# The Best Five – Do Authors' Assessments of their own Publications Correspond to Metric Based Rankings?[1]

Marion Schmidt schmidt@forschungsinfo.de / iFQ Institute for Research Information and Quality Assurance / Schützenstr. 6a / Berlin, 10117 (Germany)

Jörg Neufeld neufeld@forschungsinfo.de / iFQ Institute for Research Information and Quality Assurance / Schützenstr. 6a / Berlin, 10117 (Germany)

## Introduction

For some years now, the on-going discussion on the validity of bibliometric indicators in research evaluation and especially their potential effects on scientists – who might try to publish as much as possible – have evolved into a paradigm which can be described by the phrase "Quality not quantity"[i]. Funding agencies like the German Research Foundation (DFG), the European Research Council (ERC), the American National Science Foundation (NSF) as well as the National Institute of Health (NIH) have adopted policies which demand scientists to cite or highlight only a limited number of their most important publications in funding proposals. Reviewers of funding decisions are to be focussed on those selected papers as well as on the proposal itself. The respective selection of papers is supposed to be based on the scientists' subjective ratings which are not expected to correspond completely with bibliometric impact measures. Applicants therefore have the possibility to highlight publications that are especially innovative – which may not be adequately reflected by bibliometric indicators. However, as far as researchers adhere to metrics – or believe the funding agency or the reviewers being adhered to metrics, they will simply base their selection on just these metrics. On the other hand, subjective researcher ratings and metrics may conform with each other although the respective researchers' selection is independent from metrics.

On a wider level, the funding agencies' policy change aims to generally counteract the importance of metrics in evaluation of science, allocation of funding and appointments.

How do selected papers actually fare when evaluated with bibliometric indicators? Subjective ratings of their own papers by researchers have been analysed quite rarely although they are apparently getting more and more important to funding decisions. Based on a sample of the publication corpora of highly cited scientists, Aksnes (2006) asked them how they assessed the scientific contribution of their papers, and to categorize their papers as empirical, theoretical, methodical, or review. He observes overall correlations of 0.56 and 0.52 between raw and field-

1 This work was supported by the German Federal Ministry of Education and Research (BMBF)

normalized citation counts, respectively, and authors' perception of their papers' scientific contribution. For the empirical, theoretical and methodical papers, there is no significant difference to the overall value, while for the reviews, the correlation is weaker.

Porter et al. (1988) gather two cohorts of Sloan Chemistry Fellows and ask them to augment publication data and nominate their three best papers. In a basic probability estimate, they calculate a ratio of observed to expected overlap as 34% to 11% between best and most cited papers in the 1974 cohort. They use content analysis in order to categorize papers as theoretical, empirical and methodological papers with the result that most cited papers are methodological to a larger extent, whereas best ones are more often theoretical and empirical.

A number of studies have compared bibliometric indicators with peer ratings:

Rinia et al. (1998) report significant correlations between peer ratings with citation-based indicators, notably citations per publications and the field normalized citation rate, in physics research groups. Van Raan (2006) compares the h-index and the field normalized citation rate with peer judgments. Both indicators discriminate very well between excellent or good chemistry research groups on the one side and less good on the other side. Similar results are reported by Moed (2005).

Bornmann and Leydesdorff (2013) correlate peers' ratings of papers based on the online service *Faculty of 1000* with citation-based indicators, of which the ratio of highly-cited papers correlates best. Franceschet & Costantini (2010) report significant correlations between peer ratings and raw citations, and, to a lesser degree, journal impact factors in a study comprising several fields.

Some studies are concerned with scientists' general perception of citations:

Aksnes and Rip (2009), in a follow-up study of Aksnes (2006), conclude that, on the individual paper level, citations are not a reliable indicator as quite a few scientists perceive own papers as under- or overcited. An older survey of academics' views of citations and evaluative bibliometrics is presented by Collins (1991). Most respondents accepted quantitative indicators as part of research assessment, but favoured productivity indicators over citation indicators. Hargens and Schuman (1990) analyse the relationship between biochemists' and sociologists' use of citation data and the field-normalized medium citation score of their own publication corpus. The number of citations a researcher receives is only weakly correlated with his usage of citation indexes in case of biochemists, but the correlation is significantly stronger in case of sociologists.

