Chris Callaghan's criticism of the National Research Foundation's rating methodology: A rebuttal

AUTHOR:
Christo Boshoff[1]

AFFILIATION:
[1]Department of Business Management, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Christo Boshoff

EMAIL:
cboshoff@sun.ac.za

KEYWORDS:
peer review; NRF rating; bibliometrics

HOW TO CITE:
Boshoff C. Chris Callaghan's criticism of the National Research Foundation's rating methodology: A rebuttal. S Afr J Sci. 2018;114(7/8), Art. #a0278, 4 pages. http://dx.doi.org/10.17159/sajs.2018/a0278

PUBLISHED:
30 July 2018

© 2018. The Author(s). Published under a Creative Commons Attribution Licence.

No scientific endeavour – its methodologies, processes, procedures or anything related to it – should *ever* be above reproach, critical evaluation or reconsideration, including the methodology of South Africa's National Research Foundation (NRF) for rating those researchers who apply for a rating. Chris Callaghan's[1] views on the NRF rating methodology must thus be welcomed.

However, any such criticism must be fair, balanced, objective, properly justified and uncontaminated by personal grievances.

Callaghan's critical review of the NRF rating methodology falls short on a number of these grounds. Firstly, as a management *scientist*, he should have done a far better job of understanding the methodology before embarking on a critical review. Secondly, his recommendations for improvement unfortunately fall foul of the same criticism that he levels at the current NRF methodology. Thirdly, the entire rating system cannot reasonably be completely wrongheaded, with no positive consequences at all. Fourthly, he ought to have been more forthright with his readers, the reviewers of his paper and the editor of the *South African Journal of Science* by declaring his personal experience with the rating methodology. My rebuttal of Callaghan's criticism will be structured around these points, but limited to the domain of Management Sciences.

Key features of the NRF rating process

It must be noted at the outset that the NRF rating methodology is a peer review methodology. The entire process is thus constructed on this key feature and must be understood in that context.

A more careful review of the NRF rating process would have revealed that each applicant submits six names of potential reviewers to the NRF. The members of the Specialist Committee or expert panel (the members are appointed for 4-year terms) then select three of these nominated reviewers. To this list, a further three independent reviewers not nominated by the applicant are added to review the application (Callaghan's view that 'the NRF system works through reviewers chosen by the person being rated' is clearly not entirely correct). At least six reviewers are thus asked to review each applicant.

All reviewers are, of course, expected to be objective and fair in their evaluation, regardless of who nominated them. The nominated reviewers are carefully selected to ensure that they have the required knowledge and expertise to conduct the evaluation. In order to eliminate potential bias, the application form explicitly asks the applicants about their relationship with the nominated reviewer, and this explanation is screened to exclude cases in which the relationship is considered to be too close (supervisor, PhD student, research team colleague or 'life-long friends' in Callaghan's parlance). In a further attempt to preclude bias, applicants may ask the panel not to send their application to reviewers who they feel may not be objective in their assessment.

Once the peer reviewers' reports have been returned to the NRF, they are screened for suitability by the Specialist Committee, an Assessor and a Chairperson. Excessively negative and excessively positive reviewer reports are discarded as potentially marked by bias. The reasons for the rejection are formalised for auditing purposes and to counteract the 'gatekeeping' phenomenon.

The members of the Specialist Committee then consider the peer reviewers' reports and attempt to reach consensus on what an applicant's rating should be. Once the evaluations and suggested ratings of the Specialist Committee have been concluded (all members must read all the reviewers' reports), an independent Assessor, who is uninformed about what the Specialist Committee's recommendations are, enters the meeting. The Assessor will also have been tasked to review all the applications *independently* before the meeting. The Specialist Committee and the Assessor then reach consensus on the appropriate rating for each applicant (including, of course, a 'rating unsuccessful' decision, if deemed appropriate). Once this consultation has been completed, an independent Chairperson, who has also *independently* evaluated the applications of all the applicants, enters the meeting. The Chairperson's suggested ratings are then compared with those of the Specialist Committee and the Assessor; and in most cases agreement is reached on an applicant's appropriate rating. If not, the applications are referred to an Executive Evaluation Committee for review. The members of the Executive Evaluation Committee are the six Chairs of the different evaluations panels, two Convenors of the Specialist Committees, and three NRF executives, including the Deputy CEO of the NRF (who also chairs the meeting).

