make no sense.
- (3) Homogeneity: Gingras argues that heterogeneous indicators are invalid. An example is the h-index. It is an indicator proposed by Jorge Hirsch in 2005 (De Bellis, 2009, p. 201). This indicator is invalid, because two dimensions are combined in one number. Numbers of papers are combined with numbers of citations: If a researcher has an h-index of n , it means that she has published n papers each of which has been cited at least n times. De Bellis (2009) agrees that the h-index has weaknesses, which can be easily identified (p. 203). He is a bit more reluctant to disprove the practical use of the h-index (De Bellis, 2009, p. 202). “The above shortcomings are currently keeping scientometricians busy in the search for corrections, supplements, complements, or alternative to the original formulation. On the whole, although giving the impression that it is just a matter of time until someone new wakes up in the morning inspired by a fresh variant or an ad hoc corrective factor, the latest profusion of such creative efforts reinforces the conviction that Hirsch’s proposal incorporates some of the most desirable and long-overdue properties of an indicator of research performance” (De Bellis, 2009, pp. 203-204). Instead of totally rejecting, De Bellis (2009) stressed the need for using more than one number in bibliometric

evaluations (p. 2006), and, as Ludo Waltman (2016) points out, these h-indexes play an important role in evaluation of research nowadays. This is not the same as to say that this indicator is without problems or sometimes misused. The h-index describes what happens to the publication after it is published. We will include another perspective on the work of the researcher with focus on the time of publishing. We include the percentage of the researcher's publications, which are in the top 10 percent journals according to the metric CiteScore[3]. CiteScore is an inhomogeneous, journal metric – citation-based. By including this metric, we let the researcher's ability to publish in prestigious journals have a place in the overall picture. Furthermore, studies show that too much weight on h-index variants may be disturbing because the variants only add redundant information (Bornmann *et al.*, 2011). We leave the Scopus h-index out the investigation for that reason, but we use the GS and WoS h-indexes. The reason for using these two is that the GS index may help in fields with lower degrees of coverage in WoS (Prins *et al.*, 2016), which, for many years, was the data source for bibliometrics.

Gingras' (2014) criterion of homogeneity is argued for by this example: "Combining different indicators into a single number is like transforming a multidimensional space into a zero-dimension point, thus losing nearly all the information contained on the different axes" (p. 116). Recall our example with three physical dimensions collapsed into volume. We do not find that volume is an invalid representation – it just represents something else than the three numbers indicating height, width and depth. Another example is thermal energy. This is mass multiplied with temperature and a substance specific constant. Very different dimensions combined into something meaningful.

We will need a definition of "performing better than" (Gingras, 2014). We saw that we might not find an absolute yardstick – at best a relative (De Bellis, 2009, p. 211). When ranking researchers, we hit one of the cornerstones in bibliometric descriptions: Bibliometric distributions are skewed (De Bellis, 2009, p. 209). Many authors have a few citations and low impact, but they take part in "everyday, normal research." A few have many citations and high impact – maybe they do ground-breaking work. There is no normal distribution, where researchers are spread out on both sides of the average. The median is a better baseline. The use of the median together with percentiles may further overcome the skewness in bibliometric distributions (Bornmann *et al.*, 2012). We will use quartiles in our analysis. Then another perspective on performance is that researchers may be "good in different ways." They may reach out to different impact dimensions (Moed and Halevi, 2015).

In the field of bibliometrics, it is often mentioned that we should compare "like with like" – apples with oranges (De Bellis, 2009, p. 199; Leydesdorff and Bornmann, 2016). We cannot directly compare a STEM researcher with an SSH researcher because of the differences of their fields. If we would make such a direct comparison of citations, for instance, we would not be comparing the impact of the two researchers but the citation behavior of the two fields. Some fields traditionally have long reference lists and cite each other in high degree – others have a short reference list and fewer citations. A young researcher has not had the same time to publish and collect citations as an older colleague (De Bellis, 2009, p. 203). Some fields traditionally publish books, which are not yet sufficiently included in citation databases. Many diversities between fields may be mentioned.

