How Should Peer-Review Panels Behave?

Daniel Sgroi and Andrew J. Oswald

No 999

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

THE UNIVERSITY OF
WARWICK

How Should Peer-Review Panels Behave?

Daniel Sgroi* and Andrew J. Oswald**

26 May 2012

Abstract

Many governments wish to assess the quality of their universities. A prominent example is the UK's new Research Excellence Framework (REF) 2014. In the REF, peer-review panels will be provided with information on publications and citations. This paper suggests a way in which panels could choose the weights to attach to these two indicators. The analysis draws in an intuitive way on the concept of Bayesian updating (where citations gradually reveal information about the initially imperfectly-observed importance of the research). Our study should not be interpreted as the argument that only mechanistic measures ought to be used in a REF.

Key words

University evaluation; RAE Research Assessment Exercise 2008; citations; bibliometrics; REF 2014 (Research Excellence Framework); Bayesian methods.

JEL Codes

I23, C11, O30

*Department of Economics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL. Email: daniel.sgroi@warwick.ac.uk; Telephone: 02476 575557.

**Department of Economics and CAGE Centre, University of Warwick, and IZA Institute, Bonn.

We thank Daniel S. Hamermesh and five referees for very useful comments. Support from the ESRC, through the CAGE Centre, is gratefully acknowledged.

1. Introduction

This paper is designed as a contribution to the study of university performance. Its specific focus is that of how a research evaluation procedure, such as the United Kingdom's forthcoming 2014 REF (or "Research Excellence Framework"), might optimally use a mixture of publications data and citations data.

The paper argues that if the aim is to work out whether universities are doing world-class research it is desirable to put some weight on a count of the output of research and some weight on a count of the citations[1] that accrue to that output. The difficulty, however, is how to choose the right emphasis between the two. A way is suggested here to decide on those weights. Bayesian principles lie at the heart of the paper.

Whether there should be a REF is not an issue tackled later. Its existence is taken as given.

The background to our paper is familiar to scholars. Across the world, there is growing interest in how to judge the quality of universities. Various rankings have recently sprung up: the Times Higher Education global rankings of universities, the Jiao Tong world league table, the US News and World Report ranking of elite US colleges, the Guardian ranking of British universities, and others. This trend may be driven in part merely by newspapers' desire to sell copies. But its foundation is also genuine. Students (and their parents) want to make informed choices, and politicians wish to assess the value that citizens get for taxpayer money.

Rightly or wrongly -- see criticisms such as in Williams (1998), Osterloh and Frey (2009) and Frey and Osterloh (2011) -- the United Kingdom has been a leader in the world in formal ways to measure the research performance of universities. For more than two decades, the country has held 'research assessment exercises' in which panels of senior researchers have been asked to study, and to provide quality scores for, the work being done in UK universities. The next such exercise is the REF. Peer review panels have recently been appointed. These panels are meant to cover all areas of the research being conducted in UK

[1] In this article, citations data will be taken from the Web of Knowledge, published by Thomson Reuters, which used to be called the ISI Science Citations Index and Social Science Citations Index. Other possible sources are Scopus and Google Scholar.

universities. In the field of economics, the panel will be chaired by J Peter Neary of Oxford University.

Like the 2008 Research Assessment Exercise before it, the 2014 REF will allow universities to nominate 4 outputs -- usually journal articles -- per person. These nominations will be examined by peer reviewers; they will be graded by the reviewers into categories from a low of 1* up to a high of 4*; these assigned grades will contribute 65% of the panel's final assessment of the research quality of that department in that university. A further category for 'impact' will be given 20%. A category for 'environment' will garner the remaining 15%. Impact here means the research work's non-academic, practical impact.

As explained in the paper's Appendix, in many academic disciplines, including economics and most of the hard sciences, the REF will allow peer-review panels to put some weight on the citations that work has already accrued. Citations data will thus be provided to -- many of -- the panels. Citations data in this context are data on the number of times that articles and books are referenced in the bibliographies of later published articles. Because peer-review panels are to assess research that was published after December 2007, in a subject like Economics it is likely that only a small number of articles or books will have acquired more than a few dozen citations. For example, at the time of writing (May 2012), the most-cited Economic Journal articles since 2008 are Dolan and Kahneman (2008) and Lothian and Taylor (2008), which has each been cited approximately 60 times in the Web of Knowledge. In the physical sciences, by contrast, some articles are likely to have acquired hundreds of citations. For example, the article on graphene by Li et al (2008) is, at the time of writing, the most-cited article in the journal Science over the REF period. It has been cited over 1000 times.

There have been many advocates of bibliometric data. In 1997, Oppenheim's study of three disciplines found the following:

> The results make it clear that citation counting provides a robust and reliable indicator of the research performance of UK academic departments in a variety of disciplines, and the paper argues that for future Research Assessment Exercises, citation counting should be the primary, but not the only, means of calculating Research Assessment Exercise scores.

So these kinds of discussions are not new.

Butler and McAllister (2009) echo the point:

> The results show that citations are the most important predictor of the RAE outcome, followed by whether or not a department had a representative on the RAE panel. The results highlight the need to develop robust quantitative indicators to evaluate research quality which would obviate the need for a peer evaluation based on a large committee. Bibliometrics should form the main component of such a portfolio of quantitative indicators.

Using different data, Franceschet and Costantini (2011) promulgate the same kind of argument. Taylor's work (2011) makes the additional interesting point -- one that should perhaps be put to all critics[2] of the use of citations numbers -- that data on citations may help to restrain peer-reviewers' personal biases:

> The results support the use of quantitative indicators in the research assessment process, particularly a journal quality index. Requiring the panels to take bibliometric indicators into account should help not only to reduce the workload of panels but also to mitigate the problem of implicit bias.

However, citations data are an imperfect proxy for quality. It is known that, occasionally, 'bad' papers are cited, that self-citations are often best ignored, that citations can be bribes to referees, that occasionally an important contribution lies un-cited for years (like Louis Bachelier's thesis, 1900), and so on. These issues are not trivial matters but will not be discussed in detail. As a practical matter, it seems likely that citations data, whatever their attendant strengths and weaknesses, will become increasingly influential in academia. More on the usefulness of citations for substantive questions can be found in sources such as Van Raan (1996), Bornmann and Daniel (2005), Oppenheim (2008), Goodall (2010) and Hamermesh and Pfann (2012).

