Article

Publish and perish:
A new look at bibliometric statistics in the PBRF age

Geoffrey K. Chambers[a*], Jonathan P.A. Gardner[a] and Alastair G. Smith[b]
[a]School of Biological Sciences and [b]School of Information Management,
Victoria University of Wellington, PO Box 600, Wellington

There is no doubt that New Zealand scientists are living uncomfortably in a new age of accountability – the so-called 'audit society' (Power 1997). It is scant comfort to know that we are not alone in the world either (Lawrence 2007). This article is concerned with finding the best and most ethical way for scholars and scientists to present reliable evidence concerning the quantity and quality of their published research output and how this may influence their decisions about where to submit their manuscripts.

Introduction

There are clear obligations for transparency and accountability in association with any activity that makes use of public funds. Nothing is new about the need to recognise these values in science and science management. Practitioners have always borne greater or lesser requirements to account for laboratory management, ethical standards, student supervision and fiscal responsibility, etc. These obligations are now writ large in professional life and closely scrutinised by administrators. The latest development has been a request to provide clear evidence regarding the impact and standing of individual research efforts for promotions, etc. Even the status of a most distinguished contributor, like Professor David Penny of Massey University, is commonly reflected via statistics (see Berridge & Petrey 2009).

In New Zealand, the greatest challenge comes in the form of our regular *Performance Based Research Fund* exercises (PBRF, see Box A). This is a serious business because a significant fraction of base funding for tertiary education, c. $239 million in 2009 (www.tec.govt.nz) now comes from this source. A recognisably similar system in the UK called the *Research Assessment Exercise* (RAE) predates the PBRF, but has important differences; namely that the RAE unit of assessment is the department rather than the individual and only involves a qualitative peer review of selected work. Significantly, the RAE is presently moving towards a more quantitative statistics-based evaluation (see Noble 2010). In China, pressure derived from individual assessment practices is claimed to have led to

*Correspondence: geoff.chambers@vuw.ac.nz

Geoff Chambers is a Senior Research and Teaching Fellow in Molecular Biology and Evolution at Victoria University of Wellington (VUW). His research speciality is DNA technology applied to projects ranging from human ancestry and health through to biological systematics and wildlife conservation. Dr Chambers also has longstanding interests in science policy and management. The present article has allowed him to integrate assessment of research performance into this latter programme of scholarship.

Jonathan Gardner is a marine biologist and Director of the Centre for Marine Environmental & Economic Research at VUW. Since his appointment in 1994, Jonathan's research focus has included marine ecology, population genetics and aquaculture. He has served on VUW's internal committees for all Performance Based Research Fund (PBRF) rounds, as well as for the recent VUW internal assessment exercise. At the time of writing, according to his own records, he has published 69 peer-reviewed papers. He has no idea what his *h-index* score is because several of his papers are not listed in the usual databases.

Alastair Smith is a Senior Lecturer in the School of Information Management at Victoria University of Wellington. He teaches courses in reference work, information retrieval, and digital libraries on the postgraduate programme for librarians and record managers, Master of Information Studies. Prior to 1989, he was involved in database development at the National Library of New Zealand, and worked as a librarian in scientific and technical information services. His pre-library careers included school teaching and being a patent examiner. He has a BSc in physics, an MA on the use of expert systems, and is an Fellow of LIANZA, the Library and Information Association of New Zealand/Aotearoa. He has served on various committees of LIANZA. Alastair's web page is: http://www.vuw.ac.nz/staff/alastair_smith/

Box A: PBRF and other systems

Performance Based Research Fund (NZ)

This scheme was initiated by the *Tertiary Education Commission* (TEC: www.tec.gov.nz) in 2002 at the direction of the New Zealand Minister of Education and runs on a six-year cycle. Investigators each submit an *Evidence Portfolio* (EP) which is assessed by expert panel(s) based on their evaluation of data presented in various categories including four selected publications or *'nominated research outputs'*. Individual scholars are rated on their Research Output (70%), Peer Esteem (15%) and Contributions to the Research Environment (15%) by panels of discipline experts who award scores to the EPs. These scores range from A to C in decreasing order of scholastic merit plus R(NE) for *new and emerging researcher* or R for *research inactive*. The EPs contribute 60% to the institutional score for staff research quality plus 25% on external research income and 15% higher degree completion.The guidelines for the 2012 PBRF exercise have recently been published online (www.tec.govt.nz).

Research Assessment Exercise (UK)

This is run every five years on behalf of the various higher education research councils (www.rae.ac.uk) and scores are awarded on the basis of proportions of departmental work*… equating to attainable standards of international excellence…* The RAE ranking system seems to have changed with each succeeding round and is due to be replaced by a system of *metrics* incorporated in the new *Research Excellence Framework* following the completion of the 2008 RAE round.

Research Quality Framework (Australia)

This scheme is intended to measure quality and impact in research in the tertiary sector but has an *off again, on again* history. It seems destined to be reborn as *Excellence in Research for Australia* under the *Transforming Australia's Higher Education System* paper released in 2009 (http://www.dweer.gov.au).

a detrimental migration of scientists away from the important field of biological systematics (Jiao 2009). The PBRF scheme too has many critics. Middleton (2009) has provided a well-argued evaluation of the effect of the PBRF on a 'professional' subject, Education, and Roa *et al.* (2009) present a compelling description of the difficulties faced by Māori research.