For that reason, the normalization of citation scores using reference sets based on fields in the citation databases have become common practice in evaluative bibliometrics (Leydesdorff and Bornmann, 2016). Loet Leydesdorff and Lutz Bornmann have an ambition to find "best possible practices" because the common practice of using "built in" fields or subject categories in WoS is questionable. These categories were not built for bibliometrics and either too broad or too narrow[4]. Fields are attached to journals and, even within a journal, not all articles may relate well to one single field.

When the delimitation of research areas is accomplished or the delimitation of an “appropriate reference set” is found, some prerequisites for normalization are in place (Zhou and Leydesdorff, 2011). Normalization is the process of making the scientometric data ready for evaluation. Especially when we make cross-disciplinary comparisons, normalization is necessary.

Many of the indicators we have mentioned use standard or full counts (De Bellis, 2009, pp. 272-273). When we compare across disciplines, we may experience that in one area publications have many authors, while in others one author is not unusual.

Partial indicators

We are comparing two researchers which we assume have the same “research age” and productivity, so we will not consider size normalization. This means delimitation of and normalization according to their research areas are main points. We will use a “built in” field normalized indicator – with the implications noted! Using SciVal – which is an analytic tool for Scopus data – we have access to the Field-Weighted Citation Impact (FWCI)[5]. The FWCI considers the differences in research behavior across disciplines. A comparison is set up between a publication set and the average number of citations received by similar publications in Scopus. $FWCI = 1$ is the average and, for example, a score of 2.14 means that the publications in the set have received 114 percent more citations than expected.

It may seem fair to adjust the counting, so that the credit is shared between authors. We include fractional counting of BFI-points for this reason. BFI-points are fractionalized between universities, but we extend the fractionalization fully, so that authors from the same university must share the BFI-points.

Indicators we end up using in this exemplary evaluation:

- (1) Publication-based:
 - number and types of publications (VBN, WoS, Scopus and GS);
 - Danish Bibliometric Indicator (BFI):
 - externally fractionalized; and
 - fully fractionalized.
- (2) Citation-based:
 - H-index (WoS and GS);
 - Top 10 percent Journal Percentile (CiteScore); and
 - FWCI.
- (3) Altmetrics:
 - numbers of downloads;
 - numbers of news media coverage; and
 - funding in DKK.

To sum up indicator selection, we use publication- and citation-based evaluation. We assume that size normalization is not necessary and impose a built-in field normalization. Fractionalization is introduced on one of the indicators. Both output and process measured are included as well as indicators of both scholarly and societal impact.

Data sources

To widen coverage, we choose to use both WoS[6], Scopus and GS as the source of citation data. The BFI provides us with publication-based measures, and the local CRIS system at

Aalborg University (VBN) as well as the finance department at Aalborg University are sources of altmetric data – numbers of download, press cuttings and funding data.

The BFI was introduced in Denmark 2008 with inspiration from the Norwegian model (Ingwersen and Larsen, 2014). Central to the BFI are the “authority lists.” For a publication to be awarded BFI-points, the journal or publisher must be on one of these lists, which are made up of recommendations from 67 expert panels. The model has four levels: publications published outside the lists and three levels on the lists 1–3, with Level 3 being the most prestigious[7].

Protocols for data collection

Our starting point is a STEM researcher and an SSH researcher: two persons with a certain “research age.” They are both professors, and they both have a ResearcherID, a GS profile and they both appear in Scopus. We will investigate both in WoS, Scopus and GS regarding publications and citations. As a reference for publications and publication types and other content types, we will use the two persons’ registrations in Aalborg University’s research information system Pure – referred to as VBN[8].

X has been employed at a STEM department (Department X) at Aalborg University with small break offs since 1995 – and as a professor since 2006. A has been employed at an SSH department (Department A) at Aalborg University since 2009, where he became a professor. To rank the two researchers, we use data for all professors affiliated with the two departments X and A in 2017.