This research in progress paper evolved from a service project which was conducted by the iFQ recently and which was focussed on calculating bibliometric indicators for a benchmarking process of three German universities in chemistry and physics. As part of a publication validation process, scientists' subjective ratings of own papers have been requested and are compared with several state-of-the-art bibliometric indicators. In this paper, we analyse to which extent the best

papers according to subjective assessment are concordant with the best papers identified by metrics; and furthermore, which indicator shows best concordance.

## Data and Methods

### Data and bibliometric indicators

Publications of scientists belonging to chemistry and physics institutes of three universities have been searched in *Web of Science*[ii] by way of institutional addresses and names of currently employed scientists. After that, these scientists have been contacted and invited to validate the publication data from 2005 and 2010 searched for them, and if necessary to delete or add publications or to upload own publication lists. They also have been asked to mark five publications they considered being their best in the period from 2005 to 2010, independently of institutional affiliations. They were asked in exact wording: "In order to answer the question whether bibliometric indicators correspond to your own assessment we ask you to flag up to five publications you consider your best."[iii]

In comparison to Asknes (2006) and Porter et al. (1988), who both preselected their respective samples based on citation numbers and funding applications, our sample includes scientists in different career steps and irrespectively of their specific publication performance.

For each of the 182 scientists who have marked their best publications, all citable items[iv] between 2005 and 2009 have been selected. A threshold of 10 publications per person has been applied resulting in a sample of 109 persons. For all publications the following indicators have been calculated:

- citations per paper (within 3-year citation window), (CPP),
- field-normalized citation rate (within 3-year citation window), (FNCR),
- journal-normalized citation rate (within 3-year citation window), (JNCR),
- Journal Impact Factor, (JIF)[v].

Due to the fact that relevant portions of a year's publications are entered into WoS during the first months of the following year, we have omitted the publication year 2010 and will include these publications in an updated calculation in summer. The same holds for the possibility of considering the field-normalized ratio of highly-cited papers whose distribution among the analysed papers between 2005 and 2009 is too sparse.

**Comparing researcher's and metrics based rankings**

(1) Cohen's Kappa

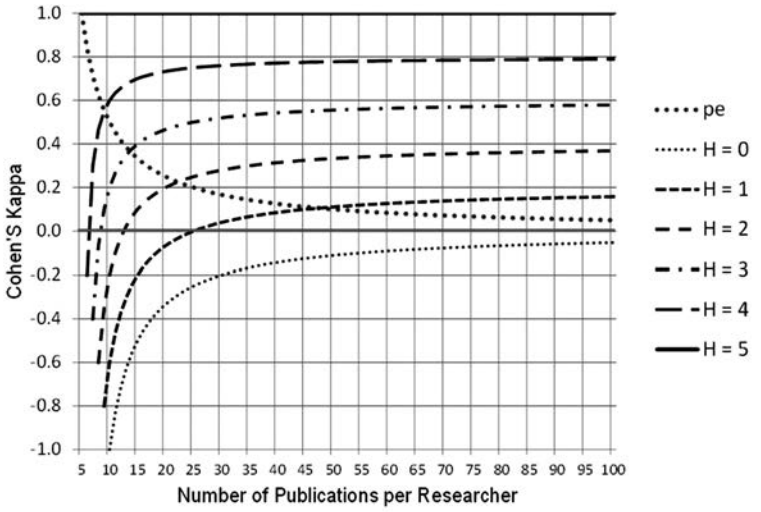
Depending on the total number of publications a researcher has published, there is a certain probability that researcher assessment and metric based ranking coincide by chance. We therefore chose Cohen's Kappa (*K*) as a measure of accordance because it takes into account this randomly expected concordance[vi].

$$K = \frac{p_0 - p_e}{1 - p_e}$$

$p_0$ = observed relative agreement
$p_e$ = concordance expected by chance

Kappa varies between -1 and +1 whereby 0 indicates the accordance of empirical concordance and the concordance expected by chance. Figure 1 illustrates how Kappa relates to the number of publications and the number of concordant judgments.

*Figure 1. Values of Kappa and expected concordance ($p_e$) depending on the number of publications per researcher and the number of hits (H) – provided that exactly five publications have been marked as best.*




Obviously with a growing number of publications the values of Kappa converge to limiting values which are equal to the respective shares of concordant judgments.