Clearly, despite the potential failings of human judgement, significant effort is built into the NRF rating methodology to minimise biased evaluations. With the considered involvement of six reviewers, four to eight Specialist Committee members, one Assessor and one Chairperson, no individual can manipulate the assessment in order, in Callaghan's words, to 'settle scores'. A rating outcome is thus not the decision of 'a small group of evaluators' only, as Callaghan contends. The 'power abuses' he refers to are simply not possible. In addition, if the rating outcome is considered inappropriate, the aggrieved applicant has the right to appeal.

Typically, an NRF rating application is thus reviewed by at least 12 different evaluators. It is just not credible that all 12 would be consistently biased against any individual applicant, even though the evaluations are not anonymous. It is thus difficult to accept Callaghan's suggestion that the process is prone to excessive subjectivity bias, especially if one considers that he regards the journal peer-review process – which typically consists of only

two reviewers and an editor – as superior because it is 'systematic'. I want to argue that the current NRF methodology is just as systematic, step-by-step, as the journal peer-review process.

In stating that a rating decision depends on a 'handful of reviewers' only, Callaghan demonstrates his deficient grasp of the NRF methodology. This fundamental misapprehension casts serious doubt on the validity of his entire assessment.

Callaghan's contention that the NRF's rating methodology is discriminatory also calls for scrutiny, as we must distinguish between discrimination and *unfair* discrimination. The sports scoreboard *discriminates* between the winner and the loser. The editor *discriminates* between the manuscript that will be published and the one that will not be published. The Nobel Prize committee *discriminates* between the recipient of the Nobel Prize and the nominees who do not win it. The football referee has *power* to decide on a penalty or not. The editor has *power* to decide on the suitability of a manuscript. The Nobel Prize committee has *power* to decide who receives the Nobel Prize. 'Discrimination', power and the associated hierarchies of distinction are facts of life. Why should academia be any different? Forms of 'discrimination' will thus be true of any rating system, regardless of the methodology used.

Callaghan's specific claim that the NRF rating methodology unfairly discriminates against non-white researchers (for which he offers no supporting evidence) is contradicted by the fact that 52% of scientific papers published in 2013/2014 were published by non-white authors[2] (the figure today may be even higher).

His contention that the NRF methodology discriminates against those who are unable to form 'personal relationships' borders on the laughable. The rating methodology does not require any applicant to have a 'relationship' with any nominated reviewer. There is no requirement of social skills – it only requires applicants to nominate experts who work in their field of endeavour and who are knowledgeable enough to judge the quality and impact of their research. It is not a matter of 'having connections'.

The argument that the NRF methodology is 'elitist' calls for scrutiny. If scholars enter a system of evaluation, are subject to the same system of evaluation and do not end up being evaluated as equal, is that a sign of elitism? Do *all* athletes participating in the Olympic Games expect to make good on their desire to win? Do *all* football teams competing in the FIFA World Cup expect to share the trophy? Do all students expect to receive the same marks for an examination? Is it elitist that a student receives a degree *cum laude*? Clearly not. Any sort of evaluation, by definition, implies differentiation; but does differentiation necessarily imply elitism? Callaghan's argument that the differentiation induced by the NRF rating methodology is elitist is simply not credible.

Callaghan's proposals and suggestions

Callaghan's proposals and suggestions for an alternative rating mechanism are, by their very nature, contradictory. Whatever he proposes will still be flawed (his term) by the use of power, 'discrimination' between those who do better than others and the inevitable creation of hierarchies. Whatever the rating system in place, some applicants will do better than others. No matter what he suggests, his preferred system would be as susceptible to the same criticism as the current system. Furthermore, from a practical execution point of view, some of his suggestions are simply unworkable.

