Journal of Informetrics 11 (2017) 888–891

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correspondence

Why not to use the journal impact factor as a criterion for the selection of junior researchers: A comment on Bornmann and Williams (2017)

Bornmann and Williams recently published an interesting study where they analysed an impressive dataset. They found trivial to moderate associations between, on the one hand, early-career publication performance in terms of number of produced publications and proportions of publications in high Journal Impact Factor (JIF) journals, and on the other hand, late-career publication performance in terms of total and mean number of citations (Bornmann & Williams, 2017). On the basis of these results, they highlighted their conclusions that the results indicate that "the JIF is able to discriminate between researchers with different publication performance later on" and that "the JIF should not be used as the sole criterion for identifying later success" (Bornmann & Williams, 2017). This prescription to not *solely* rely on the JIF when rewarding work and allocating resources implies that while JIF should not be relied on as sole criterion, the authors deem it an important criterion nonetheless. However, methodological constraints preclude drawing conclusions to that effect on the basis of the analysed dataset, and even if those methodological problems are overlooked, the results do not seem to justify enthusiasm about the JIF as diagnostic instrument to select researchers. More generally, the seemingly unquestioning acceptance of JIF as indicator of scientific quality discounts evidence to the contrary.

The dataset analysed by Bornmann and Williams (2017) is impressive, comprising three cohorts of in total 13 180 researchers. The dataset contains, for each researcher and every year from 1998 to 2012, each publication and the corresponding JIF. These data were used to compute four variables for each researcher: the total number of publications in the first five years of their career, the proportion of publications in so-called Q1 journals in those five years, and for the following years, the mean normalized citation score (MNCS) and the total normalized citation score (TNCS). Bornmann and Williams (2017) conducted their analyses separately for the cohorts of researchers whose careers started in 1998, 1999, and 2000 (and therefore, with 15, 14, and 13 'senior years'), providing some safeguard against temporal anomalies. The authors used frequent visualisation, routinely provided confidence intervals, and uploaded their manuscript to a preprint server (https://arxiv.org/abs/1706.06515). All of these are commendable research practices. At the same time, some odd choices were made. The two predictors were not used in their original continuous operationalisation, but instead are transformed into quartiles. 'Quartilization' being a slightly less harmful form of dichotomization, this is normally strongly discouraged (Altman & Royston, 2006; DeCoster, Iselin, & Gallucci, 2009; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993): quartiles are unlikely to reflect naturally occurring discontinuities, and artificially imposing such discontinuities may introduce statistical artefacts. For example, after 'quartilization', researchers in the first percentile are equal to researchers in the 24th percentile, whereas researchers in the 24th percentile are different from researchers in the 26th percentile. The authors have not fully disclosed the dataset and analysis scripts (Peters, Abraham, & Crutzen, 2012; Peters, Kok, Crutzen, & Sanderman, 2017), so I will have to assume that this quartilization did not influence the results. Apart from this odd choice (and the repeated but impossible claim that p values of 0.000 were found), however, the statistical analyses seem to have been conducted very well.

1. Correlation does not imply causation (not even with a longitudinal design and a disclaimer in the discussion)

The main problem of this paper does not have a statistical nature – nor a potential statistical solution. The main problem of this paper has a methodological nature, and boils down to the creed –cliché, almost – that correlation does not imply causation. This unfortunate methodological fact of life is often invoked when observational data are used to evoke a suggestion of causality, which is exactly what the paper by Bornmann and Williams (2017) does. The title of their paper asks whether "the journal impact factor [can] be used as a criterion for the selection of junior researchers." When selecting staff members, employers or funders mean to select staff members who have the competences required by the job at hand. When selecting researchers, these competences are the competences required to do research, or more specifically, to accurately and appropriately employ the scientific method to learn about reality (one could adopt the cynical perspective that the competences required of a researcher are not to do proper research, but instead to publish many papers in high JIF journals,

