Do bibliometrics and altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime, altmetrics, and citation data

Lutz Bornmann and Robin Haunschild

Altmetrics in research evaluation

- Altmetrics denote non-traditional metrics which represent an alternative form of impact measurement instead of using citations
- Altmetrics are views, downloads, clicks, notes, saves, tweets, shares, likes, recommends, tags, posts, trackbacks, discussions, bookmarks, and comments
- Providers of altmetrics propose them for broader impact measurement (societal or cultural attention besides academic impact)
- Main problem, the meaning of altmetrics is barely understood: What do these metrics measure?

Research questions

- We address the question of unclear meaning of altmetrics
- We study their relationship to scientific quality of papers (as measured by peer assessments) in comparison to traditional metrics
- Some years ago, F1000Prime was launched as a new type of post-publication peer-review system
- Experts identify, evaluate and comment on the most interesting papers they read for themselves each month
- (1) We analyze the underlying dimensions of measurement for traditional metrics and altmetrics – by using PCA
- (2) We test the relationship between the PCA components and quality of papers (F1000Prime assessments) – using regression analysis

Four different datasets used

- (I) Initial dataset: Papers selected for F1000Prime are rated by the Faculty members as “Good” (1), “Very good” (2), or “Exceptional” (3). Since for many cases a paper is assessed by several members, FFa (F1000 Article Factor) is given as the sum
- (II) CiteScores (journal metric) were downloaded from https://journalmetrics.scopus.com
- (III) Citation counts from WoS and Scopus as well as JIF were retrieved from in-house databases
- (IV) Altmetrics data (i.e. Mendeley, Twitter, Altmetric attention score) were used from the company Altmetric
- Altmetrics and F1000Prime data were matched with citation data via the DOI (2011-2013 were included, n=33,683)

Statistical procedures

- **Analyzing underlying components:**
- PCA is used to transform the different indicators considered here into new, uncorrelated variables
- New variables are named as principal components; each component is a linear combination of the indicators
- Indicator space can be reduced by selecting and interpreting only the first few components. These components explain most of the variance
- **Analyzing correlation between the components and F1000Prime scores (as proxies for the quality of papers):**
- We performed negative binomial regression model (NBREG)

Principal components analysis (PCA)

Eigenvalues and (cumulative) proportions of total variance for metrics data (n=33,683)

Component	Eigenvalue	Proportion	Cumulative proportion
Comp1	**6.28**	**72%**	**72%**
Comp2	**1.13**	**13%**	**84%**
Comp3	0.71	8%	93%
Comp4	0.46	5%	98%
Comp5	0.10	1%	99%
Comp6	0.05	1%	99%
Comp7	0.02	0%	100%
Comp8	0.02	0%	100%
Comp9	0.01	0%	100%

- Components as the results of the PCA using the metrics data
- First principle component is able to explain 72% of total variance in the metrics
- There is a common cut-off point at a proportion of 80%

Principal components analysis (PCA)

Principal components analysis for metrics data (n=33,683)

Variable (logarithmized)	Comp1	Comp2
3-years citations, WoS	**0.41**	-0.08
Citations, WoS	**0.46**	-0.18
Citations, Scopus	**0.46**	-0.25
3 years citations, Scopus	**0.42**	-0.15
Tweets	0.09	**0.73**
Mendeley readers	**0.40**	0.31
Altmetric attention score	0.05	0.35
CiteScore	***0.17***	0.23
Journal Impact Factor (JIF)	***0.20***	0.25
Eigenvalues	6.28	1.13
Cumulative proportion	0.72	0.84
Cut-off point: $0.5/(\text{eigenvalue})^{1/2}$	**0.20**	**0.47**

- Loadings for the first two principal components and the metrics data – indicators with a correlation greater than the cut-off point are emphasized
- Two correlations in bold face and in italics are marginally below the cut-off point

Negative binomial regression analysis (NBREG)

- Two dimensions: (1) impact on academia (readers and citers) which is partly dependent on the citation impact of the publishing journals. (2) wider impact (beyond science) which is largely independent from academic impact (tweets).
- We are interested in the correlation between the quality of papers (measured by FFa) and the two impact components from the PCA
- We included the FFa as dependent and the scores for both components as independent variables in the negative binomial regression analysis
- We calculated the statistical and practical significance

Negative binomial regression analysis (NBREG)

Beta coefficients of and marginal effects from the NBREG (*t* statistics in parentheses, n=33,683)

	F1000Prime score	Marginal effects (+SD)
Scores for component 1	0.11***	0.62
	(51.83)	
Scores for component 2	0.09***	0.30
	(29.67)	
Constant	-0.31***	
	(-20.72)	

- The coefficients of both independent variables are significant; however, the number of papers is very high
- Marginal effects – as a measure of practical significance – are more interesting
- An average increase of a paper's expected FFa by 0.62 is related to a standard deviation change in the first component (citation and reader impact)

Conclusions

- Intention of our study: should altmetrics really be used in research evaluation practice?
- An important requirement for this use is that the metrics are related to scientific quality
- Without this relationship bad or fraudulent research might be rated high, because it simply received attention in society
- Results of the PCA show that citations (on the paper and journal level) and reads measure similar, but tweets measure different things
- Similar results have been published in other studies

Conclusions

- Results of the NBREG indicate that citations and reads are significantly stronger related to quality (as measured by FFa) than tweets
- Results indicates potential use of Mendeley reader counts but question the use of tweets for research evaluation purposes
- Possible solution for the use of Twitter data in research evaluation:
  Research evaluations using Twitter data should only include papers fulfilling certain quality standards (e.g. papers which have been frequently cited over many citing years)
- In the broad impact measurement of research in the UK REF, only high-quality research is considered