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—