Statistical Science
2009, Vol. 24, No. 1, 15–16
DOI: 10.1214/09-STS285A
Main article DOI: 10.1214/09-STS285

Comment: Bibliometrics in the Context of the UK Research Assessment Exercise

Bernard W. Silverman

Abstract. Research funding and reputation in the UK have, for over two decades, been increasingly dependent on a regular peer-review of all UK departments. This is to move to a system more based on bibliometrics. Assessment exercises of this kind influence the behavior of institutions, departments and individuals, and therefore bibliometrics will have effects beyond simple measurement.

Key words and phrases: Bibliometrics, research funding, perverse incentives.

In the United Kingdom's Research Assessment Exercise (RAE), every university may submit its research in every discipline for assessment. On this assessment rests a considerable amount of funding; indeed a number of universities, leading "research universities" in American nomenclature, gain more from this source of research funding than from government funding for teaching. Within broad subject bands, the Higher Education Funding Council for England funds teaching on a flat rate per student. So the amount of funding a student of Mathematics attracts is the same whichever university they attend. On the other hand, funding for research is selective: those departments which fare well on the Research Assessment Exercise receive more funding as a result. This is in addition to any income from grants and grant overheads.

The RAE and its predecessors have been running for over two decades, and have always been based on peer review, though numerical data on student numbers and grant income also have some input into the assessment. However, it is proposed that "metrics," which include so-called bibliometric data, will be the main part of the system which will soon succeed the RAE, though it is probable that in mathematical subjects, peer review will continue to play a considerable part. The details have yet to be worked out.

In the 2008 RAE, I was chair of the committee which reviewed Probability, Statistics and the more mathematical aspects of Operational Research. The committee's experience of conducting the assessment as a whole strengthened our view that peer review must be at the core of any future assessment of research in our area. Reliance on bibliometric and purely quantitative methods of assessment would, in our unanimous view, introduce serious biases, both into the assessment process and, perhaps more seriously, into the behavior of institutions and of individual researchers, to the detriment of the very research which the exercise is intended to support.

It is important to stress the effect of any system on the behavior of institutions. The current peer-review RAE has had clear effects on institutional behavior, some of them certainly positive, some of them perhaps less so. For example, the RAE gives explicit advantages to new entrants to the profession; those entering in the last few years are allowed to submit a smaller corpus of work for assessment, and there is also credit given within the peer review system for a subjective assessment of the general vitality of the department. Of the approximately 400 research-active faculty declared to the statistics panel in the 2008 RAE, about a quarter were new entrants since 2001, and the RAE has certainly given an impetus to this new recruitment, as it also does to the mobility of leading researchers between institutions. On a more negative note, the fixed date of the assessment encourages a "boom-bust" mentality, where some institutions hire in considerable numbers of new faculty in the period leading up to the census date; to make up for this extra expenditure, during the period after the census date there is something

B. W. Silverman is Master, St Peter's College, Oxford OX1 2DL, United Kingdom (e-mail: bernard.silverman@spc.ox.ac.uk).

of a moratorium on appointments. The consideration of grant income in the RAE gives extra gearing to the pressure on faculty to pursue grant-supported research rather than to work in a more individual fashion.

There can be little doubt that a stronger emphasis on bibliometrics (and other "metrics") in assessment exercises will affect institutional behavior, especially in systems where assessment results have both reputational and fiscal impact. Because individuals are sensitive to institutional pressures, they too will modify their behavior in response. For example, it is probably the case that there is a high correlation between h-index (say) and perceived quality and reputation of researchers. Similarly, highly-cited papers are almost always influential and important (though the converse is not necessarily true). However, basing judgment of individuals or departments on citation count rather than some assessment of underlying quality would have the obvious perverse consequences. Perhaps the obvious analogy would be with a system that counts publications: of course there is some correlation between the overall quality of a researcher's work and the number of papers she or he publishes, but the "publish or perish" mentality engendered by simple paper-counting militates against the careful and thoughtful researcher who only writes papers when they feel they have something very serious to say, or—worse still—writes books rather than papers. Perhaps the bibliometric version is "be cited or benighted"?

One of the arguments the UK university funding agencies used initially in favor of bibliometrics was that, when aggregated over whole universities, the results of "metrics-based" assessments were very highly correlated with peer-review assessments. As statisticians, we should be well placed to refute the fallacy of this argument. It makes complete sense that a strong university will have more than its fair share of highly-cited researchers right across the range of disciplines. Any errors and biases will to some extent average out. But disaggregation down to departments, and even individuals, encourages the elimination, or downgrading, of disciplines and sub-disciplines which do not generate large amounts of citations. Within disciplines, there is a risk of undervaluing individuals whose work is deeply influential in ways that do not show up in short-term citation counts. And many individual researchers would no doubt bow to perceived pressure to be "cited or benighted."

If citation counts are unreasonable, the use of impact factors seems almost indefensible. Assigning a notion of quality to a paper on the grounds of the impact factor of a journal is like assigning a notion of wealth to an individual on the basis of the average GDP of their home country. Many children growing up in England in the 1950s were under the impression that all Americans were wealthy! Of course, if one knows about the refereeing standards of a leading journal, it may, or may not, be reasonable to suppose that if a paper has passed these standards it has a good chance of being of high quality, but that is a very different thing from assessing the journal by an impact factor.

In conclusion, I would very strongly support the underlying thesis of the paper: citation statistics, impact factors, the whole paraphernalia of bibliometrics may in some circumstances be a useful servant to us in our research. But they are a very poor master indeed.

Statistical Science
2009, Vol. 24, No. 1, 17–20
DOI: 10.1214/09-STS285B
Main article DOI: 10.1214/09-STS285

Comment: Citation Statistics

Sune Lehmann, Benny E. Lautrup and Andrew D. Jackson

Abstract. We discuss the paper "Citation Statistics" by the Joint Committee on Quantitative Assessment of Research. In particular, we focus on a necessary feature of "good" measures for ranking scientific authors: that good measures must able to accurately distinguish between authors.