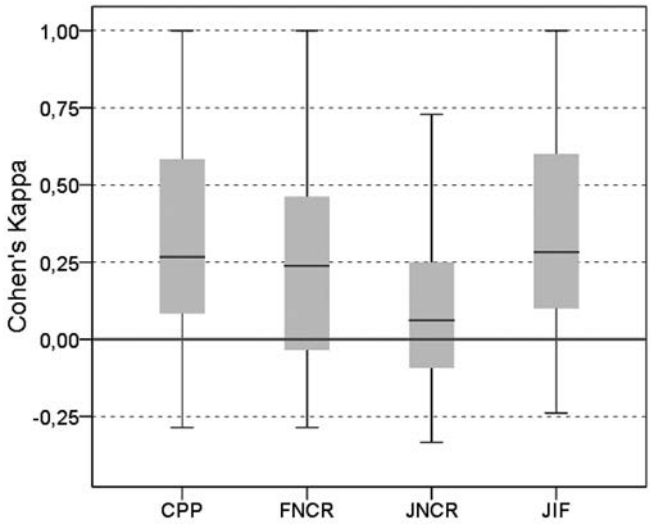
(2) Median relative rank of the best five

Kappa ignores the ordinal information, which is given by the metric based rankings. Each publication which is marked as "best five" by a scientist has its specific position in the metric based ranking of all publications of this scientist. For example, Kappa treats a best five publication which is ranked sixth regarding CPP as discordant (no hit) even if the number of publications is quite high. By contrast the *median relative rank* of the best five reflects the relative position of the best five in the metric based rankings. Medians beneath 0.5 indicate scientists who tend to evaluate their publications contrary to the metric based rankings and medians above 0.5 point to a similar/equal assessment.

## Results

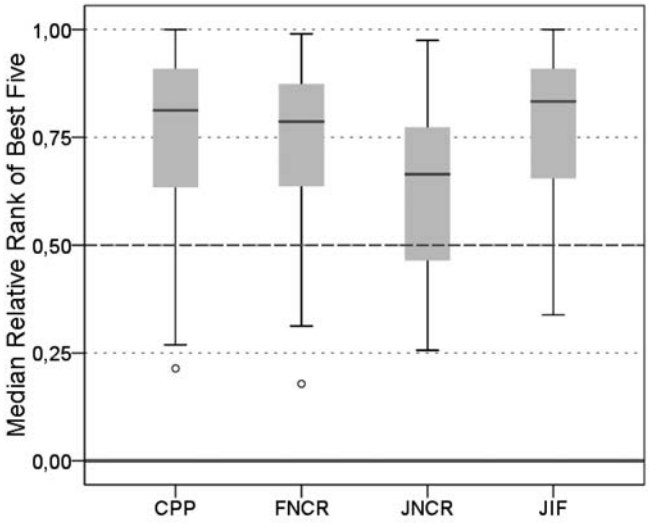
For each researcher we calculated four values of Kappa, one for each indicator (CPP, FNCR, JNCR, JIF). Figure 2 shows boxplots for each distribution of Kappa. Medians of CPP, FNCR and JIF are located near to .25. Based on usual evaluation scales for Kappa this indicates rather weak fit of scientists' judgments and metrics based rankings. However, values are quite dispersed; roughly 25 per cent of the scientists show values of .60 or higher regarding CPP and JIF. The accordance of scientists' judgments and source normalized impact (JNCR) is even smaller and barely better than would be expected by chance.

*Figure 2. Boxplots of Cohen's Kappa measuring accordance between scientists' ratings (best five) and indicator based ranking.*



The boxplots of the median relative rank in Figure 3 show the same pattern for all indicators. Again the correlation between source normalization (JNCR) and scientists' judgments is weaker than with the other indicators. Apart from the JNCR, the median relative rank of the best five publications is located in the upper third of the distributions in 75 per cent of the cases. Roughly 50 per cent are located in the top quintile (> 0.8).