The twin dilemmas

Clearly academic scientists face substantial novel complications in their present working environment. In the end, it all comes down to just two key questions: (1) What numeric indicators are most appropriate as evidence of individual scientific achievement? and (2) Where should they try to publish their research findings in order to maximise future scores? This article begins by trying to answer the second question first and includes some best-practice advice regarding the first along the way and in more detail, later. In one sense the answers to both are obvious – always pick the top journals in your field! Surely everyone is aware by gestalt of the standing of each journal and the standard of work one expects to find inside? This may feel like being back at school, where everyone knows how boys in their class rank as playground fighters and which girls are the most popular. So the optimum solution seems simple; keep your records up to date and publish lots of papers in the *best* journals and everyone knows which ones are the *best* journals. This reasoning (popular with some science managers of common experience) amounts to a covert expert system, i.e. one is actually making a series of analytical decisions without being aware of having done so. Naturally, when one tries to be explicit about such decisions and just where to find the best evidence to support them, things may turn out not to be so easy after all. Advice reported from certain managers to *… just quote any favourable statistic to support your choice of publisher...* is unhelpful at best and dishonest at worst.

Selecting a successful strategy

As argued above, the best way to select a successful strategy plus supporting facts and statistics comes down to routinely choosing to submit articles to high-quality peer-reviewed international journals and by showing metrics to prove that the journals are, in fact, high-quality. It also helps if you can show that lots of people want to publish in these chosen journals, which may be easier said than done. Alternatively, one might try to show that the editors are highly discriminating and consistently maintain a high rejection rate, which is not quite the same thing. Or perhaps demonstrate that the journals are influential, i.e. that articles in such publications are frequently cited in other papers as measured by the infamous *Journal Impact Factor* (JIF) – more on this later, and see Box B. Oh, and it helps if you make sure that you are the sole or first author. This is often an impossible condition for senior scientists and project leaders to satisfy, because modern science is frequently conducted via a contingent version of the *multidisciplinary team approach*. These collective criteria for journal selection may perhaps seem self-evident to some, but nonetheless they do bear scrutiny.

Peer review is a given, but how does one know that the reviewers are always of high quality? Here one must probably depend on the unmeasured social profile of the editorial board. Hence, there are no supportive statistics to be found here. International journals are conventionally preferred over national journals in the unvoiced belief that one is playing in a higher league. This may very well be generally true, but can never be taken as a certainty, and the dividing line is unclear. Local preference for overseas journals seems almost to be a defining characteristic of our national psyche. New Zealanders just seem to value recognition from overseas particularly highly. This may simply reflect the people of a young nation struggling for self-confidence. The result is that researchers could find themselves on the sharp end of ill-informed managerial mentoring and directed in no uncertain fashion to submit exclusively to international journals. Failure to do so might even be criticised as … *squandering time and effort*... or not being …*loyal to core institutional values and strategies*…. When followed to extremes, this practice would see the complete demise of domestic science publication!

Competition for entry to pages of a given publication is doubtless some sort of an index of impact/quality of their average article. For any given journal, this competition is some complex function of variables such as submission and acceptance rates, the number of issues per year, pages per issue, and relative numbers of different article types per issue. Admittedly, it is difficult to get comparative statistics that are equally applicable to different journals. There are next to no available statistics for such variables and besides, there may be obvious, or not so obvious, biases to acceptance. In any case, one must question if it really is intelligent behaviour to invest the required effort to submit an article to a prestige journal such as *Science* or *Nature* when one knows in advance that it will start out with a high chance of rejection. Some have even argued that good papers are often eliminated to make room for more newsworthy articles which may not represent such high-quality science (e.g. see Hilborn 2006 for comments on stories about collapsing fisheries). Hence our earlier advice to pick the best journals might be more effective when modified to: pick the highest-ranking sources in one's field and ones where there is a **reasonable** chance of getting them published.

So how can one show that one has picked the *best* journals? Here at last there is verifiable help in the form of various electronic databases and metrics. The best known of these are listed in Boxes B (metrics) and C (databases). Journal influence (or *impact*) is usually taken to be reflected by citation counts. The proposition is that those papers which have been most frequently quoted by other papers have the greatest influence on the progress of science; ditto for the journals themselves. This property is said by many to be captured by the JIF statistic (see informative comments by its inventor in Garfield 1996, 2005). The controversies surrounding this statistic and its application (or more correctly its misapplication) are legion. These include, but are not limited to, the following:

- Should the census period be longer than two years (say five years) to better reflect the regular pattern of citation history?
- Is JIF even a good measure when up to 90% of all citations may be attributed to the top 10–25% of articles?; e.g. see recent comments from the Editor-in-Chief of *Nature* (Campbell 2008).
- Should the JIF be applied exclusively to rank journals (as consistently recommended by Eugene Garfield – see above) or to rank scholars too?

