Bibliometric indicators as items

Peter van den Besselaar[1] and Charlie Mom[2]

[1] *peter@vandenbesselaar.net*
0000-0002-8304-8565
Department of Organization Sciences - Vrije Universiteit Amsterdam (Netherlands)

[2]*charlie@teresamom.com*
0000-0003-1734-1963
TMC Research, Amsterdam (Netherlands)

Abstract
The question of what bibliometric indicator indicate has been discussed for several decades. Over that period, the use of indicators has increased, the number of indicators too, but the question of what the indicators exactly measure remains to be debated. In this paper we propose to approach it from the perspective of *scale construction*. Basically, this means that we interpret the publication-based and citation-based indicators as items that measure aspects of the scientific quality, but at the same time we accept that all these indicators are characterized by error. However, several indicators together, may lead to a valid and reliable variable, representing a latent quality dimension. This approach should not be confused with composite indicators, such as deployed in university rankings.

1. Studying bias
When studying gender differences in career decisions and grant decisions, one needs valid and reliable performance measures, in order to distinguish between merit-based gender differences and non-merit-based gender differences. The latter can be classified as gender bias, but the former not, as long as one accepts that science should be a merit-driven system. Previous work (Cruz-Castro & Sanz Menendez 2019; van den Besselaar et al. 2020) showed that it is necessary to conceptually and empirically differentiate between *gender differences* and *gender bias* as outcomes of processes. If big differences between males and females are found in the outcome data, we should not conclude or assume that there is gender bias. We always need to account for *competing explanations* or, if we are constructing empirical models, control for other relevant factors, most importantly those representing *merit*. At the minimum, if we model gender differences we consider *gender bias* as the residual effect, after controlling for other factors, like merit, preferences, reputation, etc. This in order to replace *naïve residualism* (J.R Cole, 1979) with at least *sophisticated residualism* (Cole & Fiorentine 1991). However, the issue then is how to measure merit, and using the standard bibliometric indicators for that remains problematic: the problems of validity and reliability are not satisfactory answered.

2. The indicator problem
Although bibliometric indicators are used within the science system, the discussion about the meaning and use of the indicators has intensified. We do not aim to review the literature here. Overviews are available, like Glänzel et al 2019, and recently also studies have been published that report about the opinions of researchers about the use of indicators (Cruz-Castro & Sanz-Menendez 2021), and that report about the use in practice (Van den Besselaar & Sandström 2020). Well known reports (Wilsdon et al. 2015) and statements (Hicks et al. 2015) about ‘proper use’ have attracted quite some attention, and a long discussion exists on the possible negative (perverse) effects of indicator use (Butler 2003) for which the evidence is questionable (Van den Besselaar et al. 2017).

Within the fields of science studies and research evaluation the view is dominant that indicators are at best supportive and that human decision making (and in science by experts) should be central in selection processes, there is quite some literature in selection psychology showing that 'algorithmic' selection is generally better than 'holistic' selection in terms of predicting human performance (e.g. Neumann et al 2023). This is to some extent supported by the (also in science) well known fact that future performance is best predicted by past performance.
Nevertheless, the existing set of indicators is not ideal at all. Many bibliometric indicators are available (Wilgaard et al. 2014), all directly measuring some property of publications (e.g. how often cited) and of authors (e.g. how many publications), and if applied correctly always considering the context of the scientific domains and countries. Although one can argue that these indicators may reflect some underlying quality dimension, there are no good reasons to expect that the individual indicators are a reliable measure for that dimension, as there may be much error – noise - in the data, even when there is no bias (Kahneman et al. 2021).

We propose to interpret bibliometric indicators as items, and that one needs several items to construct reliable measures. This can be done using a Principal Axis Factoring of a set of individual indicators to find out what variables measure the same underlying quality dimension, and a scale analysis to find out whether the derived components (factors) create reliable scales. This by the way, should not be restricted to bibliometric indicators only, as many other merit indicators could be included. And not all reputational dimensions can be properly measured with bibliometric indicators only (Espeland & Sauder 2016).
These indicators are intended to measure aspects of scientific quality, a concept that actually covers a variety of dimensions.