A Paradox

Most researchers are aware of journal rankings and would like their work to appear in highly rated journals. Paradoxically, the rubric of the REF ostensibly requires members of the peer-

[2] Critics are not uncommon. One of the authors of this paper has sat on a university senior-promotions committee where some full professors (not in economics) of a university routinely argued that they would not countenance the use of any form of citations data or of any form of journal ranking; these professors felt they were perfect arbiters of the quality of someone else's work. Some peer-review panels in the United Kingdom's REF, such as in Political Science and in Sociology, have stipulated that they will take a related position, and have chosen not to allow the study of any citations numbers or journal rankings.

review panel not to use information on the quality of journals. For the panel concerned with the social sciences, for example, the general instruction is:

No sub-panel within Main Panel C will use journal Impact Factors or any hierarchy of journals in their assessment of outputs.

However, individual members of a peer-review panel do not have to prove that they ignored journal rankings (and arguably could not do so), and it seems possible that peer-review panel members will see such a rubric is unenforceable. Some, in our judgment, will view the above italicized statement as at best a perplexing one and at worst perhaps even a foolish one. An economist would be likely to argue that, because scientific journals have different refereeing processes and acceptance rates, it would be efficient in university evaluation exercises to put some weight on the identity of a journal. Furthermore, it is perhaps illogical of the UK's REF designers to eschew (citations-based) journal rankings and yet simultaneously to provide citations data to panels who request that information.

For this paper, we assume that in Economics the peer-review panel members are aware of journal-quality rankings. However, the paper's concerns are more general.

2. A Weighting Approach in the Spirit of Bayes

How could data on publications be efficiently combined with data on citations? One way to do this would be to employ updating methods of a kind suggested three hundred years ago by Thomas Bayes.

Our later proposal boils down to a rather intuitive idea. It is that of using citations gradually to update an initial estimate (the Prior) of a journal article's quality to form instead a considered, more informed estimate (the Posterior) of its quality. The more an article is cited by others, the more, in general, it can be said to be having important influence. Articles that quietly sink without trace in the ocean of scientific publication can be said -- at least on this kind of view -- to have had little influence.

Bayesian methods are of course fairly standard within economics. It is perhaps surprising, therefore, that a Bayesian approach to evaluating publications represents a change relative to conventional ways of thinking. Largely because of this, the model described below is designed to be as simple as possible.

From page 10 of the REF "Consultation on draft panel criteria and working methods Part 2C: Draft statement of Main Panel C" (which includes the Economics and Econometrics sub-panel) we see that the intention of the REF is to seek to determine whether a paper is "work that makes an outstanding contribution to setting agendas for work in the field or has the potential to do so" which would merit a "four star" rating or of some lesser impact (advancing or contributing to a field).[3] The simplest possible way to model this intention is to think in terms of a simple binary partitioning of the state space. Essentially, either a paper submitted to the REF is considered to be making an outstanding contribution or not. On that basis we can specify the state space to be $\omega \in \Omega = \{a, b\}$ where "a" is taken to mean "making an outstanding contribution to setting agendas" and "b" is taken to mean "not doing so".[4]

Following the more general theoretical literature on testing and evaluation, such as Gill and Sgroi (2008, 2012), a Bayesian model in this context is simply one that produces a posterior probability, p_i, that a given submitted paper indexed by i is of type a rather than type b. An external rule could then be used to further divide papers into categories: for example if we wanted four categories as indicated by the REF on page 10 of the consultation document we could make use of four ordered threshold probabilities $\{p_r\}$ where $r \in \{1, 2, 3, 4\}$. We could then assign a paper to a category by finding the largest threshold exceeded by p_i. Hence, to be considered to be a "four star" submission, our candidate paper i would need to have a posterior $p_i > p_4$. The exact values of $\{p_r\}$ could be determined after the full range of posteriors for all submissions has been calculated to create any distribution of one, two, three and four-star publications as required. For instance, it would be easy to make each rating account for 25% of all submissions.

[3] Appendix B provides a description of relevant parts of the REF consultation document.

[4] Above and beyond the intentions of the REF, we are in essence thinking about whether a product, service or even individual (for example in the context of an expert) is "good" or "not" with citations in our model perhaps best thought of as recommendations or signals in other contexts. Modelling such an evaluation process as condensing more complex information into a simple binary decision is a common idea in the literature. As Calvert (1985, p. 534) puts it: "This feature represents the basic nature of advice, a distillation of complex reality into a simple recommendation".

Consider the prior probability that a paper makes an outstanding contribution to setting agendas, denoted by $q \in (0, 1)$. This is an open interval because the prior should not be so strong as to rule out the use of citations data which would be true if $q = 1$ (or indeed if $q = 0$ since this perfectly reveals that $\omega = b$). The most obvious place to find a prior is to look at the journal which has accepted the paper for publication and use some measure of the journal's historical quality as the prior.

Consider the historically observed probability that the "typical" paper published in a given journal k has of making an outstanding contribution. Call this the prior q_k. Then, in practice, a research evaluation panel has the task of finding a reasonable set of priors for a range of journals. This is not a trivial task.

For example, the number of journal rating exercises within economics is considerable; their rankings often differ noticeably (especially among lower-ranked journals); and they use a variety of methodologies. It is not the aim of this paper to pass judgment on the methods which exist within the literature or to develop a new way to rank journals (since we are interested in individual papers). For this paper's objective, the best approach seems instead to be to present examples of existing journal ratings, briefly describe the methodology that arises from each, and leave the choice of which is most sensible to the reader.[5] The next section will present some possibilities.[6]

A journal ranking gives starting information on the likely influence of a paper published in that journal. Data on the build-up of citations then provide a form of new 'information', in the sense used in Bayesian modelling. As the citations data accumulate, the starting potential is transformed into realized fact. Through Bayesian updating, this slowly shifts the posterior probability towards what might be viewed as the true assessment of the importance of a particular journal article.

[5] As well as providing a recent ranking and rating exercise, Ritzberger (2009) also provides a survey of both the recent journal rankings literature and the extreme nature of the variation in specific well-known journal rankings.

[6] Most journal rankings are citations collection exercises themselves; they typically rate journals on the number of citations achieved per paper or per page, sometimes modified for self-cites and other possible biases. On this basis, our prior is based on accumulated past data.

Assume that citations are revealing of genuine quality, so they provide some probability that a paper makes an outstanding contribution of above 0.5, but are themselves quite imperfect, so that that probability is below 1. In other words, citations are useful signals of quality but are not perfectly revealing.[7]

Within the broader Bayesian literature, this probability is referred to as accuracy or signal strength, and henceforth we will use the term accuracy and denote it by $\alpha \in (0.5, 1)$. We can also calibrate the precise value of α (and indeed the set of priors) and this is discussed in the next section. Alternatively, it is possible to pick a value for α that is felt to represent the accuracy of citations data or the weight we wish to attach to citations.