Time frame for our analysis is generally unlimited – that is, the period where the professors have data is included. This is quite unusual, and the main reason for this is that citation data for the SSH professors would be too small in numbers to analyze, if, for instance, 2012–2016 was chosen. For BFI data, the time frame is 2012–2016 for practical reasons. Funding is related to projects approximately in a five-year period[9].

When searching persons in WoS, the following recipe is used: AU = ([last name] [first initial] ×) → select relevant “author groups” from the first names and affiliations. In a few cases, this was not possible, and the search had to be a combination of name and affiliation.

Search in Scopus is done as follows: “author search” → [last name] [first name with a period after] → one or more “author groups” are selected – relevance decided from first name, subject area and affiliation.

Searches in GS are done with “Publish or Perish,” which is an analytical tool for GS[10]. First, we try if the person has a GS profile – if not, we make a regular GS search[11]. In a few cases, we exclude journal names because of more authors with the same name.

Analysis – step by step

To decide the performance upon for X and A, we will follow this line of action:

- (1) The rank of X and A will be determined for each partial indicator. The ranking will shape the basis for finding a quartile for the researcher. This is done in Excel by sorting the indicator result for all professors at the department and then using the function QUARTILE.INC to find the quartile number. When we use the quartiles, we distribute around the median[12]. For example, the 13 professors at Department X have a BFI-score. We sort the scores descending and use QUARTILE.INC on the sorted scores. From this, we can see that X is in quartile 4, which is the highest.
- (2) We will look for convergence using a table with the quartile numbers:
 - which percentiles are the two researchers placed within at the different rankings; and
 - if placings correspond, we will say that there is convergence.
- (3) We will calculate a “q-index” (q for quartile), which is simply the sum of quartile numbers. The maximum possible value of the q-index is the number of indicators × 4.

With the q-index, we have a number we can use for the one-dimensional comparison by numbers. For the multidimensional comparison, we will create a “performance web.” This is a radar chart created in Excel. This radar chart shows A’s performance with X as index level. Both A’s and X’s values have been divided with X’s values. That is, if X has 262 publications, and A has 461, then X has 100 percent and A has 176 percent.

To sum up the evaluation of performance, we use the ranking of researchers as a relative yardstick. Quartiles are chosen because of skewness of distributions. And because researchers may relate to different impact dimensions – scholarly and societal – we use more numbers to evaluate the performance.

In the last step of our analysis, we will investigate if we can find two of the purposes for bibliometrics fulfilled: evaluation and visibility. Which of our analyses with either a multidimensional or a one-dimensional result serves the purposes of evaluation and visibility best?

Results

Researcher X publishes mainly conference articles in proceedings and journals in English. Over the years, there is an ascending tendency, increasing after being appointed as professor and then going down again. The average number of authors on publications is 4.2[13].

Researcher A writes mainly in Danish, a great deal in English and with a greater variation in publication types than we find with X. Most used types are journal articles, book chapters, books and reviews/comments/debate in newspapers. Through the years, A shows the same publication pattern regarding number of publications as X. The number increases up till and in the years after the professor appointment and then it goes down again. The average number of authors on publications is 1.3.

Coverage

Coverage is calculated as the mean of the share of publications in the citation databases compared to VBN. Results are presented in Table I. We see that our results are in line with the concerns put forward earlier regarding the difference in coverage for STEM and SSH. We see that GS has a coverage above 100 percent. This may be an indication that more material is covered in GS but, more likely, it may reveal duplicates and noise.

Ranking

Quartiles are found for our ten indicators (Table II). Red color indicates a lower quartile than the other researcher in the comparison. It is seen that A has a slightly higher q-index than X, but they are both close to the maximum of 40. Intervals in parentheses designate minimum and maximum for the partial indicators in the table[14] – for instance, professors at Department X have h-indexes in WoS ranging from 4 to 24.

Performance web[15]

From the radar chart, you get more dimensions to the comparison between X and A. In the first chart, X is evaluated against A and, in the next, it is the other way around.