We would argue that only some of those dimensions are to some extent covered by the indicator toolbox, and even those not always correctly. We do not aim to develop the concept of quality in its various dimensions here (contributing new knowledge; independence[1]; leadership of research teams; creating societal impact; supervising PhD students; teaching; etc.), nor do we aim at developing ways of measuring all these quality dimensions. The aim of the paper is methodologically, and we propose a somewhat different approach to the use of indicators. We do so by starting from the commonly used indicators, and address two in our view important aspects of the discussion about indicators, that is the *validity* and the *reliability* of the indicators. The first point relates to the question what the indicators do measure, which is especially important when discussing indicators like the Journal Impact Factor, or the H-index: The first is only indirectly related to the own performance, and the second in not field normalized, and not restricted in terms of the period covered. The second point relates to how we should measure the various quality dimensions, as there are quite some alternative indicators around. We propose to focus not on the individual indicators but on the underlying dimensions. This approach is common in test psychology (Drenth & Sijtsma 2006).

3. Measurement

Many bibliometric indicators have been developed, although most of them all measure the same small number of phenomena: productivity, impact, and (international) collaboration (co-authoring). Wilgaard et al. (2014) already distinguished 108 different variants, and since the number has grown even more. However, there is a lack of indicators for many relevant merit dimensions such as received awards, and researcher's independence. The literature comparing

[1] But we do develop a new indicator for another understudied quality dimension: The researcher's independence (section 2.3 and Annexes 5 and 6).

the various indicators and arguing which ones are better than others, is also abundant but not at all conclusive.
We would argue that most indicators measure an underlying dimension with error but cannot be conceived as a direct measurement. Therefore, we suggest treating bibliometric indicators as items measuring an underlying (latent) performance or reputation dimension, and that we should use traditional tools to assess the items and the scale they may form. This has to be distinguished from composite indicators, such as in most university rankings, where different valuation dimensions are combined into one score, without any theoretical or methodological argument. Some rankings (like the THE-ranking) present individual rankings for various evaluation dimensions, but also an overall score which is a weighted average and this can be very sensitive for the selected weights. Furthermore, as for each dimension a single indicator has been selected, one does not know how sensitive the rankings are for the measurement error. In contrast, the Leiden Ranking does not present an overall ranking and shows only the rankings per indicator. This is as such an advantage, but also here the use of single indicators for the evaluation dimensions remains a problem.
Within the bibliometrics community, Glänzel already argued a long time ago that composite indicators are not a good idea, and therefore he argued that we better stick to the single bibliometric indicators (Glänzel & Debackere 2009). However, we would argue that this are not the only two options, as the important thing would be to investigate what is measured by bibliometric (always with some error), and whether different indicators that are measuring the same underlying dimension can be combined into one more reliable (and more valid) measure.

In previous studies, we experienced that bibliometric indicators that should measure a similar dimension, lead to different outcomes of the analysis. For example, when explaining grant success, PP10 had no significant positive effect, but PP5 did have a modest positive effect, and again PP1 did not. That makes conclusions about the effect of 'top papers' on grant success difficult to draw. The alternative was to use more indicators in the statistical analysis (PP5 and PP10), but that also resulted in uninterpretable outcomes, as the one indicator suddenly had a strong positive effect, and the other a strong negative effect multicollinearity). These problems can be avoided in the proposed approach.

4. An example

For bibliometrics this is important, as currently many studies use a single indicator for each quality dimension, without much attention for issues like reliability and validity. In this example, we show how one can come to reliable measures and to a better understanding of the underlying evaluation dimension.
We have used in a study a large dataset (N=2579) of researchers from all disciplines, for which we needed bibliometric performance data. The data come from Scopus, and they were manually cleaned in order to avoid problems when more researchers have the same name (synonyms), or single researchers are in the database with more than one name (homonyms). The related bibliometric indicators were retrieved from SciVal. We won't discuss here the details of those indicators (see: Elsevier, Research metrics guidebook), as we are interested in using these indicators to develop the scales to measure the underlying quality dimensions.
For this experiment, we included nine bibliometric indicators as listed in table 1. Firstly, the z-scores at discipline (here operationalized as faculty) levels were calculated, this in order to field-normalize the values. This corrects the data for filed differences in publication and citation behaviour.