It is also necessary to think about how to define a "signal" that emerges from citations data. There are various ways to do this -- from literally thinking in terms of each citation as providing a positive signal, through to considering the overall scale of citations in a given time period. The approach taken here is closer to the latter.[8] Consider a time period *t* (for example one calendar year) and think in terms of the number of citations obtained during that time period. If the number of citations is above a certain threshold in time period *t* we can view the signal as being good and if below as being bad, in the following sense. Denote the binary signal as $x_t \in \{a, b\}$ where:[9]

$$Pr(\omega = a \mid x_t = a) = Pr(\omega = b \mid x_t = b) = \alpha \in (0.5, 1) \quad (1)$$

[7] The Research Excellence Framework de facto limits the number of years in which cites can be obtained and so we would not want to over-value those few years. More generally, as implied earlier, citations may exist to point out errors, may be of differing levels of significance (from pointing out that a paper exists to heralding it as seminal), or may be over-zealous self-cites; see vanRaan (2005) for much useful discussion. For these and other reasons, we want to keep the probability of a citation revealing that a paper makes an outstanding contribution to be significantly below 1, and the use of this probability allows us to reduce the importance of citations if we believe they are in some way suspect.

[8] Any of the suggested methods is fine, though periodicity has the added benefits of allowing comparability across different disciplines as discussed in section 4.4, by simply changing the duration of the time period in question to better fit the speed at which citations accrue in each discipline. It is also especially easy to form a binary signal as discussed.

[9] We are slightly abusing notation here by using *a* and *b* for both the signal space and the state space; but since we want a signal "*a*" to provide supporting evidence for the true state to be "*a*" this form of notation meets our need to be simple and reasonably intuitive.

In practice, one such threshold could be to find the average number of citations for any paper in a given time period (say one calendar year) denoted as β and set $x_t = a$ if the number of citations exceeds β and $x_t = b$ otherwise. Then there is a clear signal for a given *t,* and we can consider the history of such signals up to and including time *t* to be H_t. For example, if a paper has attracted a large number of citations (above the threshold for the signals to be good) in the years being considered by the REF, we would have a signal history H_5 = {a, a, a, a, a}.[10] This paper will also assume that each signal is conditionally independent.

It is necessary, finally, to specify an updating rule. That rule makes it possible to calculate the posterior conditional probability that a paper will make an outstanding contribution.

We use Bayes' Rule. For a paper published in journal *k*, with a single positive signal $x_1 = a$ on citations, the posterior probability is:

$$Pr(\omega = a|x_1 = a) = \frac{\alpha q_k}{\alpha q_k + (1-\alpha)(1-q_k)} \quad (2)$$

And for any history H_t:

$$Pr(\omega = a|H_t) = \frac{Pr\,(H_t|\omega=a) q_k}{Pr(H_t|\omega=a) q_k + Pr(H_t|\omega=b)(1-q_k)} \quad (3)$$

3. The Method in Practice

This section goes through the practical process of finding a posterior probability that a journal article makes an outstanding contribution -- starting with the acquisition of a set of priors.

3.1 Transforming Journal Ratings into Priors

Table 1, in Appendix A, presents relatively recent journal ratings. These come from Kalaitzidakis et al (2003), Palacio-Huerta and Volij (2004) and Ritzberger (2009) as columns (1) to (3) respectively. All serve as possible starting points for the prior probability that a paper makes an outstanding contribution. A brief summary of the methods used to form each rating is given in the notes below Table 1, though each represents a refinement over

[10] Submissions for the REF are drawn from the period 1 January 2008 to the end of 2013 (the submission deadline).

simply using raw citations data or even Impact Factors to rate the historical importance of each journal. The journals are ordered alphabetically. The rule used to decide inclusion in Table 1 was to first select journals (relevant for Economics) with an Eigenfactor at or above 0.006 recorded in ISI Web of Knowledge (2010 edition) plus any other journal from the top 35 in the other 3 ranking exercises. Ratings and rankings for a larger set of journals are to be found by going to the sources: Kalaitzidakis et al (2003), Palacio-Huerta and Volij (2004), Ritzberger (2009) and the ISI Web of Knowledge online database.[11]

Column 4 of Table 1 presents economics-journal ratings using the fairly new Eigenfactor approach. This approach has recently been proposed by the ISI Web of Knowledge as an alternative to the traditional 'Impact Factor'. To economists, it will be reminiscent of the Input-Output Model for which Wassily Leontief won the Nobel Prize. The Eigenfactor approach recognizes that citations create a set of ripple effects across different journals. It uses a matrix of journal importance (based on Impact Factors) to allow for the effect of each citing journal. Each coefficient in that matrix is thus a weight. If journal X contains articles which are often cited by the most-cited journals in economics (by Impact Factor), then X has a higher Eigenfactor than a journal with many cites in lower impact-factor journals.

At a technical level, a key aspect of journal ratings which will work well in a Bayesian model is that such ratings are cardinal, continuous, on the real line, and provide a reasonable amount of variation across journals. Ordinal rankings or categorized groups of journals are less useful. With this in mind, the journal ratings listed in Table 1 in Appendix A are especially well tailored to help provide a Bayesian prior.

Even a fully cardinal rating cannot simply be used directly to form a prior. The reason is that it is necessary to apply a transformation to ensure that the probability lies in the open interval *(0, 1)*. Furthermore, even where the journal rating produces a number in that interval, we might wish to apply a transformation to allow noise at the top and bottom of the distribution of ratings. When thinking about the requirement for noise it makes sense to consider the following question: *what proportion of paper published in even the very best*

[11] Any journal from these sources (or elsewhere) can be transformed into a prior by following the method discussed below. Note that Palacio-Huerta and Volij (2004) only contains ratings for 36 journals and all of these are included in Table 1.

journal in the discipline are themselves outstanding contributions? Similarly: *what is the chance that an outstanding contribution might come from a journal at the bottom of the distribution?* For example, should we think that there is a 5% chance that a paper published in the very best journal might still not make an outstanding contribution, or that there is a 5% chance that a paper published in a journal at the bottom of the distribution might make an outstanding contribution we could bound the prior to lie in the interval [0.5, 0.95]. The data in Oswald (2007) seem to show that good journals publish many 'bad' papers, and vice versa.