*Figure 3. Boxplots of median relative rank of the best five publications in metric based rankings.*




## Discussion

As the results obtained so far demonstrate, all in all researcher assessments and indicator based rankings are not perfectly consistent. However, values are dispersed, and regarding CPP, JIF and FNCR some cases show a perfect match. It still remains an open question whether these scientists actually base their perception of their own papers on the citation scores or JIF – which they can easily access via the databases' online interfaces – , or whether the concordance is merely a product of the underlying quality of the papers that leads both to citations and a good subjective rating (which in turn could also lead to submitting the paper to a journal with a higher impact factor, which in turn may lead to higher visibility and more citations). In the course of our project we would like to further investigate this issue by analysing the scientists' specific patterns. The next planned step is to survey them about the specific criteria for perceiving and defining publications as best five and to relate subjective assessments with respective author positions and estimations of their own contribution to a paper.

Regarding the funding agencies' policies the question arises if the desired effect of giving scientists the opportunity to behave blind to metrics and highlight unconventional research is actually realistic or if scientists' view of the world tends to be already shaped by bibliometrics. Funding agencies which are interested in a selection of best five publications independent of metrics should explicitly state this in their guide lines. Possibly they could even state criteria (like innovativeness). However, when, as in the case of the ERC (European Commission 2012: 20), applicants are invited to quote citation counts of the named best five, there is a risk that the selection of the best five is based just on these counts.

## References


Asknes, D. W. (2006). Citation Rates and Perceptions of Scientific Contribution. *Journal of the American Society for Information Science and Technology*. 57, 169–185.

Asknes, D.W. & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy*, 38, 895–905.

Bornmann, L. (2011). Scientific Peer Review. *Annual Review of Information Science and Technology*, 45, 199–245.

Bornmann, L. & Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics*, 7, 286–291.

Cicchetti, D. V. & Feinstein, A. R. (1990). High Agreement but Low Kappa: II. Resolving the Paradoxes. *Journal of Clinical Epidemiology*, 43, 551–558.

Collins, P. M. D. (1991): *Quantitative assessment of departmental research. A survey of academics' views*. SEPSU Policy Study No. 5. Science and Engineering Policy Studies Unit of the Royal Society and the Fellowship of Engineering, London.

DFG (2010). *Quality not Quantity – DFG Adopts Rules to Counter the Flood of Publications in Research*. Press Release. No. 7, 23. February 2010. Retrieved March 11, 2013 from http://www.dfg.de/en/service/press/press_releases/2010/pressemitteilung_nr_07/index.html.

European Commission (2012). *ERC Work Programme 2013*, European Commission, C(2012) 4562 of 09 July 2012.

Feinstein, A. R. & Cicchetti, D. V. (1990). High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, 43, 543–549.

Franceschet, M. & Costantini, A. (2010). The first Italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, 5, 275–291.

Hargens, L. L. & Schuman, H. (1990). Citation Counts and Social Comparisons: Scientists' Use and Evaluation of Citation Index Data. *Social Science Research*, 19, 205–221

Moed, H. F. (2010). New Developments in Electronic Publishing and Bibliometrics, ESSS Summer School, Berlin, 16 June 2010. Retrieved March 11, 2013 from http://www.scientometrics-school.eu/images/esss1_Moed.pdf

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Information Science and Knowledge Management, 9. Dordrecht: Springer

Porter, A. L., Chubin, D. E. & Jin, X.-Y. (1988). Citations and Scientific Progress: Comparing bibliometric measures with scientist judgements. *Scientometrics* 13, 103–124.

Rinia, E. J., van Leuween, T. N., van Vuren, H. G. & van Raan, A. F. J. (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, 95–107.

Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67, 491–502.

i http://www.dfg.de/en/service/press/press_releases/2010/pressemitteilung_nr_07/index.html

ii In a bibliometric database based on WoS raw data.

iii We abstained from a more elaborate survey because information letters and validation tool should have not been too complex.

iv Articles, Letters, Reviews published in journals.

v This indicator has been chosen because we assumed that scientists are rather attentive to the impact factors of the journals they publish in. The journal impact factor of a given journal and year is calculated by us in accordance with Thomson Reuters as ratio of the number of citations to all documents of two preceding years and the number of all articles and reviews of that journal in these two years (see Moed, 2010).

vi Note that in every case both researchers and rankings have the same number of positive (normally five) evaluated publications. Therefore the issues concerning the Kappa coefficient discussed by Feinstein & Cicchetti (1990) and Ciccetti & Feinstein (1990) – e.g. symmetry/ asymmetry of the marginal totals – are basically not relevant in this context.