Scientometrics (2013) 97:177–206
DOI 10.1007/s11192-013-0981-0

The objectivity of national research foundation peer review in South Africa assessed against bibliometric indexes

J. W. Fedderke

Received: 17 July 2012 / Published online: 2 March 2013
© Akadémiai Kiadó, Budapest, Hungary 2013

Abstract This paper examines the strength of association between the outcomes of National Research Foundation (NRF) peer review based rating mechanisms, and a range of objective measures of performance of researchers. The analysis is conducted on 1932 scholars that have received an NRF rating or an NRF research chair. We find that on average scholars with higher NRF ratings record higher performance against research output and impact metrics. However, we also record anomalies in the probabilities of different NRF ratings when assessed against bibliometric performance measures, and record a disproportionately large incidence of scholars with high peer-review based ratings with low levels of recorded research output and impact. Moreover, we find strong cross-disciplinary differences in terms of the impact that objective levels of performance have on the probability of achieving different NRF ratings. Finally, we report evidence that NRF peer review is less likely to reward multi-authored research output than single-authored output. Claims of a lack of bias in NRF peer review are thus difficult to sustain.

Keywords Peer review · Bibliometric measures

Introduction

Research has come to rely heavily on public funding. Budget constraints in turn impose the need for funding bodies to make hard choices on which research initiatives to favour, with inevitable disappointment and even disgruntlement on the part of at least some researchers dependent on such funding. The consequence is often controversy. Given that allocative decisions can carry substantial bearing on the career prospects of scholars, it is not surprising that the decision making of funding bodies is often surrounded by controversy and allegations of bias and inconsistency.

J. W. Fedderke (✉)
Pennsylvania State University, Economic Research Southern Africa and University of the Witwatersrand, Cape Town, South Africa
e-mail: jwf15@psu.edu

Funding bodies generally rely on peer review to reach decisions. Justification of peer review as the principal mechanism of quality control in scholarship, rests on claims that despite imperfections it issues in better results than any alternative, that it improves the quality of research results, avoids bias against young scholars (in contrast to output-based measures), provides feedback and positive motivation to scholars, selects reliable research findings from the large body of available research output, and is viewed positively by a majority of scholars. However, there are also significant criticisms of peer review processes. These range from claims that it has poor *reliability* (in the sense that there is rarely agreement among reviewers), poor *fairness* (recommendations are subject to a range of biases), lacks *predictive ability* (identification of the most significant contributions is weak), is *inefficient* (it is costly, time consuming, often inhibits the most innovative, unconventional and new perspectives), and that it forces researchers to follow reviews slavishly with limited opportunity of retort. Moreover, since peer review processes are generally conducted under conditions of anonymity and often in a closed review process, claims that their peer review processes issue in reliable outcomes, are inherently difficult to verify due to the lack of transparency of the process.[1]

An alternative to peer review is offered by the use of bibliometric indexes. Proponents of this approach advance evidence suggesting that bibliometric evaluation is superior to peer review in terms of robustness, validity, functionality, and the time and financial cost impact of the methodology, particularly in the context of national research assessments.[2] Since the advantages claimed for the bibliometric approach over peer review are particularly marked for comparative research evaluations, a number of public research assessment processes have moved to either a reliance on both peer review and bibliometrics, or a pure bibliometrics based approach.[3]

This paper explores whether a range of research funding body decisions based on peer review, corresponds with a range of objective measures of scholarly performance. It does so on the basis of South African data. The paper has the advantage of using verifiable data on the outcome of the national South African peer review process for 1932 scholars across the full range of scholarly disciplines represented in South Africa, which is combined with a wide set of associated bibliometric indexes of research performance for each scholar.

The principal source of funding for scholarly research in South Africa is the National Research Foundation (henceforth NRF). Critical to the funding mechanisms of the NRF are a range of peer review mechanisms. This peer review issues in a ranking of all scholars that the NRF claims is a reflection of research standing and impact, and may be associated with the award of prespecified funding grants.[4] Since the NRF makes public the list of rated

[1] The literature assessing the validity of peer review is vast. For a recent comprehensive review see Bornmann (2011). Earlier reviews of peer review in the context of grant evaluation can be found in Demicheli and Pietrantonj (2007) and Wessely (1998). For some of the evidence in support of peer review (not necessarily in relation to grant evaluation) see Goodman et al. (1994), Pierie et al. (1996), Bedeian (2003) and Shatz (2004). For critics see Eysenck and Eysenck (1992) and Frey (2003).

[2] See Abramo and D'Angelo (2011) and the discussion of the literature and evidence in Abramo et al. (2011). Again, the literature on comparing peer review and bibliometrics is large—but see for instance Horrrobin (1990), Moed (2002), Moxam and Anderson (1992), Pendlebury (2009) and van Raan (2005).

[3] The 2012 Italian evaluation exercise and 2014 United Kingdom research assessments rely on both peer review and bibliometric indicators. The Australian research excellence exercise relies purely on bibliometrics. See the discussion in Abramo et al. (2011).

[4] The impact on scholars can be substantial. While the average funding available per researcher in South Africa under NRF programs is approximately ZAR 20,000 per annum, researchers granted South African Research Chairs receive an annual budget allowance of ZAR 3 million.

scholars and holders of research chairs, we are able to match the NRF rating with data on the absolute level of research output and the citations-based impact of research output of scholars, across a wide range of alternative measures, in order to assess whether the funding body's assessment reflects the objective performance of researchers.

Findings are nuanced. The objective performance measures of scholars are found to be positively associated with the prestige-level of the NRF rating they obtain. In addition, absolute publications output as well as impact are found to raise both the probability of any specific NRF rating, as well as the level of the NRF rating of a scholar. Such findings lend support to the NRF peer review process, in the sense that the resultant ratings appear to be associated significantly with objective performance.

However, the evidence also suggests that even at high levels of objective performance measures (in both output and impact terms), lower-ranked ratings may be more probable than higher-ranked ratings, and top-ranked scholars in the NRF system have objective performance measures that correspond to minimum levels of performance of scholars at low-ranked NRF ratings. What is more, the data also suggests that there exist quite distinct probabilities of achieving the various NRF ratings categories across disciplinary groupings. Such disciplinary differences extend to the impact that improvements in the objective performance measures have on the probability of achieving any of the specific NRF ratings—and the differences are substantial.

The implication is that NRF claims that its peer review mechanisms are based on an objective consideration of research impact are incompletely supported by the data. While more highly rated scientists do report higher performance on objective measures of output and impact on average, inconsistencies across NRF ratings and across disciplines are such that they suggest the presence of non-negligible bias.

The paper proceeds as follows. In "The ratings indexes"section we present the measures and data sources employed for this study. Section "The data" presents the sources of the data that we employ for the study. In "Results" section we present our results, and "Conclusions and evaluation" section concludes.

The ratings indexes

This paper uses a number of measures of research standing. These comprise both the peer review measures of the NRF, and objective measures based on absolute research output and impact as measured by a range of bibliometric indicators.

The objective measures of scientific standing

There now exists a very wide range of bibliometric measures of research performance.[5]

In this study we employ measures that capture three distinct dimensions of scholarly output: the absolute level of output, the impact of output, and a set of composite measures that combine the level and impact of output.

The first set of objective measures of scholarly standing employed for this study, account for the absolute level of output, and are based on publications counts. We employ two distinct measures. The first is the total number of papers attributed to an author. The second corrects for multiple authorships, by recording the average number of papers per author (by Papers/Author).

[5] For a comprehensive list of the metrics, their construction and characteristics, see Rehn et al. (2007).

To account for the impact of research, we employ three raw citations-based measures of impact. The citation count records the total number of citations attributed to an author. Since the total citations count provides an implicit advantage to older scholars (they have had a longer period in which to accumulate citations), the second measure records the average number of citations per paper attributed to an author, while the third measure records the average number of citations per author.

The literature on measuring scholarly impact has also proposed a set of composite measures, that combine absolute output and citation counts. In the case of Hirsch's *h*-index, the objective is explicitly to provide a single-number metric of an academic's impact, combining both the number of publications with a measure of impact as indicated by citations.[6] In order to achieve a high *h*-index an author requires both a high number of publications, and a high number of citations.

A number of modifications have been proposed to the *h*-index, in order to correct for a number of potential weaknesses, limitations or biases to the original index.[7]

One modification to the *h*-index corrects for the distribution of publications and citations over *time*. This is given by the contemporary h_c-index, which aims to improve on the *h*-index by giving more weight to recent articles, thus rewarding academics who maintain a steady level of activity—see Sidiropoulos et al. (2006). In the case of the individual (original) *hI*-index, and individual (PoP variation) *hI*-norm, and the multi-authored h_m-index,[8] the correction is for differences in patterns of co-authorship and publication rates across disciplines. Egghe's *g*-index provides a correction to account for particularly influential contributions, by giving more weight to highly-cited articles—see Egghe (2006). Finally, Zhang's *e*-index differentiates between scientists with similar *h*-indices but different citation patterns—see Zhang (2009).[9]

An approach that offers an alternative to the combination of absolute output count and citations weight, is to adjust citations measures directly for a range of factors, most commonly the age of the research. We use three different versions of such measures. The *AWCR* measure adjusts citations for the age of the associated paper, the *AWCRpA* measure further adjusts the *AWCR* measure to account for the number of authors for each paper, while the *AW*-index adjusts the *AWCR* to allow comparison with the *h*-index—see Jin (2007).

We thus have a number of measures of research standing, which can be viewed as providing complementary evaluations of impact, compensating for the various strengths and weaknesses of the indexes.

Nevertheless, a number of weaknesses of bibliometric measures remain. Notable amongst these is the fact that appropriate measures of standing for junior academics are inherently difficult to generate. In some disciplines, the lead time to publication and hence citation is substantial. Further, bibliometric measures in general do not explicitly take into account the standing of the journals the work appears in (though more highly rated journals

[6] See Hirsch (2005). For a discussion of the properties of the *h*-index, see for example Egghe and Rousseau (2006), Glänzel (2006), Bornmann and Daniel (2005), Cronin and Meho (2006), and Van Raan (2006).