Box B: Journal Impact Factors

The Australian Journal Ranking System

The Australian Research Council (www.arc.gov.au) produced a draft ranked journals list (ARC 2008) to reflect professional gestalt for their planned research assessment exercise. It does contain a few alarming assignments but was in general accord with popular wisdom in rating journals from A* to C, at least for some disciplines, but some in our view, including marine biology, feature some highly questionable rankings. A revised version is now available as part of their 2010 ERA exercise.

Journal Impact Factor

The *Web of Science* (WoS) is an online database available from Thompson Reuters (www.thompsonreuters.com) as part of their ISI *Web of Knowledge*. The JIF for a journal is calculated by counting up all the references to articles in a particular journal contained in other journals in the Thompson Scientific database over a period of two years and dividing by the number of articles published in the focal journal over the most recent two-year census period (effectively the mean number of citations per article).

Journal Citation Reports

The WoS also provides *Journal Citation Reports* (JCR) within various sub-disciplines. While these may simply seem to represent a descending list by Impact Factor, they do provide single compellingly simple statistics. The problem with JCRs is exactly that they depend on the Impact Factor – a score that properly attaches to journals and not authors or articles. This service is now (from 2009) enhanced by inclusion of Eigenfactor assessment of journal influence. The Eigenfactor method is not yet in widespread use, but seems to have promise because it creates scores by weighting citations based on the impact score of the journal in which the citation appears, and covers a huge number of sources (Bergstrom 2007).

Scopus Ranking Systems

The Scopus-based equivalent is *SCImago Journal Ranks* (SJR) and is available direct from the Scopus webpage (http://www.scimago.es). There is a fairly strong correlation between JCR and SJR values (Thomaz & Martens 2009), but Butler (2008) reported a few rather surprising discrepancies between the two in ranking ten top biomedical journals. Their weighted contextual statistic is the Source Normalised Impact per Paper (SNIP).

Box C: Electronic databases

Scopus (Elsevier)

This database requires institutional subscription and the URL will vary depending on one's host institution but a great deal of general information can be obtained from the home page (www.scopus.com). It covers more than 15000 journal titles from 1996 to 2009.

Web of Science (Thompson Reuters)

Formerly known as the *Institute for Scientific Information* this database covers 9000 specially selected journals which the owners claim include all those of highest impact.

Google Scholar

This is a search engine (www.scholar.google.co.nz) which can recover lists of academic outputs in all sources together with citation information. It has much wider coverage than either WoS or Scopus, but is much less discriminating than either and it is difficult to refine searches effectively. It is not clear if it is safe to treat all citation scores returned under multiple listings as independent.

- Is it necessary to correct JIF scores for missing citations (i.e. those outside the Thompson Reuters database, including all books) and/or incorrectly made or incorrectly credited/duplicated citations?
- Is the JIF a non-linear metric with highly cited articles attracting ever-increasing numbers of citations – including, and perhaps especially, those that are erroneous?
- Would it be better to exclude self-citations from the lead author or his/her research group?

The last question is important in the PBRF environment and also impacts more on how scientists should present their individual citation counts (see below). Individuals are often encouraged to purge self-citations from the records that they supply for assessment exercises. This is viewed as somehow noble or at least as being the best and most ethical practice by avoiding the temptation to inflate scores by citing one's own work. Again it is not quite that simple. On the one hand, if a person wishes to show the extent to which they may have influenced others, then by all means exclude self-citations and those from the same research group. On the other hand, if one wishes to show the extent to which an article has influenced science itself, then self-citations etc. should stay in (if one can assume that authors really did behave ethically in quoting their own articles in the first place).

Indeed, it is clear that the JIF has probably become the most frequently abused bibliometric statistic. Managers often, even routinely, encourage scientists to attach JIF values to their PBRF Evidence Portfolio articles despite the inventor's insistence that this is not appropriate (see above) and as echoed by others too, e.g. by Seglen (1997) for medicine. Brischoux & Cook (2009) protest the tyranny of JIF for junior staff, reflecting similar recent comments and warnings from others (Cherubini 2008; Notkins 2008) who continue a well established critical dialogue (Colquoun 2003; Lawrence 2003). In addition, Cameron (2005) provides a librarian's perspective on the use of this tool. In short, the JIF attached to a journal says nothing for certain about the author(s) or their particular paper, except that they have convinced the editor to publish it in an outlet that has high visibility and that one suspects is also high quality with competitive entry. These latter two qualities are not guaranteed by high JIFs, since scores may depend on the composition of the journal. Those with a larger proportion of review articles may have higher, or even artificially inflated, JIF values.

Where should I submit my next paper?

This is perhaps the most important and challenging question for scientists aiming to enhance their PBRF profiles. It will be instructive to inquire at some later stage to what extent scientific publishing practice in New Zealand has actually changed in response to the introduction of this scheme. Hendy (2010) has shown that output has been static from 1995 to 2008 at the surprisingly low values of c. 0.53 papers/FTE/year for New Zealand university research staff [cf. around 0.75 for Crown research institutes, (CRIs)]. These numbers may undervalue university academic staff because graduate students are included in the FTE calculation. It is perhaps a matter of concern, or at least regret, that the student to staff ratio has declined from 3.4 to 2.7 over this period, which might be interpreted to suggest that the actual publishing outputs of universities themselves may have fallen by approximately 20% over this period as they are training fewer young researchers. It will be interesting to see if this has resulted in improved teaching quality. Interestingly enough, citation counts for both university and CRI scientists have risen steadily during this interval from 1.0 to 1.8 (confusingly called an *impact factor* by Hendy (2010) and not to be confused with JIF). However, these increases must be attributed to the appearance of increasing numbers of journals and articles between 1995 and 2008 rather than the effect of PBRF because the increase is seen in both sectors.