Denote a set of published journal ratings as *R*. Let the rating for an individual journal *k* be R_k. While any transformation is to some extent arbitrary our approach is to interfere with the ratings at the minimal possible level. To that end, we first identify the highest rated journal which we can call journal *k**. Next, we form a ratio by comparing journal *k* with journal *k**. Finally, we apply a simple function to restrict the transformation to lie within [0.05, 0.95] to allow for some noise at the top and bottom of the distribution. The transformation is therefore:

$$q_k = 0.9\frac{R_k}{R_{k^*}} + 0.05 \qquad (4)$$

This transformation is applied to each column of Table 1 to generate the respective columns (1) – (4) in Table 2. The final column in Table 2 averages the four previous columns. The journals are ordered alphabetically.[12]

[12] The implied priors are very different across the columns of Table 2; however, the *ranking* is of course retained from Table 1, and is largely consistent across columns. While the variation in priors across columns is of course entirely a function of the origin of the numbers in Table 1, we should probably not be too concerned about this variation. First, we can return to the underlying meaning of the prior: if "international quality" is considered a tough hurdle then it might make sense to choose from a column with lower suggested priors, otherwise the higher numbers might make more sense, so the choice of column should be tailored to the need. Next, since in the end since these priors will be weighted and will then be used to produce a coarse rating for each submission of a whole number up to 4, and finally these ratings will produce a GPA which ranks departments, the impact of specific numbers is less important than the fact that the priors that are used are consistent across submissions and that the rank order makes sense (so the prior on the American Economic Review should be higher than say a top field journal). Finally, we might be concerned with bias in submissions in the sense that a low-quality article that makes it into a top journal is more likely to be submitted to the REF

3.2 Calibration of Priors

There is an alternative. We could calibrate the priors in such a way as to generate sensible "matching pairs" of final posteriors for established journal articles which are considered to be at a similar level. To do this we would need two papers that are both "old" (they have reached a point where the posterior probability that they have made an outstanding contribution have stabilised) and judged to be of equal worth (their posteriors must be the same). Take, for example, two papers at the very top of the quality distribution – ones that are felt to be of close to identical worth by some fairly objective measure: for instance they may have formed the basis of a Nobel prize, or may have created a new sub-discipline. However these two papers may differ both in terms of outlet and citation numbers. In particular, one paper which is considered to be in a lesser journal but which attracted more cites might be used in conjunction with one which received fewer cites but was published in a more esteemed journal – crucially though the long-run posteriors are now considered to be identical. By varying the values attached to the priors on each journal, and the weight given to citations, it would be possible to equate the posteriors attached to the two papers. Rather than go into specifics we might be abstract and consider four journal articles: article A is considered to have made an equal contribution to article B (they have the same posterior), and article C is considered to have made an equal contribution to article D.[13] It would make sense for these example publications to be old enough for the impact of each paper to have been fully realized.

Let "citations per year relative to the average" be our sequence of signals. Imagine that all four papers were published 20 years ago. A was published in *Econometrica* and has attracted above average cites in 5 out of the 20 years, and B was published in the *Journal of Economic Theory* and has attracted above average citations in 10 out of the 20 years. Paper C was published in *Econometrica* and attracted more than average citations in 15 out of 20

---

than a high quality article that fails to do so. This might raise the hurdle for an outstanding contribution and push us towards certain columns in Table 2, though again the most important thing seems to be that an attempt to use this table in practice should involve picking a single column and sticking with that throughout the entire exercise: the precise value of the numbers used is probably less of a concern.

[13] Note that in what follows we are not assuming that articles A and B have made an identical contribution to articles C and D.

years while D was published in the *Journal of Economic Theory* and attracted more than average citations in all 20 years. Since we are now considering multiple papers, we need to complicate the notation in the model. Let $\omega_i = a$ denote the state when paper $i \in \{A, B, C, D\}$ "makes an outstanding contribution to setting agendas" and $\omega_i = b$ denote the state when paper i "does not make an outstanding contribution to setting agendas". Next we slightly alter the notation for the signal history to H_{ti} denoting the history to time period t for paper i. Recall that q_k denotes the prior probability that a typical paper at journal k "makes an outstanding contribution to setting agendas". Let $k \in \{Ecma, JET\}$ where *Ecma* denotes Econometrica and *JET* denotes the Journal of Economic Theory. All four were chosen because they were considered to be of equal worth by the REF panel, and so we can now calibrate the values of one of our journals relative to the other and the signal accuracy α based on the equivalent quality of the four papers following a simple scheme of linear equations as follows:

$$\frac{Pr\,(H_{tA}|\omega_A=a)q_{Ecma}}{Pr(H_{tA}|\omega_A=a)q_{Ecma}+Pr(H_{tA}|\omega_A=b)(1-q_{Ecma})} = \frac{Pr\,(H_{tB}|\omega_B=a)q_{JET}}{Pr(H_{tB}|\omega_B=a)q_{JET}+Pr(H_{tB}|\omega_B=b)(1-q_{JET})}$$

$$\Rightarrow \quad \frac{\alpha^5(1-\alpha)^{15}q_{Ecma}}{\alpha^5(1-\alpha)^{15}q_{Ecma}+(1-\alpha)^5\alpha^{15}(1-q_{Ecma})} = \frac{\alpha^{10}(1-\alpha)^{10}q_{JET}}{\alpha^{10}(1-\alpha)^{10}q_{JET}+(1-\alpha)^{10}\alpha^{10}(1-q_{JET})} \qquad (5)$$

And:

$$\frac{Pr\,(H_{tC}|\omega_A=a)q_{Ecma}}{Pr(H_{tC}|\omega_C=a)q_{Ecma}+Pr(H_{tC}|\omega_C=b)(1-q_{Ecma})} = \frac{Pr\,(H_{tD}|\omega_D=a)q_{JET}}{Pr(H_{tD}|\omega_D=a)q_{JET}+Pr(H_{tD}|\omega_D=b)(1-q_{JET})}$$

$$\Rightarrow \quad \frac{\alpha^{15}(1-\alpha)^5q_{Ecma}}{\alpha^{15}(1-\alpha)^5q_{Ecma}+(1-\alpha)^{15}\alpha^5(1-q_{Ecma})} = \frac{\alpha^{20}q_{JET}}{\alpha^{20}q_{JET}+\alpha^{20}(1-q_{JET})} \qquad (6)$$

Hence if we have a value of q_{Ecma} to use as our norm, perhaps drawn from something like column (5) in Table 2, we can find the values of q_{JET} and α by solving equations (5) and (6) simultaneously.

If we are happy with a set of priors, but are unsure of their accuracy, we could use any two papers considered to be of equivalent quality to find the value of α for instance by solving equation (5) with our given values for q_{Ecma} and $q_{JET.}$ Similarly, if we are happy with α but unsure about the prior probability of a paper being an outstanding contribution at a particular journal, we could use two papers considered to be of identical quality, one

published in the journal in question and one in another about which we are more confident about the prior, again simply solving equation (5) with given values for α and either journal prior would suffice to reveal the one unknown parameter. Providing we have a sufficient number of papers of equivalent worth to help form our simultaneous equations, it is always possible to set up a system of equations sufficiently large to solve out for the unknowns.[14]

3.3 Calculating Posteriors

It is also necessary to think about the period in question when considering what constitutes a signal. We might simply consider a year's worth of citations, and then compare this with the average for the given field or discipline in a year, or have some other way of interpreting citations data in mind as a means of subdividing signals into good news or bad news about the quality of the paper.[15] We can then apply Bayes' Rule in equation (3) directly to determine the posterior p_i for a particular paper *i*. Finally, we need to compare p_i with a pre-designed series of p_r thresholds to determine the star rating of the paper.