[7] In what follows only indexes actually computed for the present study are discussed. There are certainly other indexes—for instance see the discussion in Rehn et al. (2007), as well as the overview in Bornmann and Daniel (2007), and Bornmann et al. (2008, 2009a). Reasons for our choice are outlined in the discussion below.

[8] See Batista et al. (2006) and Schreiber (2008). Batista et al. (2006) show that cross-disciplinary variability under the *hI*-index is significantly reduced.

[9] See also the discussion in Thor and Bornmann (2011).

are cited more). In some instances, high citations attach to papers that contain a famous/notorious error (though this has been found to be a low frequency occurrence).

Against these caveats concerning the objective measures of scientific standing, their advantage is that they are based on objective measures of output and impact, that are transparent, verifiable, and subject to accountability.

The NRF measures of scientific standing

Research funding agencies are generally characterized as belonging to one of three distinct models. Intra-academic models rely on peer review by researchers that are themselves active in a designated area of expertise. Top-down models by contrast follow funding allocation mechanisms that are directed in terms of some overarching social and/or political objectives. Most recently, the literature has suggested that at least in some advanced industrial countries funding is allocated in terms of a "triple helix," an interaction between researchers, business and wider economic interests, and political interests, with a strong focus particularly on commercial applications.[10]

Within this characterization, the NRF follows the intra-academic model. The NRF relies on peer review, without cognizance of any objective measures of research impact. In this it differs from a number of international funding bodies,[11] and does not yet reflect the trend of responding to budgetary pressures for improvement in accountability and efficiency by increasing reliance on objective performance measures.

The NRF of South Africa provides two sets of measures of scholarly standing: the NRF ratings process, and the NRF research chair initiative.

The NRF peer reviewed science rating mechanism

The NRF conducts an evaluation of researchers that is based on peer review to benchmark researcher performance and to assist the NRF in the evaluation of its provision of research grants.

Scholars apply for an NRF rating. The application is submitted to subject-specific Specialist Committees who identify at least six, and no more than ten peer reviewers. Peers are asked to evaluate the applicant on the basis of the quality of research-based outputs over the last seven years as well as the impact of the applicant's work, and an estimation of the applicant's standing as a researcher. On the basis of the peer reports, Specialist Committees are asked to assess the standing of applicants amongst their peers and recommend a rating. No objective measures of the absolute magnitude of research output or impact is formally employed in the rating process.

Ratings can fall into a range of categories. An A-rating is held to identify researchers who are leading international scholars in their field. The B-rating identifies researchers with considerable international recognition. The C-rating identifies established researchers

[10] See the discussion in Benner and Sandström (2000), and the introduction of the triple helix concept in Leydesdorff and Etzkovitz (1996). Adler et al. (2009) document some of the associated complexities of managing research funding agencies in this type of context.

[11] We have already noted the UK, Italian and Ausralian cases. See also the discussion in Debackere and Glänzel (2003) on the Belgian funding bodies, and García-Aracil et al. (2006) sevaluations of the Valencian rating bodies to scrutiny against objective measures of performance.

with a sustained recent record of productivity. A Y-rating applies to young researchers, while an L-rating applies to researchers who faced historical discrimination.[12]

Since the peer review based ratings rely on the judgement of peers, they are irreducibly linked to the judgement of other scholars. This opens the mechanism to a number of potential sources of bias. Given that members of the Specialist Committees are themselves scholars that are rated by the NRF mechanism, there is the potential of selection bias in assessments: the favouring of old established areas of research, of well established research institutions, or of established informal research networks, while the preferences of a relatively small number of reviewers come to carry a disproportionate weight.[13] Second, given the relatively narrow disciplinary bases of the independent assessors who monitor the consistency of Specialist Committees (there are only four at any given time that cover all academic disciplines), the rigour of cross-disciplinary consistency is open to question. Third, since only the last 7 years of research output are evaluated, the price paid for immediacy of the rating is a failure to reflect the impact of life-time contributions of scholars and long-term cumulative research projects. Fourth, since the home institution of the researcher needs to support any applicant, this may provide a bias against non-conventional research loci. Finally, the process is relatively untransparent and unverifiable. Since both the peer reports and the deliberations of the Specialist Committees are confidential, the grounds for the ratings reported, the rigour and consistency of assessment cannot be assessed for objectivity and accuracy within, let alone across Specialist Committees.

The advantages to the system are that it can correct for bias against new and young scholars, excessive reliance on single high impact items of output or many low impact articles, and it can reflect the standing of the journals in which the output of scholars appears.

The NRF research chairs

The NRF also operates the South African Research Chair initiative. Its objective is described as making South Africa competitive in the international knowledge economy based on its existing and potential strengths. The core objective is specified as increasing the number of world class researchers in South Africa, and indicates a set of selection criteria that emphasizes world-class research output.

However, it is notable that correspondence between the NRF research rating and the NRF research chair evaluation mechanisms is poor. At the date of data collection, the NRF listed 80 chairs.[14] Of the 80 identifiable chairs, only 71 % are held by researchers that are even rated under the NRF peer review system, leaving 29 % of the NRF chair holders unrated. Of the rated NRF chairs, only 10 % held an A-rating, 36 % a B-rating, 23 % a C-rating, and 3 % a Y-rating.

[12] The L-rating has been discontinued as of 2010. Candidates who were eligible in this category included: black researchers, female researchers, those employed in a higher education institution that lacked a research environment and those who were previously established as researchers and have returned to a research environment.

[13] For instance, of the members of the Specialist Committees listed on the NRF website at the time of data collection, 50 % were from the University of Cape Town, the University of the Witwatersrand and the University of Pretoria; if the University of Stellenbosch and KwaZulu-Natal were added, the proportion rises to 71.43 %. By contrast, 4.76 % come from historically disadvantaged institutions.

[14] In 2012 an additional set of chairs were announced. These were not included in the analysis.

The poor correspondence between the two NRF evaluation mechanisms explains the separate treatment of NRF chairs in our analysis.

The data

For this study we employed three sources of data.

The first was derived directly from the published list of rated scholars and research chairs of the NRF.

For the range of objective measures of scholarly standing, our data was obtained from Harzing's Publish or Perish software. Every scholar reported as rated by the NRF or as holding a research chair, was entered into the Harzing software, in order to generate the range of objective citation count based measures of scholarly standing for a specified year (2009). A total of 1932 rated scholars were subjected to evaluation, and the associated rating metrics recorded. Since the underlying Google Scholar search engine identifies large numbers of references that are either spuriously attributed to specific authors, or that identify output that constitutes spurious research material, a substantial number of research hours were devoted to cleaning each of the 1932 individual records. This was assisted by the South African location of the researchers, and the resultant ability to cross reference individual biographical information.[15]

In generating the set of formal citations-based indices of scholarly standing, we specified both the surname and initials of the scholar in question. We then worked through the generated list of citations, in order to eliminate any references that were not attributable to the scholar in our ratings data base. This may generate a downward bias in the recorded performance measure. While relatively benign if the bias is consistent across scholars, it may prove more pronounced for authors that were part of multi-author teams and whose surname has a relatively low alphanumeric rank, since the Publish or Perish software may truncate long author lists which are presented in alphabetic order.

Use of an electronic search engine to generate citation counts raises a number of specific issues relating to measurement error, that should be noted at the outset.

The Harzing software relies on Google Scholar to generate the citation data and formal rating scores. The literature has generated some debate on the robustness of citation counts data based on Google Scholar relative to a range of alternatives (Scopus, ISI Web of Science). One set of studies has questioned the reliability of Google Scholar, particularly on the grounds of attribution of publications to phantom authors, inclusion of non-scholarly publications,[16] exclusion of some important scholarly journals, uneven disciplinary coverage,[17] less comprehensive coverage of publications prior to 1990,[18] and inconsistent

[15] The importance of ensuring the accuracy of author attribution is emphasised throughout the literature on bibliometrics—irrespective of search engine employed. Hence the substantial time spent on the underlying data for this study. See particularly the discussion in Pendlebury (2008, 2009).

[16] Though Vaughan and Shaw (2008) and Harzing (2007–2008) suggest this to be a relatively low source of error.

[17] Bosman et al. (2006) found Google Scholar, Web of Science and Scopus coverage generally comparable. Nonetheless they report disciplinary variations, corroborated by Kousha and Thelwall (2007, 2008) who find Google Scholar underreports the natural sciences, and Bar-Ilan (2008) who finds variation within the natural sciences.

[18] See Belew (2005) and Meho and Yang (2007).

Table 1 Comparison of ISI Web of Science and Google Scholar bibliometric results

	Papers		Citations per paper		h-Index		
	Less than or equal	Greater	Less than or equal	Greater	Less than or equal	Greater	Total
Total sample							
Proportion of scholars (%)	68	32	42	58	51	49	100
Breakdown by discipline grouping							
Biological	52	48	23	76	29	71	
Business	94	6	65	35	87	13	
Chemical	53	46	22	78	30	70	
Engineering	66	34	41	59	48	52	
Medical	51	50	29	71	33	67	
Physical	65	35	28	72	38	62	
Social	90	9	73	27	83	17	

Table reports ISI Web of Science relative to Google Scholar

accuracy.[19] However, a further set of studies suggests that Google Scholar is more robust and accurate than the ISI Web of Science database. Reasons cited are that the Web of Science database does not include citations to scholarly output that has even small mistakes in its referencing and is subject to more citation noise; it provides overrepresentation to English language and US and UK based journals; it is biased toward citations to journal articles (as opposed to books, book chapters, working papers, reports, conference papers, etc.); it significantly restricts citations to non-ISI database journals; it underreports citations in disciplines with long delays to publication; it underreports citations in general; it is sensitive to institutional subscriptions.[20] Finally, the Web of Science and Google Scholar also share some common problems, such as that names with diacritics (eg, ö, é), apostrophes (eg. O'Connell) or typesetting ligatures (eg. ff, fi, fl) cause difficulties to both search engines.

So the literature points in multiple directions on the relative reliability of the alternative search engines. Does it matter for our exercise? To assess this question we took a drawing of 617 of the total set of peer rated 1932 researchers (a 32 % sample), and derived a third data set on total publication, citations per paper, and h-index bibliometric indices from the ISI Web of Science database. As for the Google Scholar based exercise, extensive attention to the removal of spurious references and authors was mandatory. The results of the comparison are reported in Table 1.