Key words and phrases: Citations, ranking.

1. INTRODUCTION

The Joint Committee on Quantitative Assessment of Research (the Committee) has written a highly readable and well argued report discussing common misuses of citation data. The Committee argues convincingly that even the meaning of the "atom" of citation analysis, the citation of a single paper, is non-trivial and not easily converted to a measure of research quality. The Committee also emphasizes that the assessment of research based on citation statistics always reduces to the creation of *ranked lists* of papers, people, journals, etc. In order to create such a ranking of scientific authors, it is necessary to describe each author's full publication and citation record to a single scalar measure, $\mathcal{M}$. It is obvious that any choice of $\mathcal{M}$ that is independent of the citation record (e.g., the number of papers published per year) is likely to be a poor measure of research quality. However, it is less clear what constitutes a "good" measure of an author's full citation record. We have previously discussed this question in some detail [1, 2], but in the light of the present report, the subject appears to merit further discussion. Below, we elaborate on the definition of the term "good" in the context of ranking scientific authors by describing how to assign objective (i.e., purely statistical) uncertainties to any given choice of $\mathcal{M}$.

Sune Lehmann is Postdoctoral Researcher, Center for Complex Network Research, Department of Physics, Northeastern University, Boston, Massachusetts, USA and Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Harvard University, Boston, Massachusetts, USA (e-mail: lehmann@neu.edu). Benny E. Lautrup is Professor of Theoretical Physics, The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark. Andrew D. Jackson is Professor of Physics, Niels Bohr International Academy, The Niels Bohr Institute, Copenhagen, Denmark.

2. IMPROBABLE AUTHORS

It is possible to divide the question of what constitutes a "good" scalar measure of author performance into two components. One aspect is wholly subjective and not amenable to quantitative investigation. We illustrate this with an example. Consider two authors, A and B, who have written 10 papers each. Author A has written 10 papers with 100 citations each and author B has written one paper with 1000 citations and 9 papers with 0 citations. First, we consider an argument for concluding that author A is the "better" of the two.

In spite of varying citation habits in different fields of science, the distribution of citations within each field is a highly skewed power-law type distribution (e.g. see [3, 4]). Because of the power-law structure of citation distributions, the citation record of author A is more improbable than that of author B. It is illuminating to quantify the difference between the two authors using a real dataset. Here, we use data from the SPIRES database for high energy physics (see [3] for details regarding this dataset). The *citation summary* option on the SPIRES website returns the number of papers for a given author with citations in each of six intervals. These intervals and the probabilities that papers will fall in these bins are given in Table 1. The probability, $P(\{n_i\})$, that an author's actual citation record of N papers was obtained from a random draw on the citation distribution $P(i)$ is readily calculated by multiplying the probabilities of drawing the author's number of publications in the different categories, n_i, and correcting for the number of permutations.

$$P(\{n_i\}) = N! \prod_i \frac{P(i)^{n_i}}{n_i!}.$$

If a total of N papers is drawn at random on the citation distribution, the most probable result, $P(\{n_i\}_{\max})$,

TABLE 1
The search option citation summary *at the SPIRES website returns the number of papers for a given author with citations in the intervals shown. The probabilities of getting citations in these are intervals are listed in the third column*

Paper category	Citations	Probability $P(i)$
Unknown papers	0	0.267
Less known papers	1–9	0.444
Known papers	10–49	0.224
Well-known papers	50–99	0.0380
Famous papers	100–499	0.0250
Renowned papers	500+	0.00184

corresponds to $n_i = NP(i)$ papers in each bin. The quantity

$$r = -\log_{10}\left[\frac{P(\{n_i\})}{P(\{n_i\}_{\max})}\right],$$

is a useful measure of this probability which is relatively independent of the number of bins chosen. In the case of author A we find the value $r_A = 14.4$, and for author B we find $r_B = 5.33$. In spite of the fact that A and B have the same average number of citations, the record of author B is roughly 10^9 times more probable that that of author A! We do not claim that the "unlikelihood" measure r captures the richness of the full data set nor that it captures the complexity of individual citation records,[1] but we do claim that the extreme improbability of author A might convince some to choose her over author B.

On the other hand, if one believes that the most highly cited papers have a special importance—that they contain scientific results that are particularly significant or noteworthy—one might reasonably prefer B over A. A famous proponent of this view is the father of bibliometrics, E. Garfield, who dubbed such papers "citation classics" [5]. No amount of quantitative research will convince a supporter of citation classics that the improbability of a citation record is a better measure of the scientific significance of an author or vice versa; this judgment is strictly subjective.

3. DISCRIMINATORY ABILITY

As mentioned above, scalar measures of author quality also contain an element that can be assessed objectively. Whatever the intrinsic and value-based merits of the measure, $\mathcal{M}$, assigned to every author, it will be of no practical value unless the corresponding uncertainty, $\delta\mathcal{M}$ is sufficiently small. From this point of view, the "best" choice of measure will be that which provides maximal discrimination between scientists and hence the smallest value of $\delta\mathcal{M}$. If a measure cannot be assigned to a given author with suitable precision, the subjective issue of its relation to author quality is rendered moot. Below we outline how the question of deciding which of several proposed measures is most discriminating, and therefore "best," can be addressed quantitatively using standard statistical methods.

The model that authors A and B draw their citation records on the total citation distribution $P(i)$ is quite primitive. This is indicated by the fact that the numerical values of r for both A and B are uncomfortably large. It is more reasonable to assume that each author's record was drawn on some sub-distribution of citations. By using various measures of author quality to construct such sub-distributions, we can gauge their discriminatory abilities. We formalize this idea below.