Questions about journal selection also seem to be a rather more widespread issue, and the potential of real or imagined influences of bibliometric statistics on author behaviour are now causing concern (see Lawrence 2003). It would appear to be sensible, even dutiful, to choose publications that might improve one's PBRF ratings in terms of bibliometric statistics. However, simply picking journals with high JIF scores may not be the way to go, and some alternatives are explored below. Further, there is a clear tension between trying to get a paper into high-exposure journals versus a conventional *best-fit* option. In other words, how should one decide between publishing in journals that have high research visibility and journals where the publication may influence science practice, for example in professional journals and local journals? In some instances, moving up to the middle ground and submitting an article to a journal that is a recognised leader in the field may be the preferred answer. At other times sticking with a lower-ranked national journal might be a better way to go, and in our experience often is the right choice for biologists. There is a clear trade-off between quantity and apparent quality. A failed submission to a high-status journal comes at a significant cost (see Lawrence 2007 for a poignant description of the process) which may preclude sending in other articles or seriously delay, or at worst permanently deflect, publication of the original work. It may be that those who do publish in prestige journals publish fewer papers overall (see Brashier *et al.* 2005). Thus, the trade-off seems to become that between being a *team player* and supporting the objectives of employers by trying to secure high PBRF scores, and being a *good citizen* with an obligation to account to the New Zealand public for funds invested by them in the original research work by making the results available for scrutiny in a peer-reviewed source. In the end, scientists probably know what is best for them, not their managers, but it probably would not hurt anyone to get a bit more ambitious once in a while.

A related question arises in connection with authorship and papers where scientists are first (or better still sole) author. These are widely thought to be among those most prized by PBRF panels. Generally speaking, it is, or should normally be, the person who did most of the lab/field work plus perhaps the data analysis and who almost certainly wrote the manuscript that is first author. With a few notable exceptions this is impractical for most senior scientists in academic institutions who most often function as team leaders, rather than bench workers. In some circumstances they may take over the first author role if a major investigator cannot, or will not, write up their original study or if data are being combined from several sources and the team leader does the write up. Review articles are a different matter and senior researchers often take lead or sole authorship on these. Such work does attract high citations, but may not be so highly regarded by PBRF panels and may not be included among Nominated Research Outputs (NRO), despite being products of quite extensive scholarship and significant effort. This seems to be particularly perverse when the PBRF is apparently trying to promote *excellence* and *paradigm changing* science. Reviews and syntheses often provide new insights and can pull together disparate threads in a field and/or make connections across disciplines.

Citation counts and summary statistics

If one elects not to use JIF scores what then are the options? A simple and direct answer would be to use the citation counts associated with each article individually as evidence that they are receiving attention. For instance, it is well known (Garfield 2005) that at least half of all scientific articles are never cited at all, even by their own authors. Hence, if a paper receives any citations at all, it would be fair to claim that it rates as being in the top 50% of all published work. This may not be fully effective for some PBRF purposes, as their census interval of six years is quite short compared with the citation lifetime of most papers (www.isiwebofknowledge.com). Older *citation classic* type articles can be included in the Peer Esteem (PE) or Contributions to the Research Environment (CRE) sections of PBRF portfolios. Further, it may sometimes be useful to roll a full or part publication record into a single summary statistic that allows comparisons with others in the same discipline. The following sections review the various tools for doing both of these tasks and offer advice on exactly how to go about using such tools based on authentic individual experiences. Before starting out, we warn that bibliometric statistics can be very hazardous things. First, one may make an error and underestimate the value of the publication record. Second, and worse still, one may overestimate it and thereby lose credibility. So there is constantly the very real risk that somebody on an assessment panel will check up on values presented to them and come up with different numbers by inadvertently using different tools or search procedures. Such an experience could easily erode the assessor's confidence in the candidate. Hence, the most practical advice that can be offered at the outset of any such exercise is to: (1) do it the best way possible for each individual system; (2) explain exactly what was done; and (3) state when the data were first recovered.

There are three widely consulted bibliometric databases; see Box C. Each one of these databases will yield a list of papers published by particular authors, citation counts for individual articles plus summary statistics that are either generated automatically or which can be constructed by hand from the lists and counts. These are defined in Box D. There are other databases such as PubMed or Aquatic Sciences and Fisheries Abstracts that are field-specific and other less well known summary indices which will not be considered here (see Browman & Stergiou 2008 and other papers in this volume of *Ethics in Science and Environmental Politics*).