To suggest that the so-called bibliometrics (h-indices in particular) are more reliable and valid because he believes that they are more objective must be questioned. Ultimately, bibliometrics also rely on peer review and human judgement. For instance, it cannot be denied that some journal reviewers explicitly favour or reject certain methodologies. It is not uncommon to find reviewers who openly state that they will not recommend papers for publication that are based on, say, Bayesian statistics, or panel data or PLS analyses. Callaghan at least acknowledges the impact of 'paradigm beliefs'; so, by implication, he acknowledges that perfect objectivity is not possible.

In any case, to get the 'wide stakeholder consensus' among scholars that he calls for in the Management Sciences on the ideal parameters seems implausible. Questions proliferate: Under whose auspices would this 'wide stakeholder consensus' be sought? Who are the stakeholders? Who will represent the stakeholders? Which indices should be included in the assessment – all of them? Should different weights be assigned to different h-indices? Should the same combination of h-indices be used for all sub-disciplines? Any attempt to introduce bibliometrics as the basis for an alternative rating system would falter and become bogged down during the first roundtable meeting, not to mention reaching consensus across disciplines. So the 'wide stakeholder consensus' he suggests is but a pipe dream. Another problem with the bibliometrics suggestion is that a sizable h-index is built up over many years. Its use will thus favour older, well-established researchers and prejudice the younger, emerging researchers on whose behalf Callaghan claims he speaks. Having said that, there is nothing in the current NRF methodology that prevents applicants from including their h-indices in an application.

Callaghan's blanket assumption that the volume of citations provides an indication of a scholar's impact on a research domain is based on the supposition that the impact was necessarily 'positive' and that this scholarship has always made a significant contribution to the chosen domain. This is simply not true. The work of many scholars is cited for 'negative' reasons: methodologies that are flawed, results that are questionable, interpretations that are poor or not justified, and the like. In other words, some authors are widely cited because their scholarship is dubious; so they rise in the citation indices for the wrong reasons.

The idea that the review process of a scholarly journal is superior to the NRF methodology because it is 'objective' or 'unbiased' because it is anonymous is also not beyond reproach. In a small academic community such as ours in South Africa, and with very few journals in which to publish, complete anonymity cannot be guaranteed. Some authors, research units and departments specialise in certain academic domains and have developed a market reputation for their specialisation. Some universities and research institutions may have specialised facilities no other institution has and only they can publish papers on particular topics. A quick Google search is not the exclusive domain of NRF reviewers. Concealing such associations may be impossible under every circumstance.

Furthermore, it is not uncommon even for international journals to ask authors to nominate potential reviewers for their papers, which also produces a degree of subjectivity bias or collegial patronage. So for Callaghan to argue that subjectivity bias is exclusive to the NRF rating methodology is simply not valid.

Callaghan contends that 'harm' has been done by the NRF methodology without explaining what that 'harm' involved or who was harmed. More importantly, he provides no evidence whatsoever of the 'harm' caused. In the absence of any scientific evidence being provided, one must assume that the criticism is no more than his own unsubstantiated opinion.

Are bibliometrics what they are cracked up to be?

The use of bibliometrics to evaluate the research impact of individual researchers and as performance indicators of research institutions is not without its critics. Hicks et al.[3] believe that the use of research metrics has become too widespread to ignore its negative consequences. This criticism can be divided into two broad categories. Firstly is that, by using bibliometrics, scholars cede their right to peer review to data, and more particularly to the data of private sector vendors who do not have the capabilities to use it appropriately.[3,4] In the words of Laloë and Mosseri[5]: '...the implementation often seems to arise from a loss of critical and rational mind'.