We find that for the majority of the 617 researcher sample with results under both Web of Science and Google Scholar searches, ISI records fewer publications (for 68 % of the

[19] See for instance the general discussion in Bornmann et al. (2009b), Flagas et al. (2008), García-Pérez (2010), Gray et al. (2012), and Jasco (2010).

[20] See the discussion in Archambault and Gagné (2004), Belew (2005), Derrick et al. (2010), García-Pérez (2010), Harzing (2007–2008, 2008), Kulkarni et al. (2009), Meho and Yang (2007), and Roediger (2006). While Jacsó (2005, 2006a, 2006b) reports that the social sciences and Humanities are underreported under Google Scholar, larger-scale studies reverse this finding—see Bosman et al. (2006) and Kousha and Thelwall (2007). Further evidence comes from Nisonger (2004) and Butler (2006). Testa (2004) reports that ISI itself estimates that of the 2000 new journals reviewed annually only 10–12 % are selected to be included in the Web of Science.

researchers), more citations per publication (for 58 % of the researchers), while the sample is approximately evenly split between researchers who record lower and higher *h*-indices under ISI than Google Scholar. We also find marked disciplinary differences. While all disciplines record fewer publications under ISI than Google Scholar, the divergence is dramatic in the case of the business and social sciences. It is also the business and social sciences that record dramatically fewer citations per paper, and lower *h*-indices under the ISI citations system than under Google Scholar.

In our sample we therefore find that ISI Web of Science appears to underreport the output and impact of non-natural science based disciplines, and potentially dramatically so.

So how much does this matter for our exercise? This paper is concerned with the assessment of whether the relative ranking of researchers generated by the NRF peer review mechanism is consistent with evidence obtained from objective bibliometric indices. The differences between ISI and Google Scholar will be relevant only to the extent that they issue in a different *ranking* of the scholars under the bibliometric measures from the alternative search engines. Spearman's rank correlation coefficients between the bibliometric measures obtained under the ISI and Google Scholar search engines are 0.84 for the *h*-index, 0.79 to total citations, 0.75 for the number of papers, and 0.63 for citations per paper. The implication is thus that inferences drawn on the relative scholarly standing of researchers under the bibliometric indices, are unlikely to be materially affected by the use of Google Scholar or the ISI measures, with the possible exception of the citations per paper measure. This finding is consistent with other studies that suggest that the Web of Science and Google Scholar produce very similar rankings of academics.[21]

A further concern regarding our data arises from the fact that the researchers in our data are drawn from very diverse disciplines. Evidence from the literature suggests that there is strong cross-disciplinary variation in bibliometric indices.[22] For this reason we also adjusted our *h*-index by discipline-specific weights as suggested by Table 2 of Iglesias and Pecharromán (2007).[23] The Spearman rank correlation coefficient between the raw and the discipline normalized *h*-index is 0.93. In terms of the question of this paper use of the raw or discipline normalized h-index is therefore unlikely to be of material significance. Nonetheless we tested all results for sensitivity to the use of the two measures and we report significant divergence where it arises.

Since the relative ranking of scholars does not appear to be very sensitive to the search engine employed, and since the Google source provides a more comprehensive measure of impact that is likely more inclusive of the social sciences, we therefore proceed under the Google Scholar based bibliometric measures. The impact of disciplinary adjustments on the bibliometric measures are noted where relevant.

Results

We proceed in three steps. First, we present descriptive statistics on the objective performance of scholars under the alternative NRF ratings. Second, we present logit regressions that allow for the derivation of the probability of achieving any given NRF rating, conditional on the objective performance on the range of metrics that we employ for this

[21] See Saad (2006) and Meho and Yang (2007).

[22] See the discussion in Rehn et al. (2007) and particularly Iglesias and Pecharromán (2007).

[23] The weights for the discipinary categories in our study are as follows: biological 0.77; business 1.32; chemical 0.92; engineering 1.7; medical s0.625; physical 1.14; social 1.6.

Table 2 Correlation matrix for output and impact measures

	Papers	Citations	Cites per year	Cites per paper	Cites per author	Papers per author	Authors per paper	h-Index	g-Index	hc-Index	hI-Index	hI-Norm	e-Index	hm-Index	AWCR	AW-Index	AWCpA	Adj. h-index[a]
Papers	1.00																	
Citations	0.44	1.00																
Cites per year	0.38	0.91	1.00															
Cites per paper	0.14	0.25	0.23	1.00														
Cites per author	0.45	0.97	0.86	0.24	1.00													
Papers per author	0.92	0.28	0.22	0.10	0.32	1.00												
Authors per paper	−0.11	−0.003	−0.002	0.23	−0.02	−0.13	1.00											
h-Index	0.64	0.95	0.86	0.23	0.93	0.45	−0.03	1.00										
g-Index	0.65	0.94	0.84	0.24	0.93	0.47	−0.04	0.99	1.00									
hc-Index	0.64	0.94	0.86	0.23	0.91	0.46	−0.04	0.99	0.98	1.00								
hI-Index	0.64	0.92	0.81	0.22	0.94	0.51	−0.06	0.97	0.97	0.96	1.00							
hI-Norm	0.65	0.93	0.82	0.23	0.94	0.50	−0.05	0.98	0.98	0.97	0.99	1.00						
e-Index	0.65	0.92	0.82	0.24	0.92	0.48	−0.04	0.94	0.99	0.97	0.96	0.98	1.00					
hm-Index	0.64	0.95	0.84	0.23	0.94	0.48	−0.04	0.98	0.99	0.98	0.99	0.99	0.97	1.00				
AWCR	0.45	0.98	0.92	0.24	0.94	0.27	0.01	0.93	0.93	0.93	0.90	0.91	0.91	0.93	1.00			
AW-index	0.70	0.90	0.82	0.23	0.87	0.50	−0.03	0.97	0.97	0.98	0.95	0.96	0.97	0.97	0.91	1.00		
AWCpA	0.46	0.97	0.89	0.24	0.98	0.32	−0.01	0.99	0.98	0.91	0.93	0.93	0.91	0.94	0.97	0.90	1.00	
Adj. h-index[a]	0.78	0.77	0.61	0.31	0.81	0.70	0.02	0.84	0.80	0.79	0.82	0.84	0.73	0.84	0.62	0.77	0.73	1.00

[a] Denotes indices adjusted for discipline specific h-index performance metric

Table 3 Means

	Mean						
	Full sample	NRF chairs	A-rated	B-rated	C-rated	Y-rated	L-rated
Papers	77.40	70.18	148.31	80.27	45.50	33.06	25.28
Citations	515.70	679.99	1618.35	511.12	231.13	228.20	110.64
Cites per year	18.08	26.49	39.27	17.10	8.34	9.19	4.06
Cites per paper	5.83	8.03	7.93	6.02	4.34	5.26	2.95
Cites per author	231.91	350.26	664.29	221.89	123.80	82.93	47.11
Papers per author	40.38	33.28	79.35	39.76	25.19	17.23	14.14
Authors per paper	2.55	2.74	2.38	2.49	2.49	2.85	2.44
h-Index	9.01	11.05	15.51	10.13	6.45	5.67	4.09
g-Index	15.69	20.50	24.84	17.27	10.63	10.21	6.04
hc-Index	5.65	7.16	9.34	6.11	4.35	4.64	3.15
hI-Index	3.51	4.22	7.02	4.21	2.72	2.05	1.90
hI-Norm	5.86	7.19	10.96	6.55	4.42	3.70	3.08
e-Index	11.64	15.16	21.71	12.80	8.27	8.43	4.83
hm-Index	5.63	6.41	9.45	6.44	4.18	3.33	2.65
AWCR	52.14	66.10	131.56	56.34	26.60	32.85	13.06
AW-index	5.72	7.11	9.67	6.26	4.12	4.62	2.87
AWCpA	21.01	35.94	58.83	23.07	11.13	11.46	5.46
Adj. h-index[a]	8.23	12.47	17.39	11.32	7.00	5.87	4.48

[a] Denotes indices adjusted for discipline specific h-index performance metric

study. Finally, we consider the possibility that disciplines are distinct in terms of their responsiveness to objective performance measures.

Descriptive statistics

The wide range of alternative metrics designed to capture a scholar's performance fall into absolute output based measures (that count the number of distinct scholarly products of a scholar), citations based measures that attempt to assess the impact of scholarly contributions, and measures that combine output and impact.

Table 2 reports the correlation matrix between the measures of absolute output and the impact of research output of scholars we employ in the study. What is evident is that the range of direct citations-based measures of impact (citations, citations per year, citations per author) and the range of indexes designed to measure impact and absolute levels of output jointly (*h*-index, *g*-index, etc.) are highly correlated, indicating that they carry much the same information. In general, correlations between these measures exceed 0.9. The one exception is the citations per paper measure, which returns correlation coefficients with the other citations-based measures at the 0.2 level. The measures of absolute output (papers, papers per author, AW-index) again are highly correlated amongst one another (correlation coefficients exceeding 0.7). They are also correlated with the citations-based measures, though with an intermediate level of correlation (0.5–0.6).[24] Use of the *h*-index adjusted

[24] Authors per paper reports a negative (though very low) correlation with all but the citations per paper measure—given the near zero level of correlation the inference is that authors per paper does not systematically co-vary with the rest of the output and impact measures.

for discipline-specific norms does not materially change this pattern, though the correlation of the citations measures and the h-index is lower (approximately 0.8).

A consideration of the most simple descriptive statistics from our data in general supports the NRF ranking hierarchy of scholars (in descending order: A-rated scholar : B-rated scholar : C-rated scholar : Y-rated scholar : L-rated scholar). Table 3 reports the mean of the 18 measures of scholarly performance employed by this paper. We do so for the sample as a whole as well as conditional on the rating that scholars have received from the NRF, including the possibility of holding an NRF Chair. What emerges is that mean levels of absolute output, as well as mean levels of impact increase with the level of the NRF rating.[25]

The exception is provided by the NRF Chairs, which consistently rank below A-rated scholars on the output and impact measures, and in a number of the objective output and impact measures below B-rated scholars. Since the NRF claims the chairs to be the flagship programme for world-class research leaders, this outcome stands in tension to official claims.