We start by binning all authors according to some tentative indicator, $\mathcal{M}$, obtained from their full citation record. The probability that an author will lie in bin α is denoted $p(\alpha)$. Similarly, we bin each paper according to the total number of its citations.[2] The full citation record for an author is simply the set $\{n_i\}$. For each author bin, α, we then empirically construct the conditional probability distribution, $P(i|\alpha)$, that a single paper by an author in bin α will lie in citation bin i. These conditional probabilities are the central ingredient in our analysis. They can be used to calculate the probability, $P(\{n_i\}|\alpha)$, that any full citation record was actually drawn at random on the conditional distribution, $P(i|\alpha)$ appropriate for a fixed author bin, α. Bayes' theorem allows us to invert this probability to yield

$$P(\alpha|\{n_i\}) \sim P(\{n_i\}|\alpha)p(\alpha), \tag{1}$$

where $P(\alpha|\{n_i\})$ is the probability that the citation record $\{n_i\}$ was drawn at random from author bin α. By considering the actual citation histories of authors in bin β, we can thus construct the probability $P(\alpha|\beta)$, that the citation record of an author initially assigned to bin β was drawn on the distribution appropriate for bin α. In other words, we can determine the probability that an author assigned to bin β on the basis of the tentative indicator should actually be placed in bin α. This allows us to determine both the accuracy of the initial author assignment and its uncertainty in a purely statistical fashion.

[1] For example, it is possible to be an improbably bad author.

[2] We use Greek letters when binning with respect to $\mathcal{M}$ and Roman for binning citations.

While a good choice of indicator will assign each author to the correct bin with high probability, this will not be the case for a poor measure. Consider extreme cases in which we elect to bin authors on the basis of indicators unrelated to scientific quality, e.g., by hair/eye color or alphabetically. For such indicators, $P(i|\alpha)$ and $P(\{n_i\}|\alpha)$ will be independent of α, and $P(\alpha|\{n_i\})$ will be proportional to the prior distribution $p(\alpha)$. As a consequence, the proposed indicator will have no predictive power whatsoever. The utility of a given indicator (as indicated by the statistical accuracy with which a value can be assigned to any given author) will obviously be enhanced when the basic distributions $P(i|\alpha)$ depend strongly on α. These differences can be formalized using the standard Kullback–Leibler divergence. The method outline above was applied to several measures of author performance in [1, 2]. Some familiar measures, including papers per year and the Hirsch index [6], do not reflect an author's full citation record and are little better than a random ranking of authors. The most accurate measures (e.g., mean or median citations per paper) are able to assign authors to the correct decile bin with 90% confidence on the basis of approximately 50 papers. Since the accuracy of assignment grows exponentially with the number of papers, the evaluation of authors with significantly fewer papers is not likely to be useful.

4. DATA HOMOGENEITY

The average number of citations for a scientific paper varies significantly from field to field. A study of the impact factors on *Web of Science* [7] show that an average paper in molecular biology and biochemistry receives approximately 6 times more citations than a paper in mathematics. Such distinction, which are unrelated to field size or publication frequency, are entirely due to differences in the accepted referencing practice which have emerged in separate scientific fields. It is obvious that a fair comparison of authors in different fields must recognize and correct for such cultural inhomogeneities in the data. This task is more difficult than might be expected since significant differences in referencing/citation practice can be found at a surprisingly microscopic level. Consider the following subfield hierarchy:

$$\begin{aligned} \text{physics} &\rightarrow \text{high energy physics} \\ &\rightarrow \text{high energy theory} \\ &\rightarrow \text{superstring theory}. \end{aligned}$$

Study of the SPIRES database reveals that the natural assumption of identical referencing/citation patterns for string and non-string theory papers is grossly incorrect. Since its emergence in the 1980s, string theory has evolved into a distinct and largely self-contained subfield with its own characteristic referencing practices. Specifically, our studies indicate that the average number of citations/references for string theory papers is now roughly twice that of non-string theory papers in theoretical high energy physics. Any attempt to compare string theorists with non-string theorists will be meaningless unless these non-homogeneities are recognized and taken into consideration. Unfortunately, such information is not usually supplied by or readily obtainable from commercial databases.

5. IN SUMMARY

The Committee's report provides a much needed criticism of common misuses of citation data. By attempting to separate issues that are amenable to statistical analysis from purely subjective issues, we hope to have shown that serious statistical analysis does have a place in a field that is currently dominated by ad hoc measures, rationalized by anecdotal examples and by comparisons with other ad hoc measures. The probabilistic methods outlined above permit meaningful comparison of scientists working distinct areas with minimal value judgments. It seems fair, for example, to declare equality between scientists in the same percentile of their peer groups. It is similarly possible to combine probabilities in order to assign a meaningful ranking to authors with publications in several disjoint areas. All that is required is knowledge of the conditional probabilities appropriate for each homogeneous subgroup.

We emphasize that meaningful statistical analysis requires the availability of data sets of *demonstrated* homogeneity. The common tacit assumption of homogeneity in the absence of evidence to the contrary is not tenable. Finally, we note that statistical analyses along the lines indicated here are capable of identifying groups of scientists with similar citation records in a manner which is both objective and of quantifiable accuracy. The interpretation of these citation records and their relationship to intrinsic scientific quality remains a subjective and value-based issue.

REFERENCES

[1] LEHMANN, S., JACKSON, A. D. and LAUTRUP, B. E. (2006). Measures for measures. *Nature* **444** 1003.

[2] LEHMANN, S., JACKSON, A. D. and LAUTRUP, B. E. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics* **76** 369.

[3] LEHMANN, S., LAUTRUP, B. E. and JACKSON, A. D. (2003). Citation networks in high energy physics. *Phys. Rev. E* **68** 026113.

[4] REDNER, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physics Journal B* **4** 131.

[5] GARFIELD, E. (1977). Introducing citation classics: The human side of scientific papers. *Current Contents* **1** 1.