Coverage	Scopus (%)	WoS (%)	GS (%)
Department X	66	53	157
Department A	21	13	126

Table I.
Coverage – citation databases

Table II.
Ranking – quartile
distribution

Quartiles		X		A
No. of publications VBN	Q3	(57–482)	Q4	(57–461)
Top journals (%)	Q3	(15–67%)	Q3	(0–50%)
H index WoS	Q4	(4–24)	Q4	(0–22)
H index GS	Q4	(11–53)	Q4	(4–64)
BFI-points	Q4	(12–189)	Q4	(3–131)
Fractionalized BFI-points	Q4	(5–63)	Q4	(3–123)
No. of downloads	Q4	(4,280–90,078)	Q3	(0–29,735)
No. of media coverage	Q3	(0–201)	Q4	(1–685)
Funding	Q2		Q4	
FWCI	Q4	(0.1–4.0)	Q3	(0.3–2.3)
“Q index”	35		37	

The performance web is quantitative in its origin – the data behind are ten numbers for each web. To read the web, you may benefit from considering the qualitative image it presents. For instance, these statements may be considered valid readings: “Researcher X is more pronounced in the right side, where citation indexes are presented” or “Researcher A is partly represented in the left side where altmetric data are presented” or “Researcher X has to share his BFI-points with co-authors in a higher degree than Researcher A” (Figures 1 and 2).

Discussion

How to use the q-index and the performance web, respectively? We have mentioned two of the purposes which may be up front for bibliometric analysis: making research visible and evaluating research. By evaluating research, we often mean that we judge it. That is also what should be the purpose with the q-index. Are one or the other of X and A performing better? But as Spaapen and van Drooge (2011) point out evaluating can also be about learning. Hopefully, we can learn from the performance web to identify strengths and weaknesses. Will you tend to have a weak left side of the chart if you are strong on the right side and vice versa?

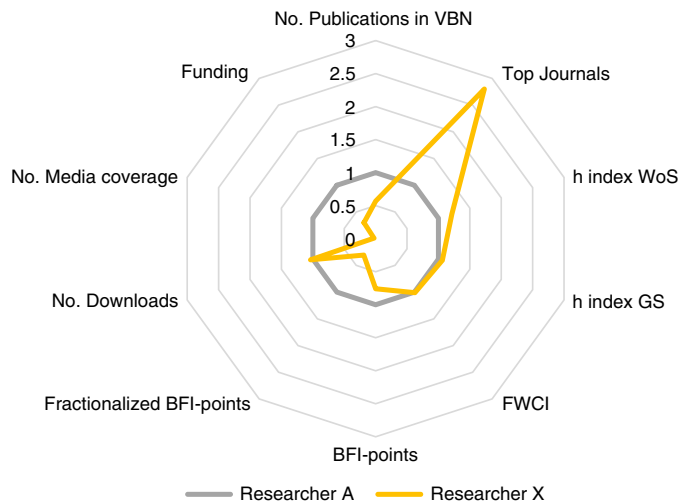


Figure 1.
Performance
web – researcher X

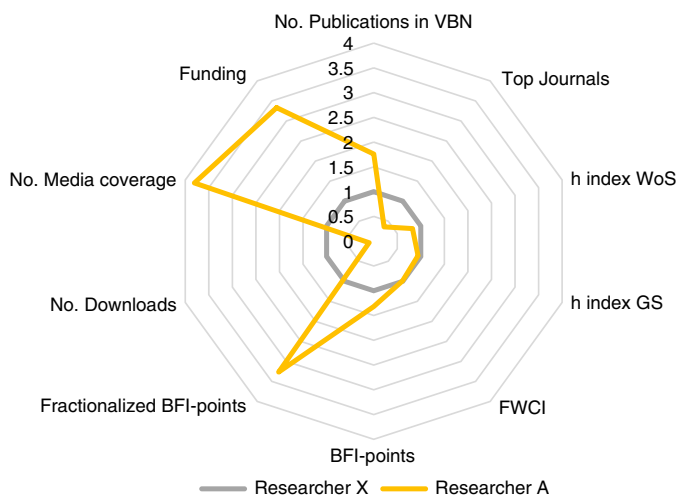


Figure 2.
Performance
web – researcher A

From a researcher perspective, the one-dimensional score may be more visible than many explanations, but, on the other hand, the score may hide important punchlines, for instance, that download statistics testify to great usage of a poorly cited publication. On the other hand, the performance web may provide learning and perspectives: a researcher does well in citations but poor in media coverage – is that something he/she should consider changing?