Practical applications of databases

In Tables 1–3, data have been tabulated for five researchers representing various career stages and various disciplines broadly focused on biology and ranging from medicine to palaeontology. We also invite readers to go on line and check their own records and see if their experiences accord with those described below. If one is feeling mischievous, one can look up records for Deans or Heads of School, or even rivals for employment or promotion, to discover if justice has been done in terms of comparative research records. Finally, people can assess their imagined standing in the field at large by running comparisons with global leaders or directors of New Zealand's Centres of Research Excellence. The data are all there to do such things and all such searches will almost inevitably turn up a few surprises.

Those who do elect to interrogate the online databases need to be very careful with respect to the search parameters that they enter and to check the outputs thoroughly; see Table 1 for typical results. Records may be missed or duplicated due to variations in the spelling of names and inclusion or otherwise of initials. One also needs to check institutional affiliations carefully to make sure that outputs all come from locations where the search subject has worked in the past and that all such locations are included for a full career record, cf. single employer tally or PBRF census. Equally, one needs to exclude references to those others who have similar names and/or work histories. In general experience, Web of Science and Scopus now seem to perform equally well in this respect. Both of them capture around 60–70% of an individual's published outputs and duplicates or work by other investigators that must often be removed by hand. This process is more easily achieved working in Scopus as individual records can be excluded. However, in fairness to both WoS and Scopus it seems that precision of performance is improving all the time. Google Scholar is seriously polluted with duplicates and cannot be used in isolation for this purpose (e.g. Smith 2008 in a recent discussion of the utility of Google Scholar in the PBRF context). The great advantage that Google Scholar does enjoy is that its output includes citations in books. This is something that neither WoS nor Scopus can do because they search only journals and some conference papers. So one could locate these extra citations using Google Scholar and add them by hand to other scores – given that one has the time to do such things. Finally, readers should be aware that Scopus only counts citations back to 1996, although it will list quite a large fraction of articles published before this date.

It is well recognised that the output counts from Scopus and WoS are strongly correlated (Harzing & van der Wal 2008) so either may be taken alone as an index of citation count. What is also widely recognised by bibliometric experts is that the lists

Box D: Bibliometric summary statistics

The *h-index* (Hirsch 2005) is the maximum number of publications that have the same number of citations. Please note that Scopus returns two versions of this metric: *Author h-index* and *Citation h-index* – see text for details. The WoS database returns only one, which is their equivalent to the *Citation h-index*, but it can be made to output a comparable version of the *Author h-index* by setting the temporal limits on the search field to begin from 1996.

The *g-index* (Egge 2006) is obtained by counting up the first *g* articles that have g^2 citations.

The citation counts returned by both WoS and Scopus are derived from around 110 000 linked journals, i.e. many more than are searched for the source author publication lists. They include historical records, including even those before 1996 for Scopus.

Table 1. Database returns for five New Zealand scientists.

Individual	Scopus			Web of Science		
	Raw	Removed	Total	Raw	Removed	Total
A	96	5	81	121	31[a]	90
B	94	6[b]	88	100	6[b]	94
C	85	0	85	109	2	107
D	51	1	50	59	1	58
E	10	0	10	9	3	6

The values in the body of this table are numbers of papers on the publication list for that individual. The scientists included in the survey include three senior academics: A – full time medical researcher; B – academic staff member with interests in medical and biological topics; and C a conservation geneticist. Individual D is a mid-career marine biology researcher and E is an earth scientist and palaeontologist. They were specially selected as illustrations to ensure representation across fields in a single discipline (nominally biology).

[a] These publications include a large number of single paragraph conference abstracts.

[b] The Scopus entries include 5 publications belonging to another scientist and one duplication compared with the WoS return, which includes 2 bogus entries and 4 minor publications.

Table 2. Set theory analysis of database returns for five New Zealand scientists.

Individual	Scopus	Common	Web of Science	Σ Global	Σ Citations
A	91	83 (91.7)	90	98 (108.3)	2503
B	88	76 (83.5)	94	106 (116.5)	2757
C	85	75 (78.1)	107	117 (121.9)	1380
D	51	47 (86.2)	58	62 (113.8)	656
E	10	3 (37.5)	6	13 (162.5)	120

The values in the body of the table are publication counts in each class and the numbers in parentheses show the values as a percentage of the arithmetical average of the Scopus and WoS scores.
The Σ Citations value is calculated by taking the Σ WoS citation score and adding to it the total citations recorded in Scopus for those publications captured by Scopus but not by WoS.
Self-citations have **not** been removed for these lists. The effect of doing so is to reduce counts and *h-index* values by up to 10–20% (data not shown).

Table 3. Bibliometric statistics for five New Zealand scientists.

Individual	Scopus				Web of Science			h_{global}
	Papers	Citations	h_{author}	h_{citation}	Papers	Citations	h-index	
A	91	1658	16 (54)	20 (91)	90	2411	26	27
B	88	1394	18 (48)	20 (88)	94	2710	27	27
C	85	1145	14 (70)	18 (85)	107	1339	19	19
D	50	424	10 (46)	13 (50)	58	633	15	16
E	10	123	4 (10)	4 (10)	6	34	2	5

Figures in parentheses after *h* values are numbers of publications from which these values are obtained.

of articles returned by WoS and Scopus overlap considerably, but not completely; see Table 2 where typical values are in the neighbourhood of 80%. Hence, the combined set of articles recovered from the two databases will nearly always be higher by up to 20% than that produced by either search individually. To these counts might be added whatever else in the way of articles and citations might be lurking in the reams of output from Google Scholar. In contrast to the lists of articles, the contents of the WoS and Scopus citation pools may only overlap by around 60% (see also Harzing & van der Wal 2008 for an in-depth commentary) – data from Table 1 were not examined in detail in this respect when preparing our present article but this fact is readily apparent from the scores in Table 2.