At a practical level, we might note that since both the prior and the data suffer from a degree of imprecision the final posteriors will carry through this imprecision. Issues such as which set of priors to use and which weights to apply only add to the potential difficulty. However, this problem is lessened by the coarse nature of the grading within the REF process (each submission is graded following a simple whole number up to 4), and much of the final discussion will likely be focussed on the grade-point average generated through averaging this coarse set of measures, resulting in an *ordinal* ranking of departments. While this does mitigate concerns about for example which set of priors is used, it also means it is essential that the method used is consistent across submissions, for example by making sure that the same set of priors and weights are used for all submissions.

4. Extensions and Generalization

The binary model suggested above has the advantage of simplicity. Indeed it may be the simplest Bayesian model for the task at hand and, given the scale of the REF, simplicity

[14] If we believe that the implicit priors used by early Research Assessment Exercises are a good starting place, we could attempt to impute these priors from earlier assessments. Mingers et al (2009) and others have done exactly this.

[15] Section 4.2 below discusses periodicity in more detail.

would seem highly advantageous. However, we can consider several generalizations and extensions.

4.1 A More Complex Model

First, we could move away from Bernoulli priors. We would need to make an alternative arbitrary distributional assumption, but we might prefer a distribution which allows us to consider more than two states, perhaps encompassing the four different star-ratings explicitly mentioned in the REF or even finer partitioning. The simple state space is justifiable partly through the heavy use of binary state spaces in the literature on evaluation and partly through the specific requirements of a four-star paper.[16] A four-star paper is required to "make an outstanding contribution to setting agendas". However, whether a paper does indeed make such a contribution is unlikely to be known for certain until well after the REF is concluded, and this is recognized with the additional caveat that a paper would also receive a four-star rating if it is perceived as having the "potential" to make such a contribution. In essence, therefore, the REF panel will be making a probabilistic judgment about the future; so it makes sense to think of a four-star paper as one where the panel feels the chances are high that it will make such a contribution, whereas lower-rated papers are those for which they feel the chances are lower. It is on this basis that the use of posterior probabilities and a binary state-space makes sense.

We might also like a more general categorization for our signals. Again, the use of simple forms of "good news" or "bad news" seems justifiable given the implicitly probabilistic judgment that the REF will have to make if they wish to determine the impact of a paper so early in its life. Moreover, since the final posterior is based on a history of citations, rather than a single signal, the final posterior will be fine enough to allow very fine grading of different submissions.

[16] For just a few from a pool of many possible examples in the literature on evaluation, see Calvert (1985), Chiao et al. (2007), Demange (2010), Farhi et al. (2005), Gill and Sgroi (2008), Lerner and Tirole (2006), Sah and Stiglitz (1986), Sgroi (2002) and Taylor (1999) all of which use simple binary models. Gill and Sgroi (2012) has some binary features but takes random draws to be from a continuous quality-dependent signal distribution and offers a way to generalize the framework in this paper though at the cost of simplicity and tractability.

It might be desirable to change α to allow different precisions for different data sources -- perhaps because we think citations from papers published in certain journals are noisier than those from other journals. In essence, we might wish to adopt something closer to the Eigenfactor approach discussed above as a method for examining individual citations as well as for forming priors. We might argue that this is less important for individual citations simply because this can already be incorporated into the prior (through the Eigenfactor approach), and hence it is likely to be of lesser importance given the short timeframe of the REF and that the complexity involved in forming an Eigenfactor measure for each individual submission would be prohibitive.

4.2 Different Periodicity and Applications outside Economics.

The methods suggested here are general. They can be applied in almost all disciplines and any process of rating individual papers. In particular, in the model section we specifically discuss what we call a data period: the period within which we wish to compare the citation record of a particular paper with some relevant norm. We can examine papers within other disciplines by altering both the period length and the norm. For example, in a discipline known for very quick recognition of new ideas (typified by some of the pure sciences) we might consider citations in terms of months rather than years and this will increase the amount of data available accordingly (and thereby reinforce the method). We might also change the norm against which a signal is measured to reflect the average number of citations in the discipline in question or follow any other convention.

Periodicity will have a profound effect: the same article might be heavily cited throughout a year and with one year as our period of measurement that boils down to one positive signal. If we consider instead the period of interest to be a month that might represent 12 positive signals. Consider for example paper *i* which is accepted at a journal *k* with $q_k = 0.6$ and with a signal accuracy of $\alpha = 0.7$ attached to the citations data. With one positive signal (above average citations for a year) that results in a posterior of $p_i = (0.7 \times 0.6) \div (0.7 \times 0.6 + 0.3 \times 0.4) \approx 0.78$ but with 12 positive signals (above average citations for twelve individual months) we would instead have $p_i = (0.7^{12} \times 0.6) \div (0.7^{12} \times 0.6 + 0.3^{12} \times 0.7) \approx 1$ for what might appear to be similar information. This adds another layer to the analysis but also

broadens the applicability of the method to the entirety of the REF and shows how the method might be used by other disciplines outside the remit of the REF itself.[17]

4.3 Incorporating Judgment

Mechanical procedures for the classification of quality are intrinsically dangerous. The method suggested here is a way of combining citations data with journal quality and seems to allow little room for individual judgments by reviewers. Judgment might be considered especially important early in the life of a paper, which would certainly apply to REF submissions.

However, this is easily altered. It can be incorporated into the construction of a Bayesian posterior, by including the assessment of a reviewer (such as a REF panel member) as an informative signal in its own right. To see how straightforward such an addition would be, consider an impartial and independent judgment (made without reference to the journal quality or citations history) to be a binary review of paper *i*, $j_i \in \{a, b\}$, with once again "*a*" representing a positive outcome and "*b*" a negative one relative to the state. We once again need to consider the accuracy of the judgment (which can also be seen to be the weight placed on judgment as opposed to journal quality and citations in the formation of the overall posterior). We can call this accuracy parameter $\gamma \in (0.5, 1)$ to keep judgments informative but not fully revealing. Then, if paper *i* was published in journal *k*, with a positive judgment $j_i = a$, the posterior probability would be:

$$Pr(\omega = a | j_i = a) = \frac{\gamma q_k}{\gamma q_k + (1-\gamma)(1-q_k)} \quad (7)$$

And with a citations history H_t as well as a judgment and prior, we have:

$$Pr(\omega = a | H_t, j_i = a) = \frac{Pr(H_t|\omega=a)\gamma q_k}{Pr(H_t|\omega=a)\gamma q_k + Pr(H_t|\omega=b)(1-\gamma)(1-q_k)} \quad (8)$$

[17] Another way to broaden applicability to other disciplines is through the citation hurdles we set to help define when a submission is at or above the outstanding-contribution hurdle. Looking at above average citations does some of this work as this will be with respect to other papers in the same discipline; but we will still need to apply discipline-specific rules. To some extent, submissions may also be judged relative to their field -- though questions of which field then become important and a different type of potential imprecision.