The research manager will have an efficient tool to rank research in the one-dimensional framework. However, as we have seen “truths” may be hidden in one number, so great care should be taken when evaluating researchers with this tool. The performance web is hard to use in ranking, but insight in the “DNA” of a research field may be provided.

The q-index is a number and can therefore make research visible in a world of research evaluation where numbers count. However, it is just one number where a download and a book have the same weight. Weighing the partial indicators making up the q-index may be a proposal for further development of the q-index. Another improvement of the q-index would be to consider the dimensions of the index. How many dimensions are needed to describe a researcher? As we have seen – a very important foundation for research evaluation is left out of our study – it is the qualitative description made by a panel for instance. On the qualitative side – how many dimensions? We have ten dimensions in this case study. A comprehensive evaluation of researchers would require a more detailed analysis. Moed (2017) lists 28 indicators in a recent study. Lorna Wildgaard (2015) finds 87 citation-based applied at the individual level.

Other limitations of our approach should be mentioned: Both the radar charts and the q-index show no development over time. It will not be possible for users of the bibliometric report to see if the evaluated researchers had their impact recently or many years ago. A way to improve the report can be to include publications and impact only three or five years back in time. In addition, the radar chart has a disadvantage in that it is difficult to read, when more than two researchers are compared. Thinking in “researcher archetypes” could be a way to improve the radar chart. If the radar chart is divided in two or more zones, each zone can denote an archetype. Archetypes examples can be the researcher with much societal influence and the researcher with much influence among peers.

What is more valuable: a researcher with many articles for download and no downloads, or a researcher with one article and a lot of downloads? What is more valuable a tube or a box? The q-index and the performance web are answers in very different ways to this question.

Notes

1. Without going too much in detail, we will use the term “bibliometrics” to cover a wide variety of metrologies of research: scientometrics, informetrics, webometrics, netometrics and cybermetrics (De Bellis, 2009, pp. 2-5).
2. An argument put forward by the company Altmetric: available at: <https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated>. The newly announced platform Dimensions also use a “collapsed” score – however, in a slightly different way, by having citations as the front figure/score: available at: www.altmetric.com/blog/dimensions-badges-a-new-way-to-see-citations/ (accessed February 8, 2018)
3. www.elsevier.com/authors-update/story/impact-metrics/citescore-a-new-metric-to-help-you-choose-the-right-journal (accessed September 27, 2018).
4. SciVal – an analytic tool for Scopus data – has recently released “Topics of Prominence.” A high degree of granularity should be one of the advantages – 97.000 topics should make comparisons possible: available at: www.elsevier.com/connect/the-dawn-of-predictive-analytics-to-measure-research-performance-scivals-topic-prominence (accessed January 10, 2018).
5. www.elsevier.com/___data/assets/pdf_file/0020/53327/scival-metrics-guidebook-v1_01-february2014.pdf (accessed January 11, 2018).
6. The Core Collection.
7. <https://ufm.dk/en/research-and-innovation/statistics-and-analyses/bibliometric-research-indicator> (accessed January 11, 2018). Level 3 is introduced in 2017, and our data do not contain Level 3 publications.
8. <https://vbn.aau.dk>. Pure is a CRIS system and an Elsevier product.
9. The periods for department X and department A are not the same. Department X lists project portfolio 2016–2017 (projects may have started earlier). Department A lists projects with a start-date five year from now and backwards.
10. <https://harzing.com/resources/publish-or-perish> (accessed January 11, 2018).
11. If there is a profile, data quality may be better because the person can access the profile and clean data. The percentage of professors with a GS profile is as follows: Department X 69 percent; Department A 39 percent.
12. <https://support.office.com/en-us/article/QUARTILE-function-93cf8f62-60cd-4fdb-8a92-8451041e1a2a> (accessed January 11, 2018).
13. The average number of authors is found by counting authors on associated publications in VBN.
14. Funding is without details for confidentiality reasons.
15. The no. of downloads have been divided by 10 in order not to have the chart axis distorted. No. of media coverage for Researcher A has been divided by 10 in order not do have the chart axis distorted.