It is clear that an opportunity exists for some enterprising webpage software engineer to use a set theory-based approach to combine these outputs and create a comprehensive and accurate citation history for authors and articles. Note that in order to achieve the requisite accuracy the service will have to remove orphan and redundant citations within and between datasets (about 5%, to judge from the data in Table 1). The proposed software would be of widespread utility as the only present alternative is to do such things by hand. If the screening part of the system were to be fast and efficient, it might even be able to recover the missing citations arising from incorrect spelling or poor citation by relaxing the search parameters.

Practical applications of summary statistics

Hirsch (2005) invented the *h-index* (see Box D) which has become the *one size fits all* bibliometric statistic of choice for many scientists. Indeed, it is now often bandied about like a golf handicap. Scientists may find themselves put on the spot by managers who wish to know the *h-index* status of interviewees. So, it may be best to know what these things are and how to recover them properly from the online databases. The *h-index* certainly has a simple enough definition (the maximum number of publications that have the same number of citations – *thus an h-index of 14 means that the author has published at least 14 papers each of which has been cited at least 14 times*), but it often proves to be one that is hard to keep an exact mental grasp on. The idea is that the best scientists publish lots of papers and these get lots of attention. Failure with regard to either of these two desirable properties is not good. Hence, high values for the *h-index* are taken as being characteristic of the very best of scholars. The obvious problem is how to judge those who have aberrant *h-index* scores: the person who has published 50 papers (aka *the unspectacular plodders*), but never received more than 3 citations for any one of them v. the person who only ever published three papers, but who nonetheless may have changed our view of the world (aka *flash-in-the-pan scholars* or worse *cold-fusion types*) – both have the same *h-index* score of just 3 even though the latter scholar may have many thousand citations to their work. The *g-index* (Egge 2006) is intended to remove such distortions – any *g-index* value greater than 10 would be excellent for a biologist (refer to Box D for a definition of the *g-index*). However, it should already be clear that it is unsafe to quote an *h* value or *g-index* unless one knows exactly how to obtain them (see below) and that they should always be accompanied by a wider view of the citation profile on which they were based, e.g. as done for David Penny (by Berridge & Petrey 2009) and see Table 3. The potential effect of *h-index* scores on future careers has been examined (Kelly & Jennions 2006) and we will return to question the value of bibliometrics to assess scientific creativity later.

A quick visit to Scopus and running an *Author Search* will return a carefully sanitised minimum publication list (see earlier). An instant *h-index* can be viewed by activating the *Citation Tracker* facility. This is actually the *Author h-index* (see Box D) although it is not shown as such on screen. On no account should one use this number in isolation because it only relates to publications going back to 1996! That is unless you are a new investigator, but even then approach with caution as it may be lower than the value indicated by visual examination of the citation scores on your publication list. Scopus claim that this form of the *h-index* provides the fairest comparative metric as it has a wide (currently fifteen year) census window and relates to more recent performance. While this may seem reasonable enough to many minds, it does scant justice to the record of mid-career and older scientists who may have achieved significant recognition prior to the Scopus start date. So, for those who might have published papers before 1996, one needs to operate the *Show Documents* function and then select **all** entries on **all** pages by ticking boxes before activating the *Citation Tracker* function. This will now return a new and, one hopes, higher *h-index* value (see Box D) based on the entire publication list, but still only counting citations back to 1996. Hence, this too is still only a minimal estimate; see Table 3 for some examples.

A similar exercise using WoS usually returns an even higher *h-index* value, as the citation scores include references in papers published before 1996. The increase in magnitude of the *h-index* value is most marked for those with long research records; see data for Scientists A and B in Table 3 compared with say Scientist D. This is not always the case, and the *h-index* score for Scientist C only goes up by one point. Finally, one might be able to make even further progress by adding in by hand any extra papers with high citation rates that were captured by Scopus and not included on the WoS list. Alternately one can run a *cited reference search* from the WoS window. This can recover additional references and citations not captured on the original WoS results page. Another way to achieve similar goals is by using Web of Knowledge (WoK) rather than WoS. The composite WoK database contains all of the WoS source databases plus several others. In some cases WoK can return significantly higher publication counts for individuals than WoS. However, although both WoK and WoS have *Analyse Results* function buttons which operate on the scores returned, **only** WoS has the all-important *Create Citation Report* button, which returns *h-index* scores, plus a full temporal record of publications. It is feasible to perform such wider search operations by hand, but the rewards are often meagre (the largest increase in Table 3 is only one point), and probably not many scientists will have the patience to filter the actual citations for the sake of adding one or two to their *h-index* scores.