From a methodological point of view, Weingart[4], in assessing the introduction of bibliometrics in the UK to assess the research performance of research institutions, concluded: 'Bibliometric measures, although quantitative and therefore seemingly objective, appeared to be theoretically unfounded, empirically crude, and dependent on data that were known to be imprecise.' The Web of Science database, on which many evaluations have been based, was not initially constructed as a data source for

research performance evaluation but '...as a literature databank designed to identify uses of knowledge and networks of researchers...'[4]. Since the commercialisation of the Web of Science database, the previous costly measures to clean the data have been stopped, leading Weingart to caution against '...the uncritical embrace of bibliometric measures'[4].

Callaghan's argument that chasing a favourable NRF rating (he refers to 'gaming' – that is, 'research is conducted for the express purpose of meeting the goals of a system') leads to the publication of poor research that is hardly read by anyone does not consider that the same reactive behaviour may be produced by bibliometrics.[5] Referring to impact factors, Lawrence points out: 'It has evolved to become an end in itself – the driving force for scientists to improve their reputation…'.[6]

The validity of bibliometrics such as h-indices depends heavily on capturing the correct data, cleaning the data, skilled employees who prepare the data, proper data-processing and proper categorisation.[7,8] The use of h-indices has been found wanting on these criteria by several authors. For instance, bibliometrics have been criticised for data processing errors[4,7,8], that they collect data from only certain journals[4], that they ignore research published in books[5] and that no distinction is drawn between the credit received by the author of a sole-authored paper and the tenth author on a multi-authored paper[5].

Ironically, a further criticism made by Weingart of the validity of bibliometric measures is the problem of the accurate categorisation of the collected data: 'In particular, *interdisciplinary* fields present a problem to proper categorization'[4]. Clearly, poor delineation between disciplines leads to mistakes in citation counts.[4]

It is clear that, in the eyes of many, the relationship between bibliometrics indices such as an h-index on the one hand, and research quality on the other hand, is tenuous to say the very least, and even unscientific.[5] More specifically, such indices are clearly not the 'strictly objective evaluation' Callaghan seems to believe they are.

Multidisciplinary and transdisciplinary research

Callaghan argues that NRF rating applicants who are involved in multidisciplinary and transdisciplinary work are penalised ('discriminated against') by the current methodology. Again, no evidence is provided to support the contention. In the case of multidisciplinary and/or transdisciplinary applications, an applicant has several options. One is to clearly position the application by focusing on the areas in which the applicant believes the most significant impact was evident. The application form has a field where the candidate can indicate whether the research is of a multidisciplinary nature, and can identify the focus of the research during the last 8 years. The second consideration is to choose the reviewers carefully to ensure that enough expertise is available among them to encompass the ambit of the researcher's research.

Callaghan tries to make the point that researchers who have 'changed trajectory' during a later phase of their research careers are prejudiced by the NRF rating methodology. This accusation is difficult to accept if one considers that the review period is the last 8 years. The narrative section of the application form should be used to point out this change of direction to the reviewers, and what impact it has had on their research portfolio. Again, the work is reviewed by peers for its significance and impact, and not for the volume of output, as is commonly believed.

In any event, the narrative part of the rating application form offers ample opportunity for the applicant to explain the potential uniqueness of the research to reviewers.

Also disputable is Callaghan's contention that monodisciplinary research is of little value and is hardly read, while multidisciplinary and transdisciplinary work (such as his own) is more valuable in the service of solving real-world problems. This is obviously a heresy.

Callaghan's contention that 'this rating system [is] acting as a catalyst to create a culture of competition which differentiates *publicly* between "winners" and "losers"' (p. 2) is somewhat misleading. Although the names of those who receive a rating are published on the NRF website, only the broad rating category (e.g. B) appears on the NRF website. It is not clear how the names of those whose rating applications were unsuccessful are publicly identified.

A more balanced assessment of the NRF rating methodology

Without question, the outcomes of the NRF rating methodology are based on the perceptions, attitudes and subsequent judgements of the people involved: the Specialist Committee, the Assessor and the Chairperson. These evaluators are involved in every single rating decision. The same applies to the Appeal Committee, which handles disputes and appeals.