In essence, then, the judgment can either be seen to be treated in a similar way to citations data. It provides an additional useful signal (another "hard fact" concerning whether the paper under consideration makes an outstanding contribution or not) although with the added feature that the importance of judgment can be made different from citations data by varying γ relative to α. Alternatively, we can think of equation (7) as generating an "updated prior" in the sense that the judgment and journal quality combine to form a prior belief *before* the arrival of "hard facts" generated from citations data. The choice of interpretation does not alter the way it is incorporated into the analysis but can change the value placed on γ.

4.4 Declining Weight on Journal Quality over Time

The method discussed here also implicitly defines the weights put, over time, on the quality-label of the journal as opposed to that on accrued citations. We have assumed throughout that citations data reflect the truth about whether a paper is an outstanding contribution or not; this is captured in our assumption that $\alpha > 0.5$. Therefore, over time, we would hope that the citations history will be predominately composed of either positive or negative signals, guiding us towards the truth about whether a paper is an outstanding contribution or not. The importance of the prior will therefore fade over time.

To think about how to measure this decline in importance, we can do a thought experiment. With no citations data, the posterior will be equal to the prior. This is the case when no time has passed, i.e. when $t = 0$. Imagine a sequence of impressive citations year after year, each making us more and more certain that a paper does indeed represent an outstanding contribution. As $t \to \infty$ the posterior will tend towards 1. In essence, each positive signal is telling us that the paper represents an outstanding contribution, but we do not accept that verdict fully until it has been confirmed many times (technically an infinite number of times) and so we continue to place some weight on the prior.

Exactly how much weight we place on the prior of the journal name as opposed to the gradually accumulating positive citations data on the article will depend upon the accuracy of the data α, the prior itself q_k and the amount of positive data that we have accumulated t. To see this in operation, consider how the posterior rises from q_k to 1 as lots of positive signals accumulate. The sequence will proceed as follows:

$$q_k, \frac{\alpha q_k}{\alpha q_k + (1-\alpha)(1-q_k)}, \frac{\alpha^2 q_k}{\alpha^2 q_k + (1-\alpha)^2(1-q_k)}, \ldots, \frac{\alpha^t q_k}{\alpha^t q_k + (1-\alpha)^t(1-q_k)}$$

Take a numerical example. Set the history to be a sequence of purely positive signals $H_t = \{a, a, a \ldots\}$ and set parameter values $\alpha = 0.55$ and $q_k = 0.6$. Now the posterior will rise as follows (starting with $t = 0$): *0.6, 0.65, 0.69, 0.73, 0.77, ...* moving towards 1 as more and more positive signals (good citations years) accumulate. One way to think about the weight on the prior is to consider the prior to be suggesting a posterior of q_k and a positive signal to be suggesting a posterior of 1, with the actual Bayesian posterior set as a weighted average of the two. For the parameter values given above, the weight on the prior will fall as follows (starting at $t = 0$): *1, 0.88, 0.77, 0.67,* and so on moving towards zero as the posterior rises towards 1. This makes intuitive sense when we think of citations data as gradually replacing the role of journal quality in helping to determine whether a paper is an outstanding contribution. The weight on the prior falls quite rapidly in this example; this is because we have assumed that the message from the stream of citations numbers is unambiguously positive. Where data are contradictory, as might well be the case in the short-run (for instance with a good citations year followed by a poor one early in the life of a paper), the prior will remain highly important for longer than in this simple numerical example.[18]

4.5 Classical statistics and normality

When data are plentiful, and satisfy certain regularity conditions, we can consider a simpler approach inspired by classical statistics. To outline such an approach, we might first collect the count data generated by citations for a particular paper and for the journal in which it appeared. We could then form a weighted average of the paper's own cites (within the assessment period) and the journal's cites (possibly over a longer period, as there is some

[18] One of the features of Bayesian updating is the well-known result that a negative signal exactly cancels out the impact of a positive signal. This is true regardless of the length of the history, so for any H_t, if the number of *a* signals equals the number of *b* signals then $p_i = q_k$. Hence the prior will continue to be highly important so long as citations data remain contradictory. Since we have assumed that citations are a good guide to quality ($\alpha > 0.5$) we would not envisage such a contradictory sequence of citations data in the long-run. However, in the short-run such a contradictory sequence is not so unlikely. The calculations of Oswald (2010), which examine the ex-post world-class research by university departments, can be thought of as a particular example of the long-run use of accrued citation rates.

persistence), with weights that depend on the variance of citations for articles published in that journal, and perhaps on the variance of citations for articles that are similar to the one considered in other respects (such as author identity or gender or age, page length, field or sub-field, etc.).[19]

This method is intuitively a simplification of the approach suggested in the main part of this paper including some ad hoc elements (the weight to use between the journal's citations and the submission's citations) and has a feel of linking a prior (the citations attached to the journal) to data (the submissions own citations) to produce a posterior. This approach provides a simple rule-of-thumb variant of the fuller Bayesian method discussed in earlier sections.

For standard statistical methods to work, normality assumptions have to hold, so there need to be sufficient observations to allow the Central Limit Theorem to apply. While it is hard to give a clear indication of what number is sufficient, it is clear that the time frame used within the REF is not necessarily at that level. To recapitulate, the REF explicitly rules out papers published prior to 2008, and so offers a 6-year window up to 2012. Taking a year as an observation, indicating whether a publication has high cites relative to the average, we do not reliably have enough to begin an application of classical statistical methods.[20] A second issue is that Laband (2011) has shown that citations data are not symmetrically distributed which also makes normality questionable (all the more so for short data periods). Laband's work shows that a large number of journals (58) published the elite 400 articles in his sample -- and this after only about one decade of cites. [Published 2001-5. Cites counted in 2011, using Google Scholar.] In other words, after only about one decade there is evidence that the journal impact factor is an imperfect measure of the really important articles.

[19] We are grateful to an anonymous referee for suggesting this alternative approach.