Table 1: Scival indicators

	Abbreviation	Indicator
1	P	Total publications
2	P frac	Total publications, fractional counting
3	C	Citations
4	C frac	Citations, fractional counting
5	C/P	Citations per publication
6	FWCI Sum	Sum field weighted citation impact
7	FWCI Average	Average field weighted citation impact
8	P10%	Number top 10% cited papers
9	P10% FN	Number top 10% cited papers, field normalized
10	P10% FN frac	Number top 10% cited papers, field normalized, frac counted
11	P10% share	Share top 10% cited papers
12	PP10%	Share top 10% cited papers
13	PP10% FN	Share top 10% cited papers, field normalized,
14	PP10% FN frac	Share top 10% cited papers, field normalized, frac counted
15	SJR	SJR
16	SNIP	Average SNIP
17	Citescore	Citescore

Secondly, a Principal Axis Factoring (PAF) with oblique rotation and Kaiser normalization was done and this resulted in the identification of three underlying dimensions (Table 2).

Table 2: The bibliometric performance indicators.

Pattern matrix	Relative impact (1)	Total impact (2)	'Reputation' Journal impact (3)
PP10% FN	0.980		
PP10% FN frac	0.950		
FWCI average	0.719		
PP10%	0.670		
C/P	0.593		
P frac		0.919	
P10% FN frac		0.801	
C frac		0.709	
SJR			0.942
SNIP average			0.703

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.
All items are panel-based z-scores
|loadings| < .30 not shown

Which factors can be distinguished?

- The first factor measures *relative impact*, the impact relative to the total output. This factor consists of the items that represent the relative indicators like 'share of top 10% highly cited papers', C/P (average citations per publication), average FWCI. The Cronbach alpha of this scale is 0.914 which is very high.
- The second factor measures *total impact*, and the following items loaded high on this component: Publications (fractionally counted), Citations (fractionally counted) and the fractional count of top 10% field normalized highly cited papers. The Cronbach alpha of this scale is 0.873, again rather high.
- The third factor measures *journal impact*, and on this component the SNIP and SJR loaded. The related Cronbach alpha is 0.851.

Why is the *second factor* called total impact, as it includes one output indicator and two citation-based indicators? Normally these indicators are differently classified as output versus impact. In contrast to this, one may argue that researchers can have impact in two different ways:

- Firstly, by bringing many new ideas (= papers) into the community, which can be assessed and become more or less adopted and used.
- Secondly, the impact in terms of the reception of these new ideas can be measured by citation-based indicators.

But as we see here, these two sides of impact are in fact one, as the factor analysis suggest. This result was not unexpected as we found elsewhere that more output leads to more highly cited papers (Sandström & van den Besselaar 2017), and as is well known also the number of publications and the number of citations do correlate with each other.

The *third factor* represents the impact of the journals in which the papers have been published. As argued elsewhere (Van den Besselaar et al 2019), the underlying variable may be *reputation*, and not so much own impact.

The meaning of the *first factor* is somewhat less clear, as it measures the share of 'good' output among all output. This has a less straightforward interpretation, as for example an early career researcher working in a good team may become co-author of two papers that end up as highly cited papers. Such a researcher would score higher than a researcher who contributed to 50 papers, of which are 30 of very highly cited. For these authors the relative impact component starts high and would be expected to drop off: as they publish more papers their share of top papers should come down. Generally, one can expected that people with *high shares* of top papers will be either on the lower end of productivity (in terms of total output), or exceptional researchers with high output and high overall impact.
If very different cases may end up scoring high on this relative impact variable, the dimension may not be so useful. Indeed, when applying these three variables in a study (Van den Besselaar & Mom, forthcoming), the variable relative impact was almost never close to statistically significant (Table 3 as example). We predict receiving an award, using among others our performance variables. And indeed, relative impact has no effect.