Nonetheless, the broad NRF claim that its peer review based rating system is a reflection of scholar's productivity and impact, is broadly supported by the descriptive statistics.

Deriving the probability of alternative ratings

To consider the impact of the range of output and impact measures of scholars' research on the probability of obtaining a specific rating under the NRF system of peer review we estimate:

$$J_i = X_i\beta + u_i \tag{1}$$

where

$$J_i = \begin{cases} 1 \text{ if } \exists \text{ an NRF rating of type } J = \{NRFChair, A, B, C, Y, L\}, \text{ with probability } \Pr(Y = 1) = P \\ 0 \text{ if } \nexists \text{ an NRF rating of } type\ J = \{NRFChair, A, B, C, Y, L\}, \text{ with probability } \Pr(Y = 0) = 1 - P \end{cases}$$

with the vector of explanatory variables X_i for each scholar i, provided by the output and performance measures of the study. Estimation results under the logit distribution are reported in Tables 4 and 5, for each of the NRF rating categories.

Note that all of the composite index measures of output and impact (h-index etc) are consistently statistically significant (generally at the 1 % level)—see Table 5. On the raw performance measures, in general it is the raw citations count, citations per year, and papers per author that prove to be statistically significant, though the absolute number of papers is sometimes weakly significant (10 % level for A-rated scholars), or replaces the raw citations count in significance (B-rated scholars)—see Table 4. Finally, for A-rated scholars citations per year, citations per paper and citations per author all are statistically significant.

The estimation results allow us to derive the associated probability density functions of realizing the various NRF ratings, conditional on the range of objective measures of scholarly performance. These densities arguably illustrate both why the NRF views its

[25] The same patterns emerge for the medians of the measures. The second moment of the distribution is generally large across all categories, reflecting a wide range of measured output, and the impact of such output.

Table 4 Logit regression: NRF ratings

	NRF chairs			A-rated scholars			B-rated scholars		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	−3.01*** (0.17)	−3.39*** (0.14)	−3.66 (0.38)	−3.79*** (0.18)	−3.66*** (0.16)	−3.38*** (0.39)	−1.57*** (0.08)	−1.42*** (0.07)	−1.46*** (0.17)
Papers	0.001 (0.001584)			0.002* (0.001)			0.003*** (0.001)		
Citations	0.0003** (0.0001)			0.0005*** (0.0001)			−2.89e−005 (8.31e−005)		
Years	−0.01** (0.01)			0.0003 (0.003)			0.004** (0.001)		
Cites per year		0.02*** (0.005)			0.01*** (0.004)			0.01** (0.003)	
Cites per paper		−0.001 (0.01)			−0.03** (0.02)			0.01 v (0.01)	
Cites per author		−0.0003 (0.0003)			0.001*** (0.0002)			0.0002 (0.0002)	
Papers per author			0.003** (0.002)			0.01*** (0.001)			0.01*** (0.001)
Authors per paper			0.16 (0.14)			−0.12 (0.15)			0.01 (0.06)
N	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932
Number scholars in category	80	80	80	77	77	77	440	440	440

	C-rated scholars			Y-rated scholars			L-rated scholars		
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Constant	0.37*** (0.07)	0.44*** (0.07)	0.71*** (0.15)	−1.28*** (0.11)	−1.76*** (0.09124)	−2.53*** (0.24)	−2.76*** (0.2)	−2.75*** (0.18)	−2.35*** (0.34)
Papers	−0.0017 (0.001)			−0.01*** (0.002)			−0.001 (0.01)		

Table 4 continued

	C-rated scholars			Y-rated scholars			L-rated scholars		
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Citations	−0.0004*** (0.0001)			0.0001 (0.0001)			−0.002** (0.0018)		
Years	0.002 (0.001)			−0.01*** (0.004)			−0.01 (0.01)		
Cites per year		−0.02*** (0.004)			0.03*** (0.01)			−0.04 (0.04)	
Cites per paper		0.004 (0.01)			0.03* (0.01)			−0.05 (0.06)	
Cites per author		−0.0003 (0.0002)			−0.004*** (0.001)			−0.003 (0.002)	
Papers per author			−0.005*** (0.001)			−0.02*** (0.004)			−0.02*** (0.01)
Authors per paper			−0.14** (0.05)			0.44*** (0.08)			−0.28** (0.14)
N	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932
Number scholars in category	1,061	1,061	1,061	255	255	255	61	61	61

Figures in round parentheses denote standard errors

*Denotes statistical significance at the 10 % level; ** denotes statistical significance at the 5 % level; *** denotes statistical significance at the 1 % level

Table 5 Logit regression: NRF ratings

Constant	NRF chairs (1)	A-rated scholars (2)	B-rated scholars (3)	C-rated scholars (4)	Y-rated scholars (5)	L-rated scholars (6)
h-Index	0.05*** (0.01)	0.12*** (0.01)	0.07*** (0.01)	−0.06*** (0.01)	−0.07*** (0.01)	−0.16*** (0.04)
Adj h-index[a]	0.05*** (0.01)	0.09*** (0.01)	0.06*** (0.01)	−0.05*** (0.01)	−0.07*** (0.01)	−0.14*** (0.03)
g-Index	0.03*** (0.01)	0.06*** (0.01)	0.04*** (0.004)	−0.03*** (0.004)	−0.03*** (0.01)	−0.09*** (0.02)
hc-Index	0.11*** (0.02)	0.17*** (0.02)	0.10*** (0.01)	−0.10*** (0.01)	−0.04** (0.02)	−0.21*** (0.05)
hI-Index	0.12*** (0.03)	0.33*** (0.03)	0.20*** (0.02)	−0.15*** (0.02)	−0.32*** (0.04)	−0.35*** (0.09)
hI-Norm	0.09*** (0.02)	0.19*** (0.02)	0.12*** (0.01)	−0.09*** (0.01)	−0.14*** (0.02)	−0.24*** (0.06)
e-Index	0.04*** (0.01)	0.07*** (0.01)	0.05*** (0.01)	0.04*** (0.01)	−0.03*** (0.008638)	−0.12*** (0.03)
hm-Index	0.07*** (0.02)	0.19*** (0.02)	0.11*** (0.01)	−0.09*** (0.01)	−0.17*** (0.02)	−0.23*** (0.06)
AWCR	0.003*** (0.001)	0.01*** (0.001)	0.003*** (0.001)	−0.01*** (0.001)	−0.002* (0.001)	−0.02*** (0.01)
AW-Index	0.09*** (0.02)	0.16*** (0.02)	0.10*** (0.01)	−0.10*** (0.01)	−0.04** (0.02)	−0.23*** (0.05)
AWCpA	0.01*** (0.002)	0.01*** (0.002)	0.01*** (0.001)	−0.01*** (0.002)	−0.01*** (0.003)	−0.06*** (0.02)
N	1,932	1,932	1,932	1,932	1,932	1,932
Number scholars in category	80	77	440	1061	255	61

Figures in round parentheses denote standard errors. Reported results are for bivariate logit specifications for the specified dependent variable

* Denotes statistical significance at the 10 % level; ** denotes statistical significance at the 5 % level; *** denotes statistical significance at the 1 % level

[a] Denotes indices adjusted for discipline specific h-index performance metric

rating processes to be legitimate, while simultaneously their rating process causes controversy among the scholarly community.

Objective performance in terms of absolute output and in terms of the impact of output, is related to the probability of the alternative ratings as one would expect. Rising absolute output (as measured by the total papers and papers per author measures), as well as rising impact of scholarly output (as measured by the citations-based impact factors), results in a *rising* probability of receiving the premier A-rating, and a B-rating—see columns 4 through 9 of Table 4. Conversely it *lowers* the probability of holding the lower ranked C-, Y- or L-ratings—see columns 10 through 18 of Table 4. In this sense therefore, the NRF is justified in claiming that the more prestigious ratings (A and B) are associated with objectively higher performance in objective output and impact dimensions of research.

But the results also show why the scholarly community can find the outcome of the NRF ratings process controversial. Fig. 1 reports the density functions for the NRF chair, A- and