References

- Aksnes, D.W. and Rip, A. (2009), “Researchers’ perceptions of citations”, *Research Policy*, Vol. 38 No. 6, doi: 10.1016/j.respol.2009.02.001.
- Bawden, D. and Robinson, L. (2016), “Information’s magic numbers: the numerology of information science”, in Sugimoto, C. (Ed.), *Theories of Information and Scholarly Communication*, ISBN 9783110298031, de Gruyter, Berlin, pp. 180-196.
- Bloch, C. and Schneider, J.W. (2016), “Performance-based funding models and researcher behavior: an analysis of the influence of the Norwegian publication indicator at the individual level”, *Research Evaluation*, Vol. 25 No. 4, pp. 371-382, doi: 10.1093/reseval/rvv047.

- Bornmann, L. and Haunschild, R. (2018), "Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data", in Dorta-González, P. (Ed.), *PLoS ONE*, Vol. 13 No. 5, p. e0197133, doi: 10.1371/journal.pone.0197133.
- Bornmann, L. and Marx, W. (2014), "Distributions instead of single numbers: percentiles and beam plots for the assessment of single researchers", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 1, pp. 206-208, doi: 10.1002/asi.22996.
- Bornmann, L., Mutz, R., Hug, S.E. and Daniel, H.D. (2011), "A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants", *Journal of Informetrics*, Vol. 5 No. 3, pp. 346-359, doi: 10.1016/j.joi.2011.01.006.
- Bornmann, L., Bowman, B.F., Bauer, J., Marx, W., Schier, H. and Palzenberger, M. (2012), "Standards für die Anwendung der Bibliometrie bei der Evaluation von Forschungsinstituten im Bereich der Naturwissenschaften", *Zeitschrift für Evaluation*, Vol. 11 No. 2, pp. 233-260.
- De Bellis, N. (2009), *Bibliometrics and Citation Analysis – From the Science Citation Index to Cybermetrics*, The Scarecrow Press, Lanham, p. 417, ISBN: 9780810867130.
- Gingras, Y. (2014), "Criteria for evaluating indicators", in Cronin, Blaise and Sugimoto, Cassidy R. (Eds), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Intent*, ISBN: 9780262525510, MIT Press, Cambridge, MA, p. 466.
- Hammarfelt, B. (2016), "Beyond coverage: toward a bibliometrics for the humanities", in Ochsner, M., Hug, S. and Daniel, H.D. (Eds), *Research Assessment in the Humanities*, Springer, Cham, pp. 115-131, doi: 10.1007/978-3-319-29016-4_10.
- Harzing, A.W. and Alakangas, S. (2016), "Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison", *Scientometrics*, Vol. 106 No. 2, pp. 787-804, doi: 10.1007/s11192-015-1798-9.
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. and Rafols, I. (2015), "The Leiden Manifesto for research metrics", *Nature*, Vol. 520 No. 7548, pp. 429-431, available at: www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351
- Ingwersen, P. and Larsen, B. (2014), "Influence of a performance indicator on Danish research production and citation impact 2000–12", *Scientometrics*, Vol. 101 No. 2, pp. 1325-1344, doi: 10.1007/s11192-014-1291-x.
- Leydesdorff, L. and Bornmann, L. (2016), "The operationalization of 'fields' as WoS subject categories (WCs) in evaluative bibliometrics: the cases of 'library and information science' and 'science & technology studies'", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 3, pp. 707-714, doi: 10.1002/asi.23408.
- Martin, B.R. (1996), "The use of multiple indicators in the assessment of basic research", *Scientometrics*, Vol. 36 No. 3, pp. 343-362, doi: 10.1007/BF02129599.
- Melchiorson, P. and Thidemann, N. (2016), "Full-fledged research analysis at Aalborg University", Poster at 21st Nordic Workshop on Bibliometrics and Research Policy, Copenhagen.
- Mingers, J. and Leydesdorff, L. (2015), "A review of theory and practice in scientometrics", *European Journal of Operational Research*, Vol. 246 No. 1, pp. 1-19, doi: 10.1016/j.ejor.2015.04.002.
- Moed, H.F. (2016), "Altmetrics as traces of the computerization of the research process", in Sugimoto, C.R. (Ed.), *Theories of Informetrics and Scholarly Communication (A Festschrift in Honour of Blaise Cronin)*, pp. 360-371, ISBN 9783110298031, Walter de Gruyter, Berlin/Boston.
- Moed, H.F. (2017), *Applied Evaluative Informetrics*, Springer, Cham, ISBN 978-3-319-60522-7, doi: 10.1007/978-3-319-60522-7.
- Moed, H.F. and Halevi, G. (2015), "Multidimensional assessment of scholarly research impact", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 10, pp. 1988-2002, doi: 10.1002/asi.23314.
- Moed, H.F., Linmans, J.A.M., Nederhof, A.J., Zuccala, A., López Illescas, C. and de Moya Anegón, F. (2010), "Options for a comprehensive database of research outputs in Social Sciences and Humanities", (Annex 2 of the report "Towards a bibliometric database for the social sciences and humanities-a European scoping project"), Science and Technology Policy Research Unit, Sussex.