[6] HIRSCH, J. E. (2005). An index to quantify an individual's scientific output. *Proc. Natl. Acad. Sci.* **102** 16569.

[7] THOMPSON SCIENTIFIC. Web of science, Address retrieved October 2008. http://scientific.thomson.com/products/wos/.

Statistical Science
2009, Vol. 24, No. 1, 21–24
DOI: 10.1214/09-STS285C
Main article DOI: 10.1214/09-STS285

Comment: Citation Statistics

David Spiegelhalter and Harvey Goldstein

Key words and phrases: Research Assessment, citation indices, institutional comparisons.

We welcome this critique of simplistic one-dimensional measures of academic performance, in particular the naive use of impact factors and the h-index, and we can only extend sympathy to colleagues who are being judged using some of the techniques described in the paper. In particular we welcome the report's emphasis on the need for careful modeling of citation data rather than relying on simple summary statistics. Our own work on league tables adopts a modeling approach that seeks to understand the factors associated with institutional performance and at the same time to quantify the statistical uncertainty that surrounds institutional rankings or future predictions of performance. In the present commentary we extend this approach to an analysis of the 2008 UK Research Assessment Exercise (RAE) for Universities.

Before we describe our analysis it is important to comment on an important modeling problem that arises in the analysis of citation data, alluded to but not discussed in detail in the report, nor, as far as we know, elsewhere. A principal difficulty with indices such as the h-index or simple citation counts is that there are inevitable dependencies between individual scientists' values. This is because a citation is to a paper with, in general, several authors, rather than to each specific author. Thus, for example, if two authors nearly always write all their papers together, they will tend to have very similar values. If they belong to the same university department then their scores do not supply independent bits of information in compiling an overall score or rank for that department. Currently this issue is recognized in the RAE, albeit imperfectly, by the requirement that the same paper cannot be entered more than once by different authors for a given university department. In a citation based system this would also need to be recognized.

David Spiegelhalter is Winton Professor of Public Understanding of Risk, Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK. Harvey Goldstein is Professor of Social Statistics, University of Bristol, 35 Berkeley Square, Bristol BS8 1JA, UK.

In addition, if our two authors were in different, competing departments, we would also need to recognize this since the dependency would affect the accuracy of any comparisons we make. We also note that this will, to some extent, affect our own analyses that we present below, and it will be expected to overestimate the accuracy of our rankings. Unfortunately we have no data that would allow us to estimate, even approximately, how important this is. To deal with this problem satisfactorily would involve a model that incorporated "effects" for each author and the detailed information about the authorship of each paper that was cited. Goldstein (2003, Chapter 12.5) describes a multilevel "multiple membership" model that can be used for this purpose, where individual authors become level 2 units and papers are level 1 units.

The UK Research Assessment Exercise was published on 18th December 2008, covering the years 2001–2008. 52,409 staff from 159 institutions were grouped into 67 "units of assessment" (UOA): up to 4 publications for each individual were considered as well as other activities and markers of esteem. Panels drawn from around 1000 peer reviewers then produced a "quality profile" for each group, summarizing in blocks of 5% the proportion of each submission judged by the panels to have met each of the following quality levels: "world-leading" (4*), "internationally excellent" (3*), "internationally recognized" (2*), "nationally recognized" (1*), and "unclassified." This procedure is notable in terms of its use of peer judgment rather than simple metrics, and allowing a distribution of performance rather than a single measure. All the data is available for downloading (Research Assessment Exercise, 2008).

Figure 1 shows the results relevant for most statisticians: the 30 groups entered under UOA22: "Statistics and Operational Research." These have been ordered into a league table using the average number of stars which we shall term the "mean score," which is the procedure adopted by the media. Also reported is the number of full-time equivalent staff in the submission. Controversy surrounds this number as it is unknown how selective institutions were in submitting staff—