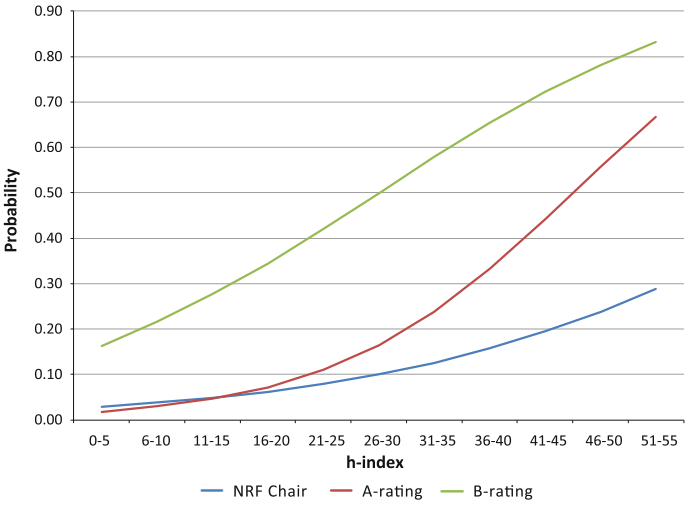

Fig. 1 Impact of adjusted h-index on probability of rating

Table 6 Minima and 25th percentile

	Minimum						25'th Percentile					
	NRF chairs	A-rated	B-rated	C-rated	Y-rated	L-rated	NRF chairs	A-rated	B-rated	C-rated	Y-rated	L-rated
Papers	1.00	1.00	1.00	1.00	1.00	1.00	18.75	13.00	15.00	9.00	6.75	9.00
Citations	1.00	1.00	1.00	1.00	1.00	1.00	48.00	75.00	46.00	19.00	17.00	9.00
Cites per year	0.25	0.11	0.11	0.10	0.11	0.15	2.89	2.53	2.80	1.47	1.70	0.88
Cites per paper	0.50	0.27	0.13	0.10	0.10	0.14	2.58	2.31	2.25	1.66	1.79	1.00
Cites per author	0.33	1.00	0.33	0.20	0.25	0.33	25.50	32.79	25.99	12.93	7.50	6.17
Papers per author	0.57	0.58	0.25	0.20	0.25	0.33	8.25	18.75	11.51	5.33	3.33	3.27
Authors per paper	1.00	1.00	1.00	1.00	1.00	1.00	2.24	1.53	1.90	1.88	2.36	1.83
h-Index	1.00	1.00	1.00	1.00	1.00	1.00	4.00	5.00	4.00	3.00	2.00	2.00
g-Index	1.00	1.00	1.00	1.00	1.00	1.00	6.75	4.00	6.00	4.00	3.00	2.00
hc-Index	1.00	1.00	1.00	1.00	1.00	1.00	2.50	3.00	2.00	2.00	2.00	2.00
hI-Index	0.25	0.20	0.25	0.20	0.20	0.20	1.98	3.56	2.45	1.33	0.94	1.00
hI-Norm	1.00	1.00	1.00	1.00	1.00	1.00	3.00	5.00	3.00	2.00	2.00	2.00
e-Index	1.00	1.00	1.00	1.00	1.00	1.00	5.34	8.12	4.30	3.16	3.16	1.73
hm-Index	0.25	0.20	0.25	0.20	0.20	0.20	2.40	2.50	2.83	1.95	1.33	1.33
AWCR	0.20	0.11	0.11	0.11	0.20	0.22	8.13	13.86	6.65	3.17	3.46	1.59
AW-Index	0.25	0.18	0.13	0.16	0.12	0.28	2.61	2.77	2.70	1.86	1.92	1.48
AWCpA	0.10	0.11	0.10	0.12	0.12	0.22	3.29	4.97	3.43	1.72	1.58	1.24

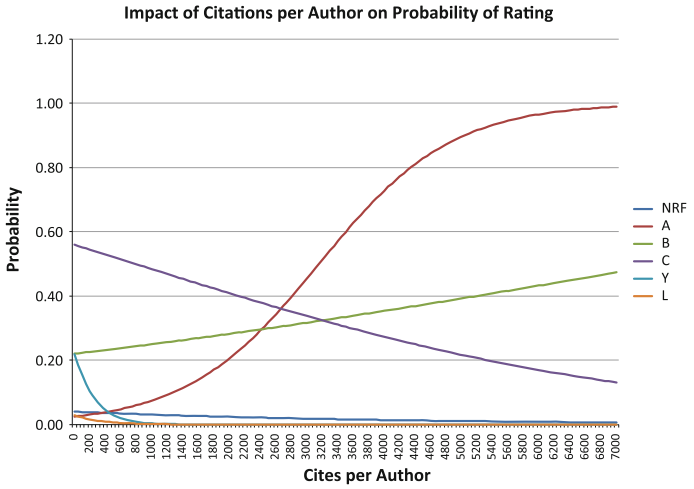

Fig. 2 Impact of citations per author on probability of rating

B-ratings for the *h*-index adjusted for discipline specific weights.[26] While we report only the *h*-index density, results are symmetrical for the papers, citations per year, citations per paper, the *g*-index, the h_c-index, the *e*-index, *hI*-norm, the *AW*-index. Thus in terms of virtually all of the objective performance measures (absolute or impact based), the probability of receiving an A-rating as well as a B-rating rises. This is as one would expect. But what is arguably controversial is that even at the very highest levels of performance in terms of most of the various objective measures employed by this study, the probability of receiving a B-rating remains higher than the probability of receiving an A-rating. Thus, even for high levels of output, and output that objectively has (citations-based) high impact, chances are that the rating that the NRF awards is B, while other scholars of directly comparable levels of output and impact factors receive an A-rating.

It is easy to understand why perceptions of inconsistency and favoritism follow.

This point compounds by virtue of the fact that scholars at distinct NRF ratings, overlap substantially in terms of their performance in terms of the objective measures of research output and impact. Consider the evidence of Table 6, which reports the minimum and 25th percentile values of the objective performance measures used by this study, conditional on the range of NRF ratings. What emerges is that there exist scholars under high NRF ratings (A-rated, NRF-chairs), that have recorded objective levels of output (eg. papers published) and impact (eg. citation counts), that fall well below those of scholars with much lower NRF-ratings (for instance, which lie below the maximum score of B-, C-, and Y-rated scholars), and are comparable to those scholars holding the *minimum* scores at the lower NRF-ratings.

The anomaly that the probability of a B-rating consistently remains higher than that of an A-rating applies to the majority of measures. However, it does not do so in the case of a

[26] While in general the implied probability of receiving a specified rating is invariant to the use of the raw *h*-index or the discipline-adjusted *h* -index (there are only marginal differences in the plied densities), in the case of the A-rating significant differences do emerge—with the discipline adjusted *h*-index generating considerably lower probability values of the A-rating than the raw measure. The reason for this is that the probability of receiving an A-rating under any given objective performance in terms of bibliometric measures is not invariant to discipline—see the discussion below.

Table 7 Logit regression: NRF-ratings—the impact of disciplines

	NRF chair	A-rated	B-rated	C-rated	Y-rated	L-rated
Constant	−3.26094*** (0.1866)	−3.54558*** (0.2555)	−1.30658*** (0.08858)	0.283165*** (0.07427)	−2.07663*** (0.1131)	−3.05395*** (0.1948)
Biological	−0.164197 (0.2550)	−0.882714** (0.4224)	0.115959 (0.1167)	−0.145624 (0.09894)	0.249914* (0.1427)	−0.477634 (0.3006)
Business	0.181729 (0.3872)	−0.439190 (0.8067)	−0.0302239 (0.1973)	0.108848 (0.1658)	0.0348814 (0.2427)	0.0396942 (0.4469)
Chemistry	0.212528 (0.2992)	0.192461 (0.4987)	0.0352571 (0.1504)	0.172928 (0.1294)	−0.160081 (0.1922)	−0.424691 (0.4447)
Engineering	0.0384059 (0.2802)	0.0471599 (0.4028)	0.0395479 (0.1369)	−0.111045 (0.1166)	0.327358** (0.1670)	−0.867661** (0.4171)
Medical	−0.0158679 (0.3311)	−0.987201 (0.5935)	−0.129794 (0.1539)	−0.0350914 (0.1258)	0.339486* (0.1735)	−0.127950 (0.3605)
Physics	0.648681*** (0.2984)	0.0100034 (0.4848)	0.403992** (0.1606)	−0.385758*** (0.1450)	−0.0912565 (0.2217)	0.0474070 (0.4205)
LL	−325.697065	−315.06442	−1,022.22956	−1,311.77531	−744.482238	−265.404762
Baseline LL	−329.1173	−319.6227	−1,026.791	−1,317.224	−749.1871	−270.1957
N	1,913	1,913	1,913	1,913	1,913	1,913
Chi-square	6.8405	9.1166	9.1227	10.898*	9.4097	9.582

Figures in round parentheses denote standard errors

LL denotes Log-Likelihood

*Denotes statistical significance at the 10 % level; ** Denotes statistical significance at the 5 % level; ***Denotes statistical significance at the 1 % level

number of the measures: the citations measure, citations per author, the *hm*-index, the *hI*-index, papers per author, *AWCR*, *AWCRpA*. See Fig. 2 by way of example. What is notable is that most of the second category of measures, in which the probability of an A-rating does come to exceed that of a B-rating at high levels of objective scholarly performance, correct for the *number* of authors that publications carry (though the pure citations count also performs as expected). It thus appears that the NRF ratings processes favour scholars that work alone, rather than as part of collaborating teams of researchers, either because it does not value collaboration or because its peer review struggles to assess the contributions of authors who are part of larger research teams. If this is the case, the NRF evaluation process appears biased against disciplines in which multi-author publications are the norm, as well as multi-disciplinary work, which is inherently collaborative.

Do disciplines matter?

One possible response to the findings of anomalies in the ratings probabilities given objective levels of performance noted in the preceding subsection, is that since the results are derived for the community of scholars in general, the anomalies are due to disciplinary differences in output and citations performance. A given level of objective performance may carry different significance across disciplines.

Conversely, a persistent source of controversy surrounding the NRF peer review mechanisms concerns allegations that disciplines receive differential treatment by the NRF process.

For both sets of reasons, we explore in more detail whether disciplinary differentials across disciplines are present in our data. To do so we begin by considering the following set of logit regressions:

$$J_i = \alpha_0 + \sum \alpha_k D_{k,i} + u_i \tag{2}$$

where

$$J_i = \begin{cases} 1 \text{ if } \exists \text{ an NRF rating of type } J = \{NRFchair, A, B, C, Y, L\}, \text{ with probability } \Pr(Y = 1) = P \\ 0 \text{ if } \nexists \text{ an } NRF \text{ rating of type } J = \{NRFchair, A, B, C, Y, L\}, \text{ with probability } \Pr(Y = 0) = 1 - P \end{cases}$$

where $D_{k,i}$ is a dummy variable denoting the disciplinary classification k of scholar i, such that $k = \{$biological sciences, business sciences, chemical sciences, medical sciences, engineering, medical sciences, physical sciences, social sciences$\}$. We employ the social sciences as the reference variable. All rated scholars were assigned to one of these categories.[27] Scholars were assigned classifications according to the nature of their home department at the time of data collection. Estimation results are reported in Table 7.

What emerges is that there are statistically significant differential probabilities of achieving different NRF classifications across disciplines. Specifically, relative to the social sciences, scholars in the physical sciences have a statistically significant greater probability of obtaining an NRF chair and a B-rating, and a statistically significantly lower probability of obtaining a C-rating. Scholars in the biological sciences have a statistically significantly lower probability of realizing an A-rating, and a statistically significantly higher probability of realizing a Y-rating than social scientists. Engineers have a statistically significantly greater probability of obtaining a Y-rating, and a lower probability of realizing an L-rating than social scientists, while scholars in the medical sciences have a statistically significantly greater probability of obtaining a Y-rating than social scientists.

The upshot of these findings is that there do exist statistically significant differences in the probabilities of achieving alternative ratings by the NRF across disciplines. The limitation of the evidence is that it is unconditional on the underlying objective performance of the scholars under the various disciplines. Thus, for instance, scholars under the biological sciences might simply not produce the same number of papers, or citations as social scientists, thereby explaining the lower probability of an A-rating in the biological disciplines.