Only very few workers are likely to be able to find the time to collate up-to-date spreadsheets showing WoS and Scopus citation scores for all of their publications, even though it might be a wise thing to do. All of the preceding observations serve to show the need for caution in these affairs. Judgements and decisions can only hope to be as good as the data upon which they are based. Those who wish to use the *h-index* for evaluations should bear in mind that, like any metric, the index number has inherent inaccuracies due to citation variations, etc. Also it is extremely unwise to compare *h-indices* (or any other citation measures for that matter) that have been calculated using different databases and/or different methodologies, or to compare *h-indices* between different disciplines: different disciplines have different practices. Furthermore, the *h-index* is an accumulating statistic so that a long-time researcher will necessarily have a higher *h-index* than a newer researcher of equal ability. It is also quite difficult to calculate *h-index* values at previous points in time, so that an institutional requirement for a staff member to regularly increase their personal *h-index* at some predetermined rate will be difficult to audit retrospectively. Anxiety driven by unthinking administration of such targets by managers may even tempt some desperate individuals to use strategic self-citation in manuscripts to push the citation counts for selected past papers over the critical count value required to raise their *h-index* by one or two points.

How can scientists best protect their reputations?

It is incumbent upon working scientists in New Zealand to support our institutions by finding the best, most ethical and safest way to advertise the quality of their research records by means of bibliometrics. The system demands it. The core problem is that no tool is perfect and no statistic tells the whole story, or even perhaps the whole truth. However, at least the presently available tools do all seem to agree with one another to a first approximation. Doing it all the strictly correct way, i.e. exhaustively by hand, is far too time-consuming. So a secure heuristic approach is required, i.e. in the form of a commonly agreed index of research output and impact (true quality is a very different kettle of fish – see below). Finally, it comes down to reporting. Here best practice would seem to involve quoting the highest *h-index* values obtained from Scopus and/or WoS as required. These should always be accompanied by database publication counts (including as a percentage of *curriculum vitae* publication count) and citation count(s) including specific figures for all publications with over 100 citations (or perhaps just the five highest counts if, like investigator C, none of the papers has received this many). It is essential to include exact details of the search procedure that produced the data, i.e. pretty much following the fine example set by Berridge & Petrey (2009). Note that these procedures are valid for presenting career achievements and as some sort of measure of the person as a scientist. They are not helpful for comparing achievements over the very short PBRF census window, because too few citations will be accumulated.

In all of the preceding we have taken for granted that citation counts are a fair reflection of valid citations and good practice. This is necessary because it is well beyond all reasonable practical means to check all one's citations for accuracy. However, when this has been done carefully (Todd *et al.* 2010) it becomes apparent that as many as 25% of citations may be inappropriate. This serves to inflate citation counts across the board and may even compensate to some extent for omissions due to ignorance, professional envy, etc. True omission rates must probably remain unknown as there seems to be no obvious way to estimate them. Many scientists do feel that their work has been unjustifiably overlooked especially by particular rivals, but again there is probably no reliable way to judge if this is a fair assessment on their part.

Of course, the general citation assessment exercise will only work if everyone agrees to do it as recommended above, or perhaps better still if at least they can be evaluated independently to ensure a standard methodology. The alternative is anarchy, featuring random quotation of dubious impact factors accompanied by bogus statistics of uncertain origin. This is pretty much the *status quo*, in fact. There is one further possibility and that is to compare one's citation profile with the WoS global or discipline-specific profile. Hence, one might be able to claim that the publication sample obtained as their own database output maps on to the 90–95th citation percentile interval in the global distribution (www.isiwebofknowledge.com). Here, the career evaluation will still not be comprehensive, because not everything on the *curriculum vitae* is included, but at least it will be a large and high-quality sample, given WoS claims regarding the nature of its database. However, this type of approach can be applied in principle to subsets of publications (e.g. between institutions and/or intervals of years) provided that each set contains a reasonably representative sample of citations.

Conclusions

To date, scientists may all have thought they knew what they were doing when they sent in numbers to PBRF, etc., but did they really? This article has shown just how careful one has to be when collecting supportive bibliometric statistics and includes some suggestions for best-practice reporting. This is not to suggest that failure to follow the advice set down here will be damaging to otherwise promising careers, but the authors do feel that it provides a measure of added security. The world of science will have to await the development of new software and the adoption of standard reporting practice before assessors can be fully confident with respect to the process.

We have left aside the deeper questions of whether it is fair, decent, or even sensible to assess careers in this fashion, because those working in New Zealand simply must respond in some way. For instance, how are we to judge participation in huge, long-term, multi-disciplinary and multi-centre studies which may be required in a rigorous clinical trial and which may only ever produce one major report? Other distinguished writers have taken up this theme. For instance, in the field of molecular biology Lawrence (2007) has shown how poorly Watson & Crick's work on DNA and Ed Lewis's pioneering studies of *Drosophila* would have fared under such a regime. Noble's (2010) review of Gillies (2008) builds the bigger picture. Readers are encouraged to consult these sources to learn how research progressed in a previous and perhaps happier and more enlightened scientific environment than under the well intentioned inquisitors of the PBRF.