-
- Nicolaisen, J. (2007), "Citation analysis", *Annual Review of Information Science and Technology*, Vol. 41 No. 1, pp. 609-641, doi: 10.1002/aris.2007.1440410120.
- Olmos-Peñuela, J., Bennenworth, P. and Castro-Martínez, E. (2014), "Are 'STEM from Mars and SSH from Venus'? – challenging disciplinary stereotypes of research's social value", *Science and Public Policy*, Vol. 41 No. 3, pp. 384-400, doi: 10.1093/scipol/sct071.
- Prins, A.A.M., Costas, R., Leeuwen, T.N. and Wouters, P.F. (2016), "Using Google Scholar in research evaluation of humanities and social science programs: a comparison with Web of Science data", *Research Evaluation*, Vol. 25 No. 3, pp. 264-270, doi: 10.1093/reseval/rvv049.
- Spaapen, J. and van Drooge, L. (2011), "Introducing 'productive interactions' in social impact assessment", *Research Evaluation*, Vol. 20 No. 3, pp. 211-218, doi: 10.3152/095820211X12941371876742.
- Waltman, L. (2016), "A review of the literature on citation impact indicators", *Journal of Informetrics*, Vol. 10 No. 2, pp. 365-291, doi: 10.1016/j.joi.2016.02.007.
- Wildgaard, L.E. (2015), *Measure Up!: The Extent Author-Level Bibliometric Indicators are Appropriate Measures of Individual Researcher Performance*, Københavns Universitet, Det Humanistiske Fakultet, Copenhagen.
- Zhou, P. and Leydesdorff, L. (2011), "Fractional counting of citations in research evaluation: a cross- and interdisciplinary assessment of the Tsinghua University in Beijing", *Journal of Informetrics*, Vol. 5 No. 3, pp. 360-368, doi: 10.1016/j.joi.2011.01.010.
- Zunde, P. (1971), "Structural models of complex information sources", *Information Storage and Retrieval*, Vol. 7 No. 1, pp. 1-18, doi: 10.1016/0020-0271(71)90023-4.

Corresponding author

Poul Meier Melchiorson can be contacted at: pmm@aub.aau.dk