But consider the evidence of Table 8, which records the means and medians of scholars in the range of disciplinary classifications we employ, over the measures of absolute output and impact of research we have generated for the present study. What emerges is that in terms of the absolute measures of output (the pure papers based measures), consideration of the discipline-specific means top-ranks the physical, medical and biological sciences, followed by the chemical sciences and engineering, while the business and social sciences are consistently bottom-ranked. For the impact-based measures the inference is largely the same (either the raw citations-based measure, or the many indexed derivatives that we consider for this study). The medical, biological and physical sciences are consistently top-ranked in terms of the mean measure, the chemical sciences and engineering are mid-

[27] Use of the more disaggregated classifications that the NRF Specialist panels consider is precluded by considerations of sample size in the case of a number of categories that contain a relatively small number of scholars. In the case of some researchers assignment was to mulptiple categories: for instance biochemists might be recorded both in the biological and the chemical sciences.

Table 8 Means of performance by discipline

	Means						
	Biological sciences	Business	Chemical sciences	Engineering	Medical sciences	Physical sciences	Social sciences
Papers	61.01	52.80	54.18	53.00	62.20	66.99	43.08
Citations	461.90	231.56	283.78	280.49	515.07	373.95	184.55
Citations per year	16.05	8.96	11.09	9.62	16.27	10.67	5.56
Citations per paper	6.62	4.22	4.58	4.42	7.15	4.39	3.18
Citations per author	197.84	151.74	113.71	130.27	191.72	178.39	130.98
Papers per author	28.76	30.14	26.79	28.68	25.10	31.39	32.92
Authors per paper	2.84	2.08	2.93	2.50	3.09	2.80	1.92
h-Index	9.39	6.20	7.75	6.70	9.15	7.66	5.47
g-Index	15.48	9.83	13.07	11.55	15.86	12.86	8.51
hc-Index	5.84	4.11	5.14	4.74	6.16	5.03	3.54
hI-Index	3.55	3.13	2.82	2.97	3.04	3.04	3.02
hI-Norm	6.05	4.33	4.73	4.73	5.66	5.05	4.13
e-Index	11.58	7.85	9.84	9.44	12.06	10.59	6.81
hm-Index	5.72	4.22	4.60	4.58	5.00	4.51	3.93
AWCR	48.71	35.00	37.24	37.22	45.94	35.93	20.89
AW-Index	5.93	4.21	5.02	4.74	5.96	4.85	3.46
AWCRpA	19.89	15.29	11.96	14.82	17.40	14.80	12.65
Adj. h-index[a]	7.47	8.21	7.50	12.18	6.11	8.34	7.97

[a] Denotes indices adjusted for discipline specific h-index performance metric

ranked, while the business and social sciences are bottom-ranked.[28] Under these data, finding the lower probability of an A-rating for the biological than the social sciences does become surprising. In terms of either objective absolute (paper-based) measures of output, or measures (citations-based) of research impact, scholars in the biological sciences outperform scholars in the social sciences in every dimension, by a substantial margin, and nevertheless have a consistently lower probability of obtaining an A-rating at any objective level of performance.

Consideration of the discipline-normalized *h*-index, suggests that disciplines fall into four levels of performance—see Table 8. On average, researchers in Engineering have *h*-indices above 12; those in the physical and business sciences fall into the 8–9 range of *h*-index; researchers in the biological, chemical and social sciences range from 7–8 in the *h*-index; while those in the medical sciences report *h*-indices below seven on average. The implication is that the discipline normalized *h*-index measures serves to equalize the reported performance across disciplines. But it does not explain why there should emerge differential probabilities of receiving ratings across disciplines. For instance, since

[28] Throughout, use of the median measure of central tendency leaves inferences unchanged.

Table 9 Means for the h-index across alternative ratings and disciplines

	Biological sciences	Business	Chemical sciences	Engineering	Medical sciences	Physical sciences	Social sciences
Means: baseline h-index							
NRF chairs	12	9	9	10	12	12	10
A-rated	22	16	11	12	26	14	8
B-rated	13	10	11	9	12	11	7
C-rated	8	5	7	6	8	6	5
Y-rated	7	4	6	6	6	4	4
L-rated	6	5	4	4	5	7	4
Means: discipline adjusted h-index							
NRF chairs	10	13	10	14	8	14	13
A-rated	20	17	14	18	20	16	12
B-rated	12	14	12	14	10	12	10
C-rated	7	7	7	8	6	6	7
Y-rated	6	5	6	10	4	4	4
L-rated	5	6	4	4	5	5	5

the performance of researchers in the biological and social sciences is directly comparable, it is not clear why the former have a considerably lower probability of receiving an A-rating.

However, we have not yet considered the possibility that those scholars in the social sciences that have achieved an A-rating, have achieved very high levels of output and research impact, a performance which is hidden by the measures of central tendency that capture the performance of all social scientists. To consider this possibility, in Table 9 we report the means of the *h*-index measure across the alternative NRF rating categories by disciplinary attribution of the rated scholar. The choice of the *h*-index is motivated by four considerations. First, it is amongst the most widely studied of the bibliometric indices. Second, as a result its characteristics, strengths and weaknesses are well understood. Third, as already noted above in our sample of researchers it is highly correlated with a wide range of other bibliometric measures. Finally, it is the measure for which the literature has specified a clear disciplinary adjustment factor.

The results deepen the suggestion that there are strong differences across the disciplines in terms of the ratings, even when considering objective data of performance of rated scholars. A-rated scholars in the biological and medical sciences on average have an *h*-index between 4 and 5 times as high as A-rated scholars in the social sciences. Indeed, on average, C-rated scholars in the biological sciences have the same *h*-index as A-rated scholars in the social sciences.[29] While consideration of h-index measures normalized by discipline ameliorate these gradients across disciplines, they continue to be evident—compare the social and biological sciences on the adjusted *h*-index in Table 9.

[29] It is perhaps worth reminding ourselves that if anything, the methodology by means of which the *h*-index is compiled favours the social, rather than the natural sciences. Thus the cross-disciplinary performance differential is, if anything, understated.

Table 10 Logit regressions for baseline and discipline adjusted h-indexes

	NRF chair		A-rated		B-rated		C-rated		Y-rated		L-rated	
	Baseline	Adjusted	Baseline	Adjusted	Baseline	Adjusted	Baseline	Adjusted	Baseline	Adjusted	Baseline	Adjusted
Constant	−3.65***	−3.65***	−4.55***	−4.52***	−1.88***	−1.86***	0.67***	0.66***	−1.28***	−1.26***	−2.51***	−2.52***
	(0.18)	(0.18)	(0.22)	(0.22)	(0.09)	(0.09)	(0.07)	(0.07)	(0.11)	(0.11)	(0.19)	(0.19)
h-index	0.06**	0.05**	0.10***	0.11***	0.06***	0.06***	−0.05***	−0.04***	−0.08***	−0.09***	−0.21***	−0.21***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.06)	(0.06)
Biological x h-index	−0.03	−0.02	0.02	0.04**	0.01	0.02**	−0.02	−0.03**	0.01	0.01	0.03	0.002
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.05)
Chemistry x h-index	0.02	0.02	−0.01	−0.002	0.03**	0.02*	−0.003	−0.0004	−0.01	−0.02	−0.09	−0.08
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.10)	(0.10)
Engineering x h-index	0.01	−0.01	0.02	−0.03*	0.01	0.01	−0.03*	−0.003	0.04*	0.07***	−0.16	−0.07
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.11)	(0.08)
Medical x h-index	−0.03	−0.01	−0.01	0.03*	−0.02*	−0.004	0.01	−0.01	0.02	0.01	0.06	0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.05)	(0.06)
Physics x h-index	0.03	0.03	0.05**	0.03	0.04**	0.03**	−0.05***	−0.04***	−0.07**	−0.05*	0.08	0.08
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.04)	(0.03)	(0.06)	(0.06)
Social Sci x h-index	0.02	0.0004	0.02	−0.03*	0.01	−0.02	−0.0003	0.02	−0.18***	−0.09***	0.11**	0.15***
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.05)	(0.03)	(0.05)	(0.05)
LL	−320.37	−320.06	−272.30	−274.02	−978.90	−980.95	−1284.98	−1285.37	−720.00	−717.84	−250.06	−250.34
Baseline LL	−333.06	−333.06	−323.58	−323.58	−1036.59	−1036.59	−1329.80	−1329.80	−753.77	−753.77	−270.81	−270.81
N	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932	1,932
Chi-square	25.38***	26.002***	102.55***	99.13***	115.37***	111.26***	89.65***	88.87***	67.53***	71.86	41.50***	40.94***

Figures in round parentheses denote standard errors

LL denotes Log-Likelihood

* Denotes statistical significance at the 10 % level; ** denotes statistical significance at the 5 % level; *** denotes statistical significance at the 1 % level

The implication of these findings is that the cross-disciplinary differences in the likelihood of realizing high ratings by the NRF, are at best weakly, at worst perversely related to objective measures of research output and impact of scholars.

This leaves the question of how much these disciplinary differences matter. To investigate this question we estimate:

$$J_i = \alpha_0 + \alpha_1 R_i + \sum \gamma_k \left(R_i D_{k,i}\right) + u_i \tag{3}$$

where

$$J_i = \begin{cases} 1 \text{ if } \exists \text{ an NRF rating of type } J = \{NRFChair, A, B, C, Y, L\}, \text{ with probability } \Pr(Y = 1) = P \\ 0 \text{ if } \nexists \text{ an NRF rating of type } J = \{NRFChair, A, B, C, Y, L\}, \text{ with probability } \Pr(Y = 0) = 1 - P \end{cases}$$

where R_i denotes the objective performance measure (we report the results for the *h*-index, though the results are consistent across the alternative composite performance measures). For the disciplinary binary variables, $D_{k,i}$, we treat the business sciences as the reference category. Estimation is under the logit distribution, with results reported in Table 10. We estimate under both the raw *h*-index (baseline), and the *h*-index adjusted by discipline specific normalization factors. This specification allows us to establish how changes in objective performance impacts the probability of alternative NRF ratings, and whether this impact is differentiated across disciplines.