The NRF methodology is:

- not anonymous – it is based on past research output
- voluntary
- a peer-review process
- qualitative and subjective in nature
- as valid a measure of any researcher's peer-reviewed assessment of past research output as is humanly possible

Neither the developers nor the NRF has ever tried to pretend that it is anything else.

Its purpose is to evaluate the significance and impact of a scholar's sustained work over an extended period of time. Unlike a single paper, it evaluates the quality and impact of a 'body' of output and its significance. And, unfortunately, it is impossible to evaluate past (published) work without identifying the author. As Tijsen et al.[9] point out, excellence is by definition a matter of the *ex ante* assessment or the *ex post* evaluation of research performance.

If universities use the NRF rating as a measure of research progress and the recognition of academic standing across disciplines, it means that the system is accepted as reasonably reliable and valid. Fedderke's empirical results confirm the validity of the NRF rating methodology by stating: 'Scholars with higher NRF ratings record higher performance on average against the objective measures of absolute output and the impact of their research, than scholars at lower rating'[10]. These results obviously confirm the broad validity of the rating methodology, and seem to contradict Callaghan's point about inconsistency 'across individuals'.

Callaghan's assessment of the NRF methodology is insistently negative. A more balanced assessment would have uncovered some positives as well. Against this background, one must note that the per capita output at South African universities has increased from 0.51 in 2006 to 0.88 in 2015.[2] I argue that the NRF rating system and its associated incentives have played a significant role in this substantial improvement. The number of research publications produced by South African universities has grown from 5540 in 1994 to 15 542 in 2014 – a threefold increase in a decade (information supplied by the Department of Higher Education and Training). Again, I argue that the NRF rating system and its associated incentives have played a significant role in this substantial improvement.

Callaghan's own interaction with the NRF

To reach a balanced assessment of Callaghan's criticism of the NRF rating methodology, his readers would need to know whether he himself has applied for a rating, given his strong views of how flawed the system is. If he *has* applied, one would reasonably like to know why he applied, given his abhorrence of an unfair, discriminatory system that does no good and only causes harm.

If he did *not* apply, presumably for the reasons cited in his paper, this would add credibility to his critique. Either way, for someone who claims the moral high ground by 'having the courage to speak to power', he ought to have declared his own interaction (or otherwise) with the NRF rating methodology to all readers of his paper. This information is essential in helping readers to judge Callaghan's contribution to the debate on the validity of the NRF rating system.

In summary

Callaghan challenges the validity of the NRF rating methodology, but his own criticism lacks the validity he calls for as a result of his poor understanding of the entire rating process. A more thoughtful analysis of the rating methodology would have acknowledged its purpose and *raison d'être*, and would have revealed the considered checks-and-balances that are in place to minimise – and even avoid – the subjectivity bias to which Callaghan attributes it.

His 'analysis' is flawed because of his poor understanding of the process; he makes sweeping statements (about harm done, subjectivity, gamification, research that is not read, unfair discrimination) without offering any evidence to support these claims.

His recommendations, which suggest that the volume of citations and the peer reviews of journals will lead to bibliometric indices that are unbiased, valid measures of researchers' impact on a research domain, are not without their own limitations. Furthermore, his recommendations will not overcome his own criticism of elitism, the supposed favouring of the 'Ivy League' universities and his view that the process leads to unfair discrimination.

Callaghan contradicts himself by suggesting an alternative rating system. He advocates alternative, better measures and procedures to 'rate' researchers; but any methodology is 'elitist' by definition, and leads to a hierarchy founded on forms of differentiation. If his suggestions were to be implemented, the productive 'elites' would simply get to the top via a different route. What he regards as undue power will simply shift from 12 NRF-commissioned evaluators to (usually) two reviewers and an editor and finally to the questionable database of a commercial vendor. The evaluative hierarchy would not disappear: it might even be constrained and consolidated into an even tighter cluster of authority.