Table 3. Award by gender, faculty and quality*

	B	S.E.	Wald	df	Sig.	Exp(B)	95% CI Exp(B)	
							Lower	Upper
Sex of PhD candidate(1)	-0.714	0.228	9.824	1	0.002	0.49	0.313	0.765
faculty			33.673	5	0			
faculty(1)	-0.702	0.258	7.373	1	0.007	0.496	0.299	0.823
faculty(2)	-1.806	0.402	20.154	1	0	0.164	0.075	0.362
faculty(3)	0.004	0.268	0	1	0.988	1.004	0.594	1.697
faculty(4)	-1.677	0.495	11.462	1	0.001	0.187	0.071	0.494
faculty(5)	-1.125	0.764	2.164	1	0.141	0.325	0.073	1.453
PhD year	-0.041	0.024	2.967	1	0.085	0.96	0.916	1.006
relative impact	0.02	0.1	0.042	1	0.838	1.021	0.839	1.241
total impact	0.465	0.066	49.777	1	0	1.592	1.399	1.812
journal impact	0.474	0.093	26.051	1	0	1.606	1.339	1.926
Constant	-2.143	0.265	65.64	1	0	0.117		

* Logistic regression; all PhD students in the selected faculties; ** Indicators calculated over the t-3 ~ t+3 period;
N=2579; Nagelkerke R Square: 0.177

We also tested the same approach including the full count variants of the various variables (publications, citations, various variants of the P10% papers and the sum FWCI) instead of only the fractional counts, but the results were very similar. These four additional variables loaded on the same component as their fractionalized counterparts, and they do not influence the results of the Principle Component Analysis. Also the Cronbach Alphas did hardly change (See table 5 below for another example).

The factor scores were saved using 'regression' and the three resulting dimensions are included as past performance measures. The correlations between relative impact and total impact, and between relative impact and journal impact are moderate. But between total impact and journal impact, the correlation is low (table 4).

Table 4: Factor correlations

Factor	Relative	Total
Total	0.349	
Journal	0.463	0.162

For further validating the approach, we used the same method on two other datasets, and this resulted in the same three components, which suggest that the approach is robust. The advantage of using the underlying factors to measure performance and reputation over using single indicators seems clear: the stability and reliability of the measurement is strongly improved.

One of the other two datasets had a similar size, also covering all fields on a somewhat more granular level. We again used *Scopus* data, and the bibliometric indicators available in *Scival*. We won't discuss here the details of those indicators, as we are particularly interested in the underlying quality dimension. We included now fifteen bibliometric indicators, so also the non-fractional counts of citations and publications. This resulted again in a three-factor solution, with a similar structure (Table 5). Note that the total impact factor is now the first one, as it consists of much more indicators than in the previous example.

The following factors were found:

- Firstly, the *total impact* factor, and the following items loaded high on this component: Publications (total), Publications (fractionally counted), Citations (total), Citations (fractionally counted), Sum of the FWCI (Field Weighted Citation Impact), and several fractional and full counts of top 10 highly cited papers. The Cronbach alpha of this scale is very high: 0.957.
- The second factor is *journal impact*, and on this component the SNIP, SJR and the CiteScore loaded. The related Cronbach alpha is also very high: 0.933.
- The third factor is *relative impact*, with the items that represent the relative indicators like 'share of top 10% highly cited papers', C/P (average citations per publication), average FWCI. The Cronbach alpha of this scale is 0.924.

The correlations between component 1 and component 2 are small. Both component 1 and 2 correlate moderately strong but negative with component 3 (Table 5). This indicates that those with a high relative impact (so with among their papers many highly cited ones) have only a small oeuvre with a lower absolute impact.
We also tested whether the large first component does have a strong influence on the factor structure, and we therefore reran the PCA with only four out of the eight items like in the first

example above: only the fractional counted indicators were included. As in the first example, this did not change the factor structure.
The factor scores were saved using 'regression' and the three resulting dimensions are included as past performance measures in studies we did (e.g., Van den Besselaar & Mom, forthcoming; see also Table 3).

Table 5. The bibliometric performance indicators.