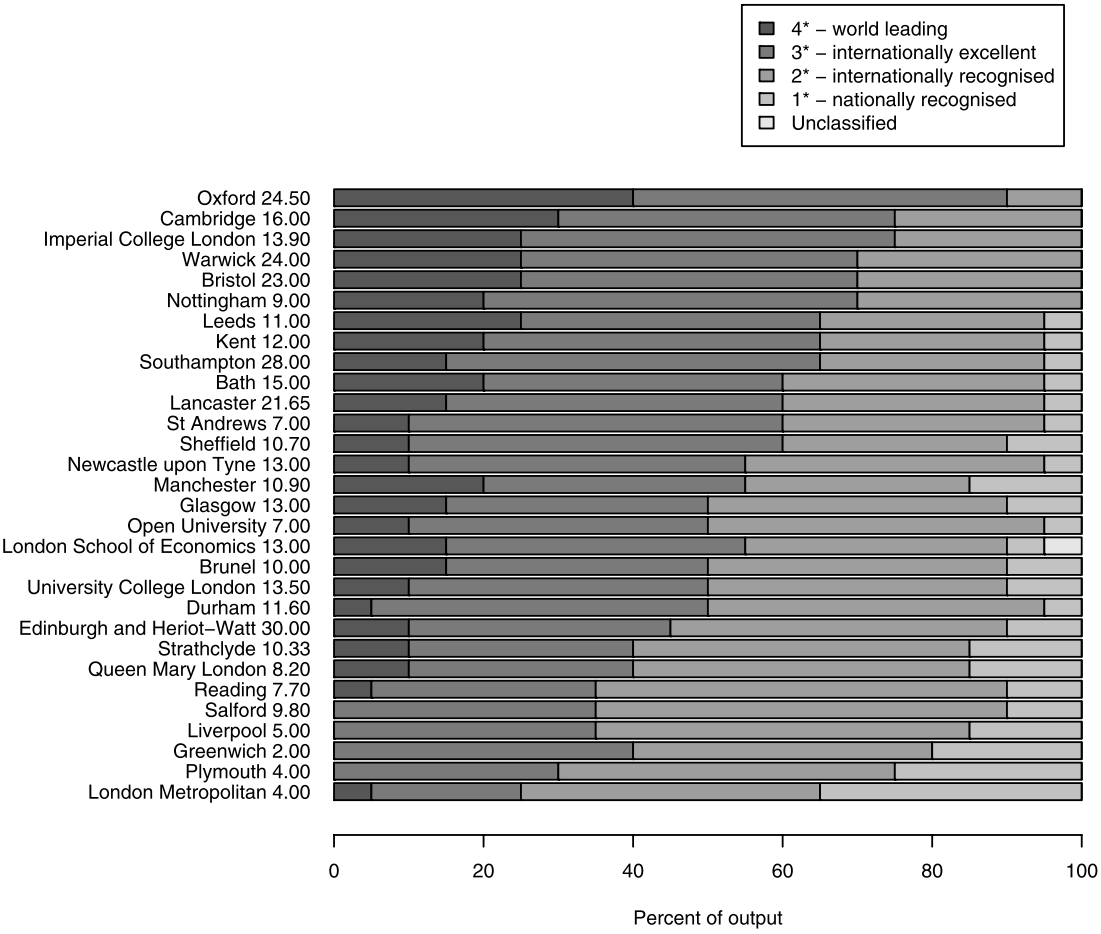

FIG. 1. *"Quality profiles" for* 30 *groups under UOA*22 *"Statistics and Operational research"*: *UK Research Assessment Exercise* 2008, *ranked according to mean score*: *numbers of staff taken into account are shown*.

it was originally intended that the total pool of staff would also be reported but late in the day there were objections raised as to the definitions of eligibility and this requirement was dropped.

The financial consequences of this whole exercise concern the distribution of around £1.5 billion of future funding. After publication of the quality profiles it was revealed that for funding purposes $4^*, 3^*, 2^*, 1^*$ outputs would be weighted proportional to $7, 3, 1, 0$: in further analysis we consider the "mean funding score" as $7p_4 + 3p_3 + p_2$, where p_i is the proportion of outputs given i stars.

In their report, Adler and colleagues argue that statistical analysis of performance data requires some concept of a model, and the provision of a quality profile rather than just a single number suggests it could be used for this purpose. We might first view the quality profile as representing the sampling distribution of material arising from each group, in fact a single Multinomial observation with probability (p_4, p_3, p_2, p_1): if, in the spirit of a bootstrap, we simulate from these distributions and rank the institutions at each iteration, we can produce a distribution for the predicted rank of a random future output from each group as shown in Figure 2.

We note the substantial overlap of the distributions: in fact the rank distributions are highly multimodal due to the extreme number of ties at each iteration, which explains the somewhat anomalous results for some groups in which the median rank order is substantially different from the mean-score order in which the institutions are plotted.

We are not, however, particularly interested in a single output and instead we may want to focus on the accuracy with which a summary parameter, such as the underlying mean funding score, is known: we treat this as an illustration of a general technique for analyzing any summary measure arising from a specified weighting. It then seems reasonable to take into account the quantity of information underlying the quality profile:

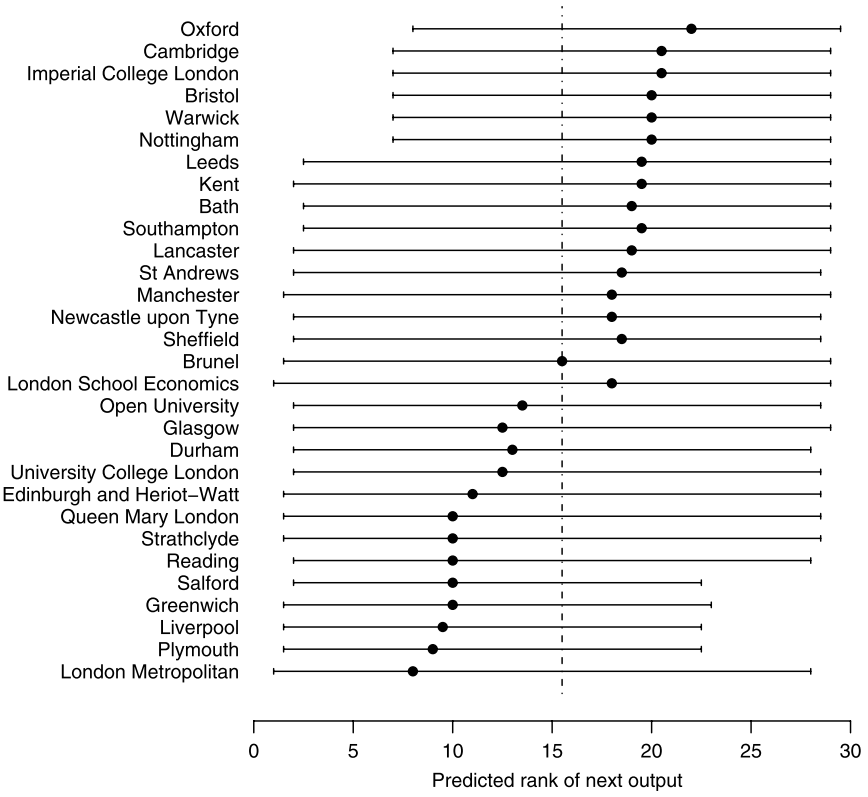

FIG. 2. *Predicted rank of a future output from each group*: *median and* 95% *intervals are shown based on* 10,000 *iterations*.

each individual contributes 4 publications and the publications count for 70% of the quality profile, and so we shall take a rough "effective sample size" as 6 outputs per staff member. Note that this does not mean that we are treating the publications as being a random sample from a larger population, but as relevant information connected through a probability model with some underlying parameter which may, in our particular illustration, be interpreted as the expected funding score of future outputs.