We continue to find that improved performance of scholars, as measured by the *h*-index raises the probability of realizing an NRF chair, an A- and a B-rating, but that it lowers the probability of realizing a C-, Y-, and L-rating, as already discussed under "Deriving the probability of alternative ratings" section . The finding is invariant to the use of the raw baseline or discipline-normalized *h*-index.

In addition, it emerges that there are strong disciplinary differences in terms of the impact that improved performance under the *h*-index measure has on the probability of achieving the different NRF ratings, or an NRF chair. In terms of statistical significance, particularly the physical sciences are marked by a more rapidly rising probability of an A- or a B-rating with a rising h-index, and a more rapidly falling probability of realizing a C-, Y- or L-rating than the reference category (business sciences). The chemical sciences report a more rapidly rising probability of a B-rating in response to improved performance under the *h*-index than the reference category (business sciences) which is statistically significant, and the same holds for engineering under the Y-rating, and the social sciences under the L-rating. Finally, improved performance under the *h*-index reduces the probability of a C-rating more rapidly relative to the reference category (business sciences) for engineering and the biological sciences, and in the Y-rating for the social sciences. Again these findings are invariant to the use of the raw or discipline normalized *h*-index.

The inference is that the same increases in performance under the *h*-index, will result in more rapid transition from the lower ratings (L, Y and C), and a more rapid rise through the higher ratings (B, A and NRF Chair) in the physical sciences than in the business sciences. Transition out of the C-rating is also "easier" (as measured by performance under the *h*-index) than for the business sciences for engineering and the biological sciences, and out of the Y-rating for the social sciences, while achieving a B-rating is more readily achieved under the business sciences.

Cross-disciplinary differences in terms of the relation between ratings outcomes and objective measures of performance thus undoubtedly emerge from the data.

To illustrate how substantial these disciplinary differences are, we report the implied probability density from the estimations of achieving the A-rating under varying levels of

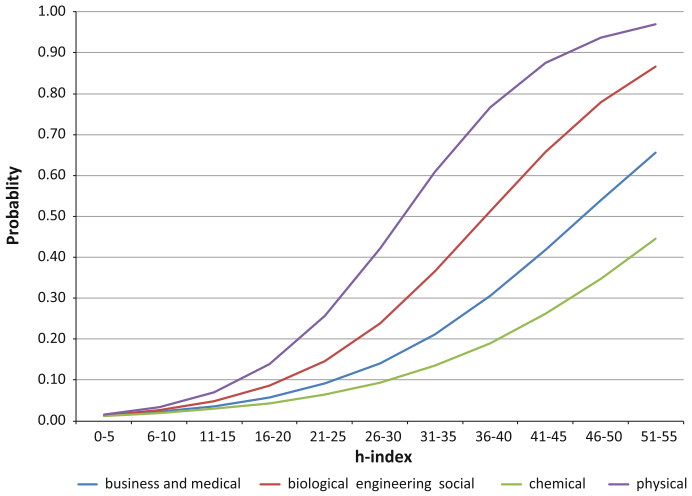

Fig. 3 Impact of h-index on probability of A-rating

performance in terms of the *h*-index, across the disciplines in Fig. 3 (the implied probability values are invariant across the raw and discipline normalized *h*-index).[30] Disciplines fall into four categories: in the physical sciences the probability of an A-rating rises most rapidly with improvements in the h-index, closely followed by the biological, engineering and social sciences. The probability of an A-rating responds much more weakly to improved h-index performance in the business and medical sciences, and most weakly of all in the chemical sciences.[31] It is worth noting that the responsiveness of the various disciplines not only vary, but vary substantially in terms of the probability of realizing higher ratings in response to increases in the objective performance measures (*h*-index). By way of illustration, in the case of an A-rating, an *h*-index rating of 55 translates into a probability of approximately 90 % for an A-rating in the case of the physical and social sciences, but of only 70 % in the case of the business and medical sciences.

Dramatic differences across disciplines also emerge with respect to the NRF chair category, which carries the highest level of financial research support that the NRF offers. In this instance there is a marked difference between the results obtained from the raw and the discipline normalized *h*-index. Table 11 reports the probability of observing an NRF chair against values of the *h*-index. Noticeably, once the discipline normalized measure is employed, the increase in the probability of observing an NRF chair as the h-index increases is dramatically lower. Even more startling, for some disciplines (business, biological, medical) the increase in the probability of observing an NRF chair against a rising *h*-index is either negligible or small (less than 10 %). Only in the physical and social sciences does there appear to be an appreciable response in the probability of an NRF chair with rising *h*-index performance. Once again, not only do there appear to be strong

[30] We report only the highest rating category, since this carries the greatest prestige and funding implications. Results for the remaining ratings categories are available from the author.

[31] While we do not report the densities explicitly, in the case of a B-rating, the strongest probability response to a rising *h*-index again emerges for the physical sciences, followed by the social and chemical sciences, then engineering and the biological sciences, then business sciences and finally the medical sciences.

cross-disciplinary differences in the way in which the NRF awards its research chairs—but in the case of the business, biological and medical sciences the award of research chairs seems to be entirely divorced from objective underlying research performance as measured by the *h*-index. A startling finding for the NRF category targeted specifically at world leading researchers (both in terms of status and in funding support).

In short, what the evidence implies is that the disciplines in which it is easiest to translate objective performance into a higher NRF rating are the physical sciences, closely followed by the social sciences. Engineering and the chemical sciences are the next most responsive to improvements in objective performance measures. The business sciences, the biological and the medical sciences appear the most demanding in terms of their requirements in terms of objective performance in the NRF rating system. Thus in the physical and social sciences high ratings are "easy" to obtain in the sense that little objective performance evidence is required. While in the business, biological, the chemical (in the case of the A-rating) and medical sciences high ratings are "hard" to obtain in terms of the same metric.

Our analysis cannot determine which of these alternative degrees of stringency in terms of objective performance is the correct one. But the study can and does note that disciplines do appear to differ not only statistically, but in terms of objective probabilities of alternative ratings emerging on the basis of similar objective research impact performance. An inference of cross-disciplinary bias is difficult to avoid.

Conclusions and evaluation

This paper has examined the strength of association between the outcomes of a research funding agency's peer review based rating mechanisms, and a range of measures of performance of scholars in terms of both absolute output (principally counts of publications in either raw or normalized form), as well as measures of the impact of research output (principally citation counts, either in raw or normalized format). The analysis is conducted on 1932 scholars that have received a rating or a research chair by South Africa's NRF.

Concern of the analysis is to address the reliability of peer review and related perceptions that it is subject to bias and inconsistent standards.

Our findings are mixed.

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 ratings. In addition, the higher the performance of scholars against all objective measures of absolute output and impact, increases the probability of an A- or B-rating and of holding an NRF research chair, and lowers the probability of a C-, Y and L-rating, which accords with the implicit ranking amongst the various ratings the NRF awards.

Such evidence accords with NRF claims that its peer review mechanisms reflect the scholarly standing of researchers, and in particular that ratings reflect impact as well as absolute levels of productivity.

But there is countervailing evidence. First, we find that on a range of objective measures of performance, the probability of achieving a B-rating remains higher than that of achieving an A-rating even at the very highest levels of recorded performance for South African scholars. This is only reversed for objective measures of performance that undertake downward corrections for output generated by multiple authors. The inference is that the NRF peer review either does not, or cannot, value output generated by larger teams of researchers. Thus researchers in disciplines where collaborative research is the norm, or

Table 11 Probability density for NRF chairs by discipline

	Business		Biological		Chemical		Engineering		Medical		Physical		Social	
h-Index	Baseline h-index	Adjusted h-index	Baseline h-index	Adjusted h-index	Baseline h-index	Adjusted h-index	Baseline h-index	Adjusted h-index	Baseline h-index	Adjusted h-index	Baseline h-index	Adjusted h-index	Baseline h-index	Adjusted h-index
0–5	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.03	0.03
6–10	0.04	0.03	0.03	0.03	0.05	0.04	0.04	0.04	0.03	0.04	0.05	0.05	0.05	0.05
11–15	0.05	0.04	0.04	0.03	0.06	0.05	0.06	0.05	0.04	0.04	0.08	0.07	0.07	0.06
16–20	0.07	0.04	0.04	0.03	0.09	0.07	0.08	0.06	0.04	0.04	0.12	0.10	0.10	0.08
21–25	0.09	0.04	0.05	0.03	0.13	0.08	0.12	0.07	0.05	0.05	0.18	0.14	0.14	0.10
26–30	0.12	0.05	0.05	0.04	0.17	0.10	0.16	0.08	0.06	0.06	0.26	0.19	0.20	0.13
31–35	0.15	0.05	0.06	0.04	0.23	0.12	0.21	0.09	0.07	0.06	0.36	0.24	0.27	0.17
36–40	0.20	0.05	0.07	0.04	0.31	0.14	0.27	0.11	0.08	0.07	0.47	0.32	0.36	0.21
41–45	0.25	0.06	0.08	0.04	0.39	0.17	0.35	0.13	0.09	0.08	0.58	0.40	0.46	0.27
46–50	0.31	0.06	0.09	0.04	0.48	0.21	0.43	0.15	0.10	0.08	0.69	0.48	0.56	0.32
51–55	0.37	0.07	0.10	0.04	0.58	0.25	0.52	0.17	0.12	0.09	0.78	0.57	0.66	0.39

interdisciplinary research which by its nature is collaborative, will face greater challenges in achieving higher ratings than researchers who work on their own, and in pure core disciplines.

Second, we find that the variance of objective performance under each rating category is large. The result is that scholars who have received the highest ratings (A-ratings or NRF research chairs) record objective levels of research output and impact of their research that are no different from the *minimum* levels of objective performance at much lower NRF ratings (eg. the C-rating). This finding is particularly striking with respect to the category (NRF research chairs) that is advertised as being associated with attracting world class scholars, and which is tied to the very highest level of funding grants (an automatic ZAR 3 million per annum, for a minimum of five, but possibly 15 years).