http://dx.doi.org/10.1016/j.joi.2017.08.001

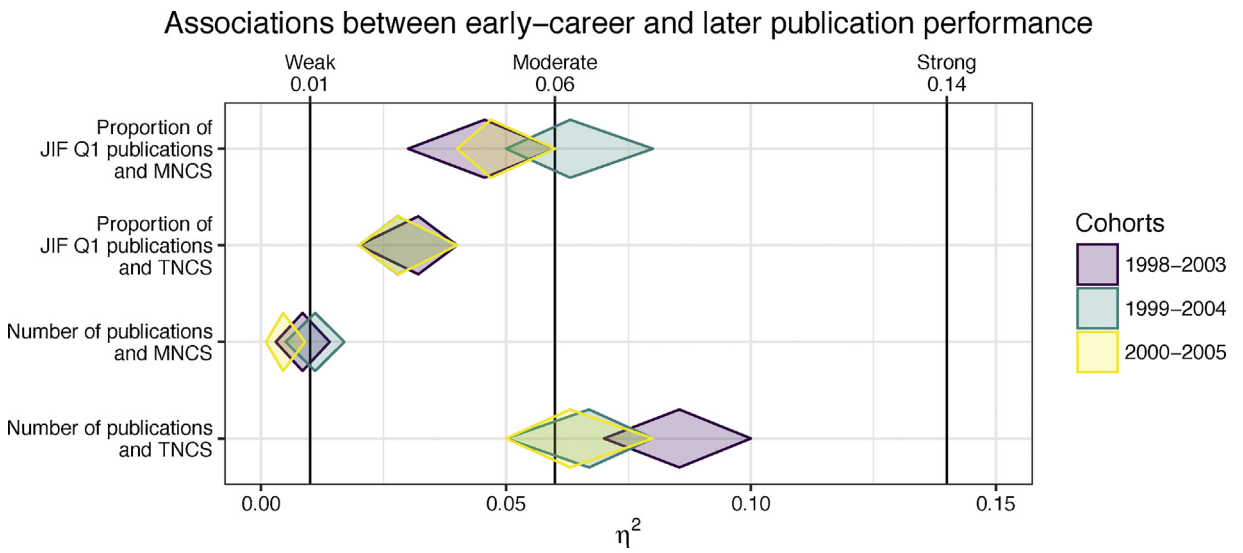

Fig. 1. The effect sizes of the Bornmann and Williams (2017) paper visualized with their confidence intervals and with vertical lines indicating 'weak', 'moderate' and 'strong' effect size values.

but I will assume that the selecting agent aims to contribute to, rather than corrupt, science). Instruments used in selection of researchers, therefore, derive their value from identifying those competences in a researcher. This means that answering Bornmann and Williams' (2017) titular question requires establishing whether early-career high JIF publications are indicative of scientific competence.

Bornmann and Williams (2017) discuss this in terms of two competing hypotheses: the "sacred spark" theory and the "accumulative advantage" theory (both discussed in Allison & Stewart, 1974). The first postulates that early publication performance is indicative of "ability and motivation to do creative scientific research" (Allison & Stewart, 1974, p. 596); the second that early publication performance conveys resources (e.g. opportunities, employment, funding, recognition) that secure later publication performance. In their discussion of these explanations, Bornmann and Williams (2017) argue that "whether the correlation is causal or spurious, our research shows that early success is a predictor of later success" (Bornmann & Williams, 2017, p. 796). This statement is true with the statistical definition of 'prediction' (i.e. synonymous with 'association'): in that definition, it is equally true to say that ice cream sales predict shark attacks. But the statement is false if the general definition of 'prediction' is used, which implies that the knowledge of the predictor grants some advantage when trying to obtain (or avoid) the predicted variable. If researchers' early publication performance subsequently provides them with more resources and opportunities which ultimately result in later publication performance, this means that if selecting actors (e.g. funders or employers) select researchers based on early publication performance, they are *creating* the causal link, instead of leveraging causality that exists outside of their decision as a 'selector'. In other words, there is no advantage for the selector. If later publication performance is a causal consequence of the afforded opportunities instead of scientific competence, affording such opportunities to a researcher (by selecting the researcher) creates the publication performance. In that scenario, selecting researchers on the basis of early-career publication performance affords no advantages to the selector. By selecting researchers based on their early-career publication performance, the selector will simply create a self-fulfilling prophesy, and in doing so, forego opportunities for selecting on variables that *are* indicative of relevant competences.

2. Predictive value of early-career publication performance

In my view, this first methodological problem in itself should suffice to refrain from attaching any practical recommendation to results inferred from this dataset. However, even when assuming that the accumulative advantage theory does not hold, these results still do not mean that early-career publication performance is a promising or valuable selection tool. The effect sizes (i.e. association strengths between early-career publication performance and later publication performance) are generally low, as illustrated in the diamond plot (Peters, 2017a, 2017b) in Fig. 1, where black vertical lines represent the thresholds for weak, moderate and strong associations (the R code to produce this plot is available at the Open Science Framework at https://osf.io/pj8vz). Of the six associations between proportion of Q1 JIF publications and MNCS and TNCS, only one point estimate qualifies as a 'moderate' association ($\eta^2 > 0.06$; see Fig. 1), and when acknowledging sampling variability by looking at confidence intervals, the evidence suggests that the association between Q1 JIF publications and both later publication performance indicators are likely weak in the population. When looking at number of publications, only one estimate firmly suggests that a moderate association in the population may exist ($\eta^2 > 0.06$ for both bounds of the confidence interval for TNCS), whereas the association with MNCS may well be trivial (i.e. $\eta^2 < 0.01$) in the population.