If one argues that fair discrimination between different levels of academic performance is unacceptable, then the evaluative judgement of scholarly performance is, by its very nature, the unethical use of 'power', and that the resultant hierarchies are unacceptable (because we should all be equal), then Callaghan's proposals and suggestions are not the answer. In fact, there is clearly only one answer. Extending Callaghan's arguments on elitism, power (ab)use and hierarchies, *any* method or form of 'rating' would be immoral.

I want to argue that the NRF rating system has been a huge success. It has been widely accepted by the academic community. The number of applications received by the NRF since 1984 is 6744; currently 3889 researchers hold a valid rating, of whom 196 are rated researchers in the domains of economics, management, accountancy and public administration (information supplied by the NRF). From anecdotal evidence – and many personal conversations – I know that an NRF rating has proved aspirational to many and that the prospect of a (higher) rating has motivated many researchers, both at universities and at other research institutions.

While we should all strive to find a valid and reliable measure of the standing of a researcher, it is beyond question that such evaluations should be fair, balanced and uncontaminated by personal considerations and emotions.

Finally, a rating should not be construed as a barrier: it should be seen as an objective means to demonstrate your progress as a researcher; and I want to encourage all scholars, especially young academics and researchers, to take up the challenge rather than to wallow in a pit of complaining and blaming others for 'injustices'.

Future debate

Callaghan calls for 'further research and discussions' on this topic. Some of his criticism of the NRF methodology clearly raises a number of further questions:

1. Can a voluntary system violate academic freedom?
2. Can only multidisciplinary and transdisciplinary research be to the benefit of societal stakeholders?
3. Is monodisciplinary research by definition inferior, leading to 'wasteful publications'?
4. Why are journal reviewers 'knowledgeable peers' but NRF reviewers are not?
5. Will bibliometrics demonstrate the benefit of research to societal research?

Addressing these contentious arguments in Callaghan's review could form the basis for future debate.

References

1. Callaghan C. A review of South Africa's National Research Foundation's rating methodology from a social science perspective. S Afr J Sci. 2018;114(3/4), Art. #2017-0344, 7 pages. https://doi.org/10.17159/sajs.2018/20170344
2. Centre for Research on Evaluation, Science and Technology, Stellenbosch University. SA Knowledgebase [database]. No date [cited 2018 Jul 02].
3. Hicks D, Wouters P, Waltman L, De Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520(7548):429–431. http://dx.doi.org/10.1038/520429a
4. Weingart P. Impact of bibliometrics upon the science system: Inadvertent consequences? Scientometrics. 2005;62(1):117–131. https://doi.org/10.1007/s11192-005-0007-7
5. Laloë F, Mosseri R. Bibliometric evaluation of individual researchers: Not even right… Not even wrong. Europhys News. 2009;40(5):26–29. https://doi.org/10.1051/epn/2009704
6. Lawrence PA. Rank injustice. Nature. 2003;415:835–836. https://doi.org/10.1038/415835a
7. Braun T, Glanzel W, Schubert A. How balanced is the Science Citation Index's journal coverage? A preliminary overview of macrolevel statistical data. In: Cronin B, Atkins HB, editors. WoK: A Festchrift in Honor of Eugene Garfield. Medford, NJ: Information Today Inc. & The American Society for Information Science; 2000. p. 251–277.
8. Zitt M, Ramanana-Rahary S, Bassecoulard E. Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. Scientometrics. 2003;56(2):259–282. https://doi.org/10.1023/A:1021923329277
9. Tijssen RJW, Visser MS, Van Leeuwen TN. Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference. Scientometrics. 2002;54(3):381–397. https://doi.org/10.1023/A:1016082432660
10. Fedderke J. The objectivity of National Research Foundation peer review based ratings in South Africa. ERSA Working Paper 300 [document on the Internet]. c2012 [cited 2018 Jul 02]. Available from: https://econrsa.org/system/files/publications/working_papers/wp300.pdf