Third, the probability of obtaining alternative NRF ratings is statistically significantly different across alternative disciplines. For instance, researchers in the physical sciences have a statistically significantly higher propensity to realize an A-rating or an NRF research chair, while those in the biological sciences have a significantly lower probability to do so. Such differences persist even when one considers the objective performance in terms of research output and impact of scholars in disciplines.

In fact, in terms of the impact of objective measures of performance on the probability of achieving alternative NRF ratings, we find differences across disciplines that are both statistically significant and substantial in terms of implied probability density. In summary, high ratings are "easy" to obtain in the physical and social sciences, and relatively "hard" to obtain in the business, biological, chemical and medical sciences, in the sense that in the case of the former disciplines high ratings are awarded at low to moderate levels of objective output and impact relative to the latter group of disciplines.

What is more, in the premier NRF chair category, in the business, biological and medical sciences the award of research chairs seems to be entirely divorced from objective underlying research performance.

The set of findings under disciplinary differences under the NRF rating mechanisms presents a direct challenge to NRF claims that its peer review mechanisms are designed to ensure that no cross-disciplinary biases emerge. Such claims do not accord with the data. Since research-active scholars are likely to have an understanding of the impact of the work of their peers in cognate disciplines, when the outcome of NRF peer review begins to bear weak association with objective measures of output and impact, it is not surprising that the legitimacy of the review process is brought into question.

In summary, therefore, the empirical findings of the paper clarify both why the NRF suggests that its peer review mechanisms reflect underlying research performance of scholars, and why researchers hold perceptions of bias and inconsistency in the application of the NRF rating mechanisms. Both sets of claim have a basis in fact.

Acknowledgements The author acknowledges the research support of Economic Research Southern Africa in completing this paper.

References

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics, 87*(3), 499–514.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). National research assessment exercises: Acomparison of peer review and bibliometrics rankings. *Scientometrics, 89*(3), 929–941.

Adler, N., Elmquist, M., & Norrgre, F. (2009). The challenge of managing boundary-spanning research activities: Experiences from the Swedish context. *Research Policy* 38, 1136–1149.

Archambault, E., & Gagné, E. V. (2004). *The use of bibliometrics in social sciences and Humanities*. Montreal: Social Sciences and Humanities Research Council of Canada (SSHRCC).

Bar-Ilan, J. (2008). Which h-index?: A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*(2), 257–271.

Batista, P. D., Campiteli, M. G., Kinouchi, O., & Martinez, A. S. (2006). Is it possible to compare researchers with different scientific interests?. *Scientometrics, 68*(1), 179–189.

Bedeian, A. G. (2003). The manuscript review process: The proper role of authors, referees, and editors. *Journal of Management Inquiry, 12*(4), 331–338.

Belew, R. K. (2005). Scientific impact quantity and quality: Analysis of two sources of bibliographic data, arXiv:cs.IR/0504036 v1, 11 August 2005.

Benner, M., & Sandström, U. (2000). Institutionalizing the triple helix: research funding and norms in the academic system. *Research Policy, 29*, 291–301.

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

Bornmann, L., & Daniel, H. -D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics, 65*(3), 391–392.

Bornmann, L., & Daniel, H. -D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology, 58*(9), 1381–1385.

Bornmann, L., Mutz, R., & Daniel, H. -D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology, 59(5), 830–837.*

Bornmann, L., Mutz, R., Daniel, H. -D., Wallon, G., & Ledin, A. (2009a). Are there really two types of h index variants? A validation study by using molecular life sciences data. *Research Evaluation, 18*(3), 185–190.

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H. D. (2009b). Convergent validity of bibliometric Google Scholar data in the field of chemistry. Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *Journal of Informetrics*, *3*(1), 27–35.

Bosman, J, Mourik, I. van, Rasch, M.; Sieverts, E., & Verhoeff, H. (2006). Scopus reviewed and compared. The coverage and functionality of the citation database Scopus, including comparisons with Web of Science and Google Scholar. Utrecht: Utrecht University Library. Retrieved from http://igitur-archive.library.uu.nl/DARLIN/2006-1220-200432/Scopusdoorgelicht&vergeleken-translated.pdf.

Butler, L. (2006). RQF pilot study project: History and political science methodology for citation analysis, November 2006. Retrieved from http://www.chass.org.au/papers/bibliometrics/CHASS_Methodology.pdf.

Cronin, B., & Meho, L. (2006). Using the h-Index to rank Influential Information scientists. *Journal of the American Association for Information Science and Technology, 57*(9), 1275–1278.

Debackere, K., & Glänzel, W. (2003). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics, 59*(2), 253–276.

Demicheli, V., & Pietrantonj, C. (2007). Peer review for improving the quality of grant applications. *Cochrane Database of Systematic Reviews, 2*, Art. No.: MR000003. doi:10.1002/14651858.MR000003.pub2.

Derrick, G. E., Sturk, H., Haynes, A. S., Chapman, S., & Hall, W. D. (2010). A cautionary bibliometric tale of two cities. *Scientometrics, 84*(2), 317–320.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics, 69* (1), 131–152.

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics, 69*(1), 121–129.

Eysenck, H. J., & Eysenck, S. B. G. (1992). Peer review: Advice to referees and contributors. *Personality and Individual Differences, 13*(4), 393–399.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*, 338–342 doi: 10.1096/fj.07-9492LSF

Frey, B. S. (2003). Publishing and prostitution? Choosing between one's own ideas and academic success. *Public Choice, 116*(1–2), 205–223.

García-Aracil, A., Gracia, A. G., & Pérez-Marín, M. (2006). Analysis of the evaluation process of the research performance: An empirical case. *Scientometrics, 67*(2), 213–230.

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h Indices in psychology. *Journal of the American Society for Information Science and Technology, 61*(10), 2070–2085.
Glänzel, W. (2006). On the opportunities and limitations of the H-index. *Science Focus, 1*(1), 10–11.
Goodman, S. N., Berlin, J., Fletcher, S. W., & Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at annals of internal medicine. *Annals of Internal Medicine, 121*(1), 11–21.
Gray, J. E., Hamilton, M. C., Hauser, A., Janz, M. M., Peters, J. P., & Taggart, F. (2012). Scholarish: Google Scholar and its value to the sciences. *Issues in Science and Technology Librarianship, 70*(Summer). doi::10:5062/F4MK69T9.
Harzing, A.-W. (2007–2008). Google Scholar: Anew data source for citation analysis. Retrieved from http://www.harzing.com/pop_gs.htm.
Harzing, A.-W. (2008). Reflections on the h-index. Retrieved from http://www.harzing.com/pop_hindex.htm.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output, arXiv:physics/0508025 v5, 29 Sep 2006.
Horrobin, D. F. (1990). The philosophical basis for peer review and the suppression of innovation. *Journal of the American Medical Association, 263*(10), 1438–1441.
Iglesias, J. E., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics, 73*(3), 303–320.
Jacsó, P. (2005). Google Scholar: The pros and the cons. *Online Information Review, 29*(2), 208–214.
Jacsó, P. (2006a). Dubious hit counts and cuckoo's eggs. *Online Information Review, 30*(2), 188–193.
Jacsó, P. (2006b). Deflated, inflated and phantom citation counts. *Online Information Review, 30*(3), 297–309.
Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review, 34*(1), 175–191.
Jin, B. (2007). The AR-index: Complementing the h-index. *ISSI Newsletter, 3*(1), 6.
Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *Journal of the American Medical Association, 302*(10), 1092–1096.
Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055–1065.
Kousha, K, & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation index: A comparison between four science disciplines. *Scientometrics, 74*(2), 273–294.
Leydesdorff, L. & Etzkowitz, H. (1996). Emergence of a Triple-Helix of university–industry–government relations. *Science and Public Policy, 23*(5), 279–296.
Meho, L. I., & Yang, K. (2007). A new era in citation and bibliometric analyses: Web of Science, Scopus, and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105–2125.
Moed, H. F. (2002). The impact-factors debate: The ISI's uses and limits. *Nature, 415*, 731–732.
Moxam, H., & Anderson, J. (1992). Peer review. *A view from the inside. Science and Technology Policy, 5*(1), 7–15.
Nisonger, T. E. (2004). Citation autobiography: An investigation of ISI database coverage in determining author citedness. *College & Research Libraries, 65*(2), 152–163.
Pendlebury, D. A. (2008). *Using bibliometrics in evaluating research*. Philadelphia, PA: Research Department, Thomson Scientific.
Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. *Scientometrics, 57*(1), 1–11.
Pierie, J. P. E. N., Walvoort, H. C., & Overbeke, A. J. P. M. (1996). Reader's evaluation of effect of peer review and editing on quality of articles in the Nederlans Tijdschrift voor Geneeskunde. *Lancet, 348*(9040), 1480–1483.
Rehn, C., Kronman, U., & Wadskog, D. (2007). Bibliometric indicators: Definitions and usage at Karolinska Institutet, Stickholm. Sweden: Karolinska Institutet University Library.
Roediger III, H. L. (2006). The h index in Science: A new measure of scholarly contribution. *APS Observer: The Academic Observer, 19*, 4.
Saad, G. (2006). Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively. *Scientometrics, 69*(1), 117–120.
Schreiber, M. (2008). To share the fame in a fair way, h_m modifies h for multi-authored manuscripts. *New Journal of Physics, 10*, 040201-1–8.
Shatz, D. (2004). *Peer review: A critical inqury*. Lanham, MD: Rowman and Littlefield.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2006). Generalized h-index for disclosing latent facts in citation networks, arXiv:cs.DL/0607066 v1, 13 Jul 2006.

Testa, J. (2004). The Thomson scientific journal selection process. Retrieved from http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/.

Thor, A., & Bornmann, L. (2011). The calculation of the single publication h index and related performance measures: A web application based on Google Scholar data. *Online Information Review, 35*(2), 291–300.

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133–143.

Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. *Scientometrics, 67*(3), 491–502.

Vaughan, L., & Shaw, D. (2008). A new look at evidence of scholarly citations in citation indexes and from web sources. *Scientometrics, 74*(2), 317–330.

Wessely, S. (1998). Peer review of grant applications: What do we know?. *Lancet, 352*(9124), 301–305.

Zhang, C.-T. (2009). The e-index, complementing the h-index for excess citations. *PLoS ONE, 5*(5), e5429.