It would be possible to convert to ordered categorical data by multiplying the quality profile for each group by the number of publications taken into account (6 times the number of staff). Here, for the sake of simplicity, we have assumed a normal sampling distribution by estimating a standard error of the mean funding score as the square root of the sample variance of the profile divided by 6 times the number of staff.

Figure 3a shows the resulting estimates and 95% intervals for the mean scores. Treating these as normal distributions we can simulate future mean scores, rank at each iteration, and form a distribution for the "true" rank of each group. These are summarized in Figure 3b.

We see that for 14 out of 30 groups the 95% interval for the mean funding score overlaps the overall mean for all groups. Correspondingly we can identify 14 groups for which the 95% interval for their "true" rank, based on their mean funding scores, lies in either the top or bottom half. Both the mean funding scores and ranks, particularly for the smaller institutions, are associated with considerable uncertainty and this should warn against over-interpretation of either. If desired this could provide a basis for allocation into one of three groups for resource allocation purposes, although we would not necessarily recommend such a procedure.

We could, in principle, take this analysis further by noting that if we are really interested in predicting future performance, then we should be taking into account the possibility of regression-to-the-mean, recognizing the variation within each institution that would be expected over time. We could do this by fitting a hierarchical/multilevel model where conditioning takes place on the current scores (see Goldstein and Leckie, 2008, for an example using school league tables). We could adjust for background factors such as available resources in order to reduce the within-institution variability and to help satisfy relevant exchangeability assumptions, and so produce an "adjusted" institution effect. Whether we use this adjusted effect, or the fitted

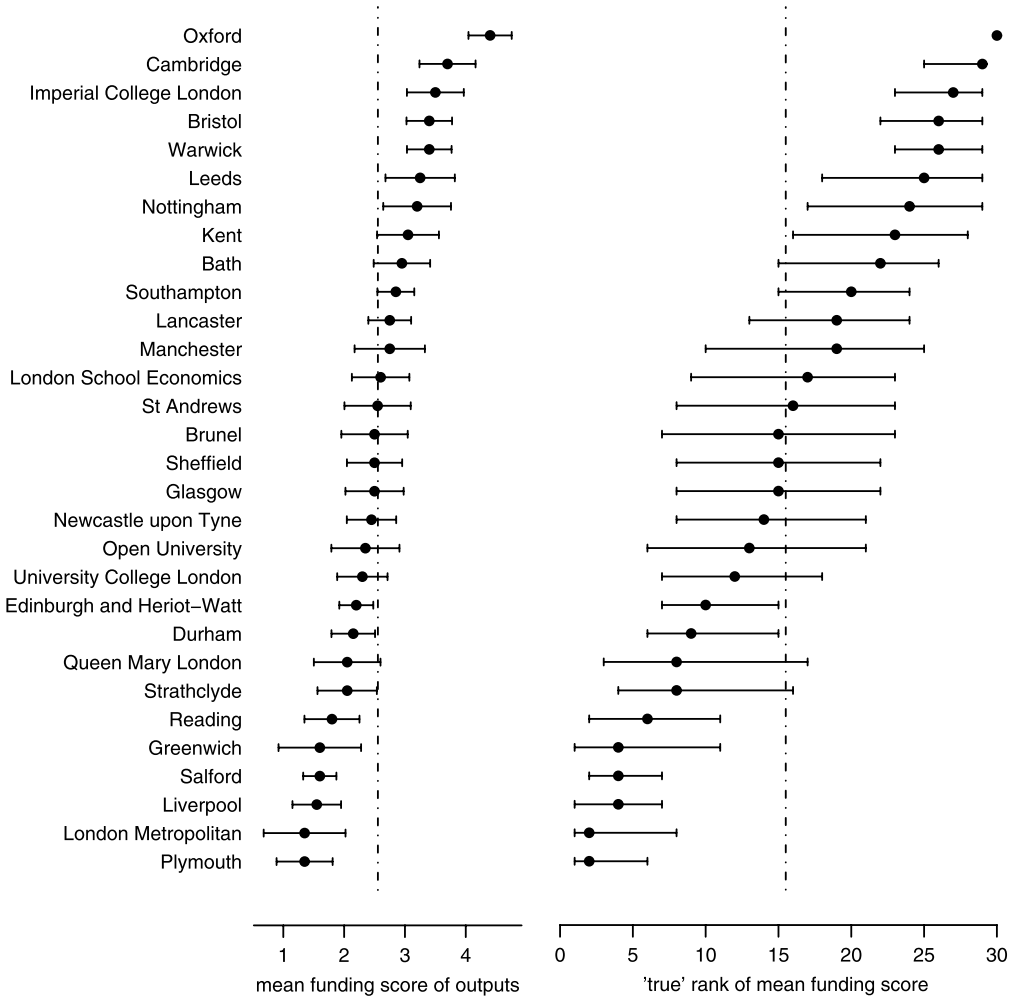

FIG. 3. (*Left*) *Estimates and intervals for expected funding score of outputs from each group*. (*Right*) *Summary of distribution of ranks of expected scores. Median and* 95% *intervals are shown based on* 10,000 *iterations*.

mean, as a basis for comparing groups would depend on the purpose: if we were university administrators wanting to know whether a group had done well given the resources available, then we would examine the adjusted affect. If, however, we wished to use the current scores simply to allocate income, then the fitted mean would be appropriate: see Goldstein and Leckie (2008) for a close examination of the potential role for different kinds of adjustments when comparing schools. In practice it is likely that such an analysis would be considered too complex.

In conclusion, we agree with the Report's strictures on the meaning of citation counts and would go further and argue that citations form a rather bizarre measure of research performance, as if the sole purpose of research was to provide material for other researchers. If they are to be used, we would argue that they be analyzed within a statistical modeling framework that fully incorporates uncertainty and dependency. As we have shown, for example, in Figure 3b, this could help to guide funding decisions by avoiding fine distinctions that may reflect little more than random noise. But citations alone, no matter how carefully analyzed, can only provide one measure of performance, and we feel strongly that they should be part of a broader profile that takes into account other measures of real world impact and is assessed using peer judgement rather than mechanistic and spuriously "objective" processes.