Acknowledgements

We thank Mike Hannah (VUW, Earth Sciences), Kim Sterelny (VUW Philosophy) for encouragement, Wayne Linklater (VUW, Biology) and Sue d'Ath (VUW, Library) for help with resources and Katy Miller (VUW, Research Office) for constructive criticism. We are grateful for valuable insights and advice from the Science Review Editorial Committee and are particularly indebted to Mike Berridge (Malaghan Institute) for sharing his personal experience of and scholarly observations on the *h-index* in theory and practice.

References

ARC 2008. Tiers for the Australian ranking of journals. Published online at www.arc.gov.au/era/tiers_ranking.htm

Bergstrom, C. 2007. Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News 68*: 314–316.

Berridge, M.; Petrey, A. 2009. Editorial. *New Zealand Science Review 66*: 1.

Browman, H.I.; Stergiou, K.I. (Eds) 2008. The use and misuse of bibliometric statistics in evaluating scholarly performance. Volume 8 Theme Section in *Ethics in Science and Environmental Politics*. Inter-Research, Oldendorf.

Brashier, T.; Mathew, I.; Symonds, R.E.; Gemmell, N.J. 2005. Publication in *Science* and *Nature* is not gender dependent. *BioEssays 27*: 858–859.

Brischoux, F.; Cook, T.R. 2009. Juniors seek an end to the Impact Factor race. *BioScience 59*: 858–859.

Butler, D. 2008. Free journal-ranking tool enters citation market. *Nature 451*: 6.

Cameron, B.D. 2005. Trends in the usage of ISI bibliometric data: Uses, abuses and implications. *Libraries and the Academy 5*: 105–125.

Campbell, P. 2008. Escape from the impact factor. Pp. 5–7 in Browman, H.I.; Stergiou, K.I. (Eds) *Ethics in Science and Environmental Politics* Volume 8. Inter-Research, Oldendorf.

Cherubini, P. 2008. Impact Factor fever. *Science 322*: 191.

Colquoun, D. 2003. Challenging the tyranny of impact factors. *Nature 423*: 479.

Egge, L. 2006. Theory and practice of the g-index. *Scientometrics 69*: 131–152.

Garfield, E. 1996. How can Impact Factors be improved? *British Medical Journal 313*: 411–413.

Garfield, E. 2005. The agony and the ecstasy – The history and meaning of the Journal Impact Factor. *Proceedings of the International Congress of Peer Review and Biomedical Publication.* Chicago available from www. eugenegarfield.org

Gillies, D. 2008. *How should research be organised?* College Publications, London.

Harzing, A.-W.; van der Wal, R. 2008. *Google Scholar as a new source for citation analysis.* Pp. 61–73 in Browman, H.I.; Stergiou, K.I. (Eds) *Ethics in Science and Environmental Politics* Volume 8. Inter-Research, Oldendorf.

Hendy, S.C. 2010. New Zealand's bibliometric record in research and development: 1990–2008. *New Zealand Science Review 67*: 56–59.

Hilborn, R. 2006. Faith-based fisheries. *Fisheries 31*: 554–555.

Hirsch, J.E. 2005. An index to quantify an individual's research output. *Proceedings National Academy of Sciences USA 102*: 16589–16572.

Jiao, L. 2009. China searches for an 11th-hour lifesaver for a dying discipline. *Science 325*: 31.

Kelly, J.K.; Jennion, M.D. 2006. The h-index and career assessment by numbers. *Trends in Ecology and Evolution 21*: 167–170.

Lawrence, P.A. 2003. The politics of publication. *Nature 422*: 259–261.

Lawrence, P.A. 2007. The mismeasurement of science. *Current Biology 17*: R583–585.

Middleton, S. 2009. Becoming PBRF-able: Research assessment and Education in New Zealand. Pp. 193–208 in Besley, A.C. (Ed.) *Assessing the Quality of Educational Research in Higher Education.* Sense Publishers, Rotterdam.

Noble, D. 2010. Funding the pink diamonds: A historical perspective. *Notes and Records of the Royal Society 64*: 97–102.

Notkins, A.L. 2008. Neutralising the Impact Factor culture. *Science 322*: 191.

Power, M. 1997. *The Audit Society: Rituals of Verification.* Oxford University Press, Oxford.

Roa, T.; Beggs, J.R.; Williams, J.; Moller, H. 2009. New Zealand's Performance Based Research Funding (PBRF) model undermines Māori research. *Journal of the Royal Society of New Zealand 39*: 233–238.

Seglen, P.O. 1997. Why the Impact Factor of journals should not be used for evaluating research. *British Medical Journal 314*: 497–511.

Smith, A.G. 2008. Benchmarking Google Scholar with the New Zealand PBRF research assessment exercise. *Scientometrics 74*: 309–316.

Thomaz, S.M.; Martens, K. 2009. Alternative metrics to measure journal impacts: Entering in a "free market" era. *Hydrobiologica 636*: 7–10.

Todd, P.A.; Guest, J.R.; Lu, J.; Chou, L.M. 2010. One in four citations in marine biology papers is inappropriate. *Marine Ecology Progress Series 408*: 299–303.