Pattern matrix	Total impact	Journal impact	Relative impact
P*	0.994		
P frac	0.952		0.376
P10%	0.844		
P10% FN	0.837		
FWCI sum	0.815		
P10% FN frac	0.773		
C frac	0.747		
C	0.747		
SJR		0.951	
SNIP average		0.928	
Citescore		0.916	
P10% share FN			-0.839
FWCI average			-0.828
P10% sgare			-0.706
C/P			-0.697

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization
Rotation converged in 9 iterations.
* all items are panel-based z-scores
|loadings| < .30 not shown

Table 5: Component Correlation Matrix

Component	Relative impact	Total impact
Total impact	-.330	
Journal impact	-.515	.142

7. Conclusions and discussion

The approach described here seems promising. It does appear that we can identify three distinct dimensions in the bibliometric data, and these are similar between the two different datasets we used. The use of these resulting variables have been used in different studies about gender bias in careers, grants and awards (Van den Besselaar & Mom 2021; Van den Besselaar & Mom forthcoming). The outcomes of those studies are less vulnerable for the choice of individual indicators. The use of the three dimensions in studies suggest that the relative impact is not very useful, as this variable gives similar scores for very different performance levels. Furthermore, the Journal Impact variable is for theoretical reasons probably more a reputation than as performance measure. Consequently, we reduced the original 17 indicators into two underlying variables with a measurement model:

- (total) impact
- (journal impact =) reputation

8. Acknowledgements

This paper is an outcome of the GRANteD project. This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824574

9. References

Butler, L. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy* **32** pp. 143–155.

Cole JR (1979) Fair science: Women in the scientific community. The Free Press.

Cole S & Fiorentine R (1991) Discrimination against women in science: the confusion of outcome with process. In: *The outer circle: Women in the scientific community*.

Cruz-Castro, L and Sanz-Menendez, L (2019). *Grant Allocation Disparities from a Gender Perspective: Literature Review*. Synthesis Report. https://doi.org/10.13039/501100000780

Cruz-Castro, L., & Sanz-Menéndez, L. (2021). What should be rewarded? Gender and evaluation criteria for tenure and promotion. *Journal of Informetrics*, 15(3), 101196. https://doi.org/10.1016/j.joi.2021.101196

Drenth P & K Sijtsma (2006) Test theory. Introduction in the theory of the psychological test and its applications (in Dutch). Boom.

Espeland W.N., Sauder M. (2016) *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*.

Glänzel W & Debackere K (2009) On the "multi-dimensionality" of ranking the role of bibliometrics in university assessments. In: Ranking Universities. Brussels: VUB.

Hicks D., Wouters P., Waltman L., de Rijcke S. & Rafols I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, April 22, 2015

Kahneman D, Sibony O, Sunstein CR, (2019) *Noise, a flaw in human judgement*. Hachett Book Group, New York.

Neumann, M., Niessen, A. S. M., Hurks, P. P. M. & Meijer, R. R., (30-Jan-2023. E-pub ahead of print). Holistic and mechanical combination in psychological assessment: Why algorithms are underutilized and what is needed to increase their use. In: *International Journal of Selection and Assessment*.

Sandström U. & van den Besselaar P. (2016). Quantity and/or Quality? The Importance of Publishing Many Papers. PLoS ONE 11(11): e0166149. doi:

Van den Besselaar P, Mom C, Cruz-Castro L, Sanz-Menendez L (eds.) (2020) *Identifying gender bias in grant allocation, and its causes and effects*. Deliverable D2.1, GRANteD project.

Van den Besselaar P, Heyman U, Sandström U, Perverse effects of output-based research funding? Butler's Australian case revisited, *Journal of Informetrics* **11** 2017, 905-918

Van Den Besselaar, P., Sandström, U., & Mom, C. (2019). Recognition through performance and reputation. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings* (Vol. 2, pp. 2065-2069).

Van den Besselaar P, Sandström U (2020) Bibliometrically disciplined peer review; using indicators in research evaluation. In: *Scholarly Assessment Reports* **1**

Van den Besselaar P & Mom C (2021) *Gender bias in grant allocation, a mixed picture*. Preprint.

van den Besselaar P & Mom C, *Gender and Merit in Awarding Cum Laude for the PhD Thesis.* Forthcoming.

Wilgaard L, Schneider JW, Larsen B (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics* 101, 125-158

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