REFERENCES

GOLDSTEIN, H. (2003). *Multilevel Statistical Models*, 3rd ed. Edward Arnold, London.

GOLDSTEIN, H. and LECKIE, G. (2008). School league tables: What can they really tell us? *Significance* **5** 67–69.

RESEARCH ASSESSMENT EXERCISE (2008). http://submissions.rae.ac.uk/Results/.

Statistical Science
2009, Vol. 24, No. 1, 25–26
DOI: 10.1214/09-STS285D
Main article DOI: 10.1214/09-STS285

Comment: Citation Statistics

Peter Gavin Hall

Key words and phrases: Bibliometric analysis, bibliometric data, citation analysis, impact factor, journal ranking, research assessment.

I remember a US colleague commenting, in the mid 1980s, on the predilection of deans and other university managers for assessing academic statisticians' performance in terms of the numbers of papers they published. The managers, he said, "don't have many skills, but they can count." It's not clear whether the management science of assessing research performance in universities has advanced greatly in the intervening quarter century, but there are certainly more things to count than ever before, and there are increasingly sophisticated ways of doing the counting.

The paper by Adler, Ewing and Taylor is rightly critical of many of the practices, and arguments, that are based on counting citations. The authors are to be congratulated for producing a forthright and informative document, which is already being read by scientists in fields outside the mathematical sciences. For example, I mentioned the paper at a meeting of the executive of an Australian science body, and found that its very existence generated considerable interest. Even in fields where impact factors, h-factors and their brethren are more widely accepted than in mathematics or statistics, there is apprehension that the use of those numbers is getting out of hand, and that their implications are poorly understood.

The latter point should be of particular concern. We know, sometimes from bitter experience, of some of the statistical challenges of comparing journals or scientists on the basis of citation data—for example, the data can be very heavy-tailed, and there are vast differences in citation culture among different areas of science and technology. There are major differences even within probability and statistics. However, we have only rudimentary tools for quantifying this variation, and that means that we can provide only limited advice to people who are using citation data to assess the work of others, or who are themselves being assessed using those data.

Peter Gavin Hall is Professor of Statistics, University of Melbourne, Victoria, Melbourne, VIC 3010, Australia (e-mail: halpstat@ms.unimelb. edu.au).

Therefore, one of the conclusions we should draw from the study by Adler, Ewing and Taylor is that we need to know more. Perhaps, as statisticians, we could undertake a study, possibly funded in part by a grant awarding agency or our professional societies, into the nature of citation data, the information they contain, and the methods for analysing them if one must. This would possibly require the assistance of companies or organizations that gather such data, for example, Thomson Reuters and the American Mathematical Society. However, without a proper study of the data to determine its features and to develop guidelines for people who are inevitably going to use it, we are all in the dark. This includes the people who sell the data, those who use it to assess research performance and those of us whose performance is judged.

It should be mentioned, however, that too sharp a focus on citation analysis and performance rankings can lead almost inevitably to short- rather than long-term fostering of research excellence. For example, the appropriate time window for analyzing citation data in mathematics and statistics is often far longer than the two to three years found in most impact factor calculations; it can be more like 10–20 years. However, university managers typically object to that sort of window, not least because they wish to assess our performance over the last few years, not over the last decade or so. More generally, focusing sharply on citations to measure performance is not unlike ranking a movie in terms of its box-office receipts. There are many movies, and many research papers, that have a marked long-term impact through a complex process that is poorly represented by a simple average of naive criteria. Moreover, by relying on a formulaic approach to measuring performance we act to discourage the creative young men and women whom we want to take up research careers in statistical science. If they enjoyed being narrowly sized and measured by bean-counters, they'd most likely have chosen a different profession.

To illustrate some of the issues connected with citation analysis I should mention recent experiences in Australia with the use of citation data to assess research

performance. In the second half of 2007 the academies, societies and associations representing Australian academics were asked by our federal government to rank national and international journals, as a prelude to a national review of research and to the development of new methods for distributing "overheads" to universities. The request was not uniformly well received by the academic community. For example, I didn't like it. However, to the government's credit it did endeavor to consult. Different fields drew up journal rankings in four tiers, using methods (e.g., deliberation by committee) that they deemed appropriate. But the conservative government that proposed this process lost office in November 2007, and a month later the Labor government that replaced it quietly but assiduously set about revising the rankings. They still used four tiers, consisting of the top 5%, next 15%, next 30% and lower 50% of the cohort of journals in a given field. (Selecting the cohort was, and is still, a controversial matter.) However, in many cases the revised rankings differed substantially from the earlier ones.

In probability and statistics, and applied mathematics, the revised rankings were worked out by the bureaucracy and by consultants whom the government employed, using five-year journal impact factors apparently computed from purchased data. The resulting ranking departed from accepted norms in a number of important respects, enough to shed significant doubt on the credibility of the whole exercise. Initially the procedures laid down by the Australian Research Council (ARC) for commenting on their revised ranking seriously restricted the ability of the probability and statistics community to respond as a body, for example through a committee. However, thanks to timely intervention by the IMS President in early July 2008, we were given an opportunity to make a submission directly to the ARC.

This enabled us to form a committee to recommend the correction of a number of serious problems. For example, the ARC's revised ranking based on impact factors had dictated that no journals in probability could be in the top tier; probabilists generally publish less, and are cited less, than statisticians. Even within statistics there were a number of what I regarded as significant errors. For example, some high impact factor journals, dedicated to specific fields of application, were placed into much higher tiers than renowned journals that focused more on the development of general statistical methodology. Still other important journals were omitted entirely. The committee set to work to remedy these problems.