[20] Using a period of below a year would be problematic, because it is much harder to measure the month in which a citation applies and moreover there will be many months in which some journals are published and others are not. Most journals have a fixed annual subscription period and so a year is the smallest reasonable period to use.

Looking beyond the REF, classical statistics might be applied when attempting to judge the contribution of older papers or perhaps when attempting to evaluate individuals over a longer time-horizon (for senior hiring decisions perhaps), and so should be considered as an alternative to the methods suggested in the main part of this paper.[21]

5. Discussion

To summarize, if we wish to calculate the probability that a given paper *i* published in a given journal *k* makes an outstanding contribution, we can first find a suitable journal rating (i.e. something cardinal rather than ordinal) which includes the journal *k*. We next convert this into a prior probability q_k that a typical paper published in journal *k* makes an outstanding contribution to setting agendas. We choose the periodicity of the citation data, for example, cites per year. Then we examine how well paper *i* performed relative to some metric, for example did it attract more citations than average in each year? This produces a series of signals $\{x_1, x_2, \ldots, x_t\}$, each of which we award an accuracy α based on how confident we are that our signals correctly identify good papers. We can also include judgments made by expert reviewers as an additional source of independent information (which should not take into account the journal quality or citations data) with a separate accuracy parameter γ. We then update our prior probability using the history of citations data and any judgments we wish to include through an application of Bayes' Rule to produce a posterior probability p_i that paper *i* makes an outstanding contribution.

For the purposes of the REF, the simple framework presented here seems to meet the requirement of offering a simple programme for forming Bayesian posterior beliefs on whether a given paper is likely to be an outstanding contribution, and a relatively easy conversion into a 1-4 star rating. As suggested in section 2, the REF could categorize based on requiring a fixed percentage of papers to be four-star rated, three-star rated and so on. Or the REF assessors might prefer to award a four-star rating for a paper that meets a fixed threshold level, and the same for three-star rating and so on. This gives a REF panel the ability to produce as much variation as they wish and so enables a clear ranking of departments and to be as fine or coarse as might be wished.

[21] We are grateful to an anonymous referee for pointing this out.

Beyond the "Economics and Econometrics" part of the REF, the methods suggested here also seem to have the potential to be useful across disciplines (as discussed in section 4.2).

These Bayesian techniques could also be used for the individual rating of scholars during hiring and tenure-review/promotion processes, or for other processes of rating where there is uncertainty concerning the true impact of a given piece of work. The Bayesian approach itself is of course applicable well beyond the process of rating individuals or papers. It has been discussed in other settings throughout economics, for instance for rating the quality of services or products (see footnote 12).

6. Conclusion

Governments want to measure the quality of their universities' research. This paper proposes a way in which evaluation panels could blend citations data with publications data. Although it should be tempered by, and combined with, the disinterested judgments of experienced experts, a Bayesian approach has been suggested here in which, to evaluate a published article,

(i) emphasis should initially be attached to the quality of the publishing journal;

(ii) as time passes, the weight given instead to the actual article's citations[22] should become dominant, and the weight on the quality of the journal should dwindle toward zero.

A Bayesian approach, as suggested above, allows these changing weights to be determined.

The paper formalizes a way of thinking that may already be held, albeit informally, by some peer-review panels and researchers. In the very short run, the quality of the journal acts as a kind of sufficient statistic; in the very long run, the number of citations plays that role. For the decades in between, weights on each have to be chosen.

[22] We would like to mention one other point, which we imagine will become increasingly recognized. The more that data on citations are discussed publicly, and stressed as a criterion for success, the more distorted, and less reliable, actual citations data will become as a signal. This is for the rather human reason that individual scholars and university departments will respond strategically to try to alter – manipulate would be a harsher word – their own citations rates. Thus, in the long run it will be necessary, as with other indicators, to ensure that citations data are constructed in such a way as to try to minimize these kinds of distortions.

To give a practical example, Table 2 reveals that articles published in the JPE seem to have a higher prior probability of making an outstanding contribution than those published in the EJ. [23] This reflects the JPE's higher Impact Factor. Imagine that a peer-review panel discovers that, after a small number of years, a specific article published in the EJ happens to have a significantly better citation record than one in the JPE. How should that panel react? In the language of this paper, the citations record of the particular EJ article constitutes a series of good signals, whereas the citations record of the particular JPE article does not. A reasonable question for the panel is then: how long should we persist in our initial belief which favoured the JPE article if the relative citations for the EJ one continue to suggest otherwise? For simple parameter values, our calculations find that Bayes' Rule would suggest that roughly 4 years of conflicting citation data are needed before the original opinion should be reversed.[24]

Although our paper's method is designed as a general one, we have been asked by referees to give a recent example from the specific field of economics of an article that is much-cited but did not appear in the journals usually placed at the head of a journal ranking.

To do this, it is perhaps natural to consider the UK's journal output over the last 2001-2007 Research Assessment Exercise[25]. One striking example is Collier and Hoeffler (2004) in Oxford Economic Papers. At the time of writing, this article has been cited approximately 450 times in the ISI Web of Knowledge. That is a larger number than any paper -- not just those from UK university researchers -- published in 2004 in the American Economic Review[26]. Even in 2007, when the RAE assessors would have been doing their assessment, it had garnered approximately 100 cites in the Web of Knowledge. We might conjecture that one reason for this article's success is that it deals with a topic that really matters.

[23] This choice is not meant to signify anything; neither of the authors of this paper currently plans to submit to the REF an output published in these journals.

[24] This calculation uses priors taken from column (5) of Table 2, which averages the four different journal ratings listed in the preceding columns, and assumes an accuracy of 0.55 attached to citations data.

[25] Related arguments and earlier data are given in Starbuck (2005) and Oswald (2007).

[26] In the Thomson Reuters eigenfactor journal rankings in economics, the AER is typically close to number 1 while OEP is typically approximately number 100.

We would like to draw to a close on a cautious note. Journal articles are the main raw material of modern science (and arguably have the advantage, whatever their faults, that they have been through a form of refereeing). Citations to them are the main marker of those articles' influence (and arguably have the advantage, whatever their faults, that they are observable in a way that is not true of the nodded approval of quietly self-selecting scientific communities). However, this paper should not be interpreted as an argument that qualitative peer-review judgments, or mature overview by experienced humans, should have no role to play in university evaluations. It has been suggested in the paper how such judgments could be incorporated within the Bayesian framework. We would not recommend that solely mechanistic procedures be used in REF-like evaluations of universities, scholars, departments, or disciplines.

References

Bachelier, L. (1900), Theorie de la Speculation. Thesis. University of Paris.

Bornmann, L. and Daniel, H.D. (2005). "Does the H-Index for Ranking Scientists Really Work?", Scientometrics, 65, 391-392.

Broadbent, J. (2010), "The UK Research Assessment Exercise: Performance Measurement and Resource Allocation", Australian Accounting Review, 20, 14–23.

Butler, L. and McAllister, I. (2009), "Evaluating the 2001 UK Research Assessment Exercise in Political Science", Political Studies Review, 7, 3–17.

Calvert, R. L. (1985), "The Value of Biased Information: A Rational Choice Model of Political Advice", Journal of Politics, 47, 530–555.

Chiao, B., Lerner, J. and Tirole, J. (2007), "The Rules of Standard Setting Organizations: An Empirical Analysis", RAND Journal of Economics, 38, 905–930.

Clerides, S., Pashardes, P. and Polycarpou, A. (2009), "Peer Review vs Metric-Based Assessment: Testing for Bias in the RAE Ratings of UK Economics Departments", Economica, 78, 565–583.

Collier, P., and Hoeffler, A. (2004), "Greed and Grievance in Civil War", Oxford Economic Papers, 56, 563–595.

Demange, G. (2010), "Sharing Information in Web Communities", Games and Economic Behavior, 68, 580–601.

Dolan, P. and Kahneman, D. (2008), "Interpretations of Utility and their Implications for the Valuation of Health", Economic Journal, 118, 215–234.

Farhi, E., Lerner, J. and Tirole J. (2005), "Certifying New Technologies", Journal of the European Economic Association, 3, 734–744.

Franceschet, M. and Costantini, A. (2011), "The First Italian Research Assessment Exercise: A Bibliometric Perspective", Journal of Informetrics, 5, 275–291.

Frey, B.S. and Osterloh, M. (2011), "Ranking Games", working paper, University of Zurich.

Gill, D. and Sgroi, D. (2008), "Sequential Decisions with Tests", Games and Economic Behavior, 63, 663–678.

Gill, D. and Sgroi, D. (2012), "The Optimal Choice of Pre-Launch Reviewer", Journal of Economic Theory, 147, 1247-1260.

Goodall, A.H. (2009), "Highly Cited Leaders and the Performance of Research Universities", Research Policy, 38, 1079–1092.

Hamermesh, D.S. and Pfann, G.A. (2012), "Reputation and Earnings: The Roles of Quality and Quantity in Academe", Economic Inquiry, 50, 1-16.

Hudson, J. (2012), "Ranking Journals", submitted for this volume.

Kalaitzidakis, P., Mamuneas, T. P. and Stengo, T. (2003), "Rankings of Academic Journals and Institutions in Economics", Journal of the European Economic Association, 1, 1346–1366.

Laband, D. N. (2011), "On the Use and Abuse of Economics Journal Rankings", submitted for this volume.

Lee, F.S. (2007), "The Research Assessment Exercise, the State and the Dominance of Mainstream Economics in British Universities', Cambridge Journal of Economics, 31, 309–25.

Lerner, J. and Tirole, J. (2006), "A Model of Forum Shopping", American Economic Review, 96, 1091–1113.

Lothian, J.R. and Taylor, M.P. (2008), "Real Exchange Rates over the Past Two Centuries: How Important is the Harrod-Balassa-Samuleson Effect?", Economic Journal, 118, 1742–1763.

Li, X., Wang, X, and Zhang, L. (2008), "Chemically Derived, Ultrasmooth Grapheme Nanoribbon Semiconductors", Science, 319, 1229–1232.

Mingers, J., Watson, K, and Scaparra, P. (2009), "Estimating Business and Management Journal Quality from the 2008 Research Assessment Exercise in the UK", Kent Working Paper No. 205.

Oppenheim, C. (1997), “The Correlation Between Citations Counts and the 1992 Research Assessment Exercise Ratings for British Research in Genetics, Anatomy and Archaeology”, Journal of Documentation, 53, 477–487.

Oppenheim, C. (2008), “Out with the Old and In with the New: The RAE, Bibliometrics and the New REF”, Journal of Librarianship and Information Science, 40, 147–149.

Osterloh, M. and Frey, B.S. (2009). “Are More and Better Indicators the Solution? Comment to William Starbuck”, Scandinavian Journal of Management, 25, 225-27.

Oswald, A.J. (2007), “An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision-Makers”, Economica, 74, 21-31.

Oswald, A.J. (2010), “A Suggested Method for the Measurement of World-Leading Research (Illustrated with Data on Economics)”, Scientometrics, 84, 99–113.

Palacio-Huerta, I. and Volij, O. (2004), “The Measurement of Intellectual Influence”, Econometrica 72, 963–977.

Ritzberger, K. (2009), “A Ranking of Journals in Economics and Related Fields”, German Economic Review, 9, 402–430.

Sah, R. K. and Stiglitz, J. E. (1986), “The Architecture of Economic Systems: Hierarchies and Polyarchies”, American Economic Review, 76, 716–727.

Sgroi, D. (2002), “Optimizing Information in the Herd: Guinea Pigs, Profits and Welfare, Games and Economic Behavior, 39, 137-166.

Starbuck, W. H. (2005), “How Much Better are the Most Prestigious Journals? The Statistics of Academic Publication”, Organizational Science, 16, 180-200.

Taylor, C. R. (1999), “Time-on-the-Market as a Sign of Quality”, Review of Economic Studies, 66, 555–578.

Taylor, J. (2011), “The Assessment of Research Quality in UK Universities: Peer Review or Metrics?”, British Journal of Management, 22, 202–217.

Van Raan, A. F. J. (1996), “Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises”, Scientometrics, 36, 397-420.

Van Raan, A. F. J. (2005), “Fatal Attraction: Conceptual and Methodological Problems in the Ranking of Universities by Bibliometric Methods”, Scientometrics, 62, 133–143.

Williams, G. (1998), “Misleading, Unscientific, and Unjust: The United Kingdom’s Research Assessment Exercise”, British Medical Journal, 316, 1079–1082.

Appendix A: Tables

Table 1: 75 Journals from Selected Published Journal Ratings

Journal	(1)	(2)	(3)	(4)
Accounting Review	-	-	13.28	0.01103
AER Papers and Proceedings	-	14.00	-	-
American Economic Review	100.00	77.90	36.14	0.10135
American Journal of Agricultural Economics	6.19	-	2.38	0.00668
Applied Economics	2.00	-	0.52	0.00720
Ecological Economics	0.89	-	0.33	0.02311
Econometric Theory	45.85	16.50	11.78	0.00868
Econometrica	96.78	102.60	100.00	0.04605
Economic Inquiry	6.03	6.30	7.40	0.00564
Economic Journal	20.71	12.20	16.78	0.02185
Economic Theory	22.43	18.70	15.30	0.01162
Economics and Philosophy	0.