As you can imagine, the redistribution of journals among tiers was not without significant debate. I received very strong email messages from, for example, a medical statistician who objected strenuously to *Statistics in Medicine* being in a lower tier than the *The Annals of Probability*. As he pointed out, the committee revising the ranking had "no objective criterion" for journal ranking other than impact factors, and in Thompson Reuters' most recent (i.e., 2007) list of those factors, *The Annals of Probability* had an impact factor of only 1.270, whereas *Statistics in Medicine* enjoyed 1.547. Then there were the upset probabilists, who objected to the large number of statistics journals in the top tier, relative to the small number of probability journals. One probabilist suggested a substantial reduction in the number of statistics journals being considered. Several argued that too much attention was being paid to impact factors. (I was unsuccessful in persuading my statistics colleagues to move far enough away from an impact-factor view of the world to put the *The Annals of Applied Probability* into the top tier, but colleagues on the applied mathematics committee generously adopted the journal and placed it in their first tier.)

As these experiences indicate, the lack of a clear understanding by the probability and statistics community of the strengths and weaknesses of citation analysis is causing more than a few problems. If the Australian government has its way, whether a paper is published in a first- or second-tier journal will influence the standing of the associated research, and will affect the "overhead" component of funding that flows to a university in connection with that work. I think this is quite wrong, but at present we do not have much choice other than to make the best of a bad deal. In that context, if our community does not have a clear and authoritative understanding of the nature, and hence the limitations, of impact factors (and more generally of citation data), then we cannot react in an authoritative way to arguments that we feel are invalid, but are nevertheless strongly held. Frankly, we need to know more about citation data and citation analysis, and that requires investment so that we can investigate the topic.

Statistical Science
2009, Vol. 24, No. 1, 27–28
DOI: 10.1214/09-STS285REJ
Main article DOI: 10.1214/09-STS285

Rejoinder: Citation Statistics

Robert Adler, John Ewing and Peter Taylor

We would like to thank the discussants for reading our report and for their insightful and constructive comments.

To start our brief response, we would like to quote Bernard Silverman's phrase "reducing an assessment of an individual to a single number is both morally and professionally repugnant." Bernard puts it strongly, but his underlying point, with which we strongly agree, is that "research quality" is not something that ought to be regarded as well-ordered.

We note the general support for the case that any analysis should be carried out in the context of a properly-defined model. Peter Hall calls for statisticians to undertake a study of "the nature of citation data, the information they contain and methods for analysing them if one must." Among the three of us, there are varying levels of enthusiasm for advocating such a project. A possible downside is the danger that such a study will add to the burgeoning number of proposals for carrying out citation analysis in a "better" way, and none of us have much enthusiasm for this. On the plus side, such a study would enable the mathematical sciences community to comment more authoritatively on citation statistics and the quantitative ranking measures that are derived from them. Given that the scientometric industry shows every sign of growing, it can be argued that it is the responsibility of the mathematical sciences, and particularly of statisticians, to develop this capability.

David Spiegelhalter and Harvey Goldstein pointed out that there is a lack of independence between individual authors' citation records due to issues of coauthorship. The effects of this lack of independence seem to be very poorly understood, and nothing in the literature that we reviewed sheds any light on them.

Robert Adler, Faculty of Electrical Engineering, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel, 32000 (e-mail: robert@ieadler.technion.ac.il). John Ewing, President, Math for America 800 Third Ave, 31st fl, New York, New York 10022, USA (e-mail: ewing@mathforamerica.org). Peter Taylor, Department of Mathematics and Statistics, University of Melbourne, Vic 3010, Australia (e-mail: p.taylor@ms.unimelb.edu.au).

In our report, we spent some time discussing the meaning of citations. Sune Lehmann, Benny Lautrup and Andrew Jackson took this point further in their discussion of the fact that there needs to be agreement on the basic meaning of a researcher's citation distribution, which is something that goes beyond merely knowing what citations mean, which itself is not clear. Their example involving researchers A and B makes this point clearly.

We would like to emphasise three final points that have more to do with human behavior than statistics, and which were not emphasised in the report itself. The first is related to Bernard Silverman's point that any measurement or ranking system will drive researcher behavior via natural feedback mechanisms. Traditionally, the mechanisms adopted in academia have been qualitative rather than quantitative. Peer review has been at the core of the system. When carefully done, peer review not only provides accurate and professional assessments of an individual's contributions, but it also provides a balanced and educated interpretation of quantitative information such as prizes and citation data. Moving to a system based purely on quantitative citation metrics will deliver feedback more frequently, more unequivocally, and in a different way. It is not at all clear that "good research" (and we realise how loaded this term is) will be encouraged by such a system. Our strong opinion is that this feedback aspect is very important.

Related to this issue is another of particular concern. In general, it is not all that easy to fool one's peers, but it takes little imagination to see how, by adopting citation policies that are different from the norm in a particular discipline or sub-discipline, a small group of individuals could easily fool an automated assessment system built on citation data. Assessment is important to all of us, as individuals, as institutions, and as representatives of disciplines. Adopting a system, for short term gains, that is so easily open to abuse is a risk to research standards in the long term.

Our final point, which has been amplified by our experiences since the report was first released, is that almost everyone is affected by conflicts of interest when the topic of research assessment comes up. For most of

us, the way our research is regarded goes to the very core of our professional identity, and it would be a rare individual who could isolate his or her opinion about a particular method of research assessment and the way that his or her own research is ranked by the method. For example, most people who do well according to h-indices tend to think that the h-index is not a bad measure of researcher quality. There are also individuals who have built careers, and companies that have profited, from undertaking research assessment in a particular way. Since we are certain that it is healthy for all disciplines to have a multitude of skills and temperaments in their research communities, this observation leads us back to where we started: "research quality" is an inherently multidimensional object and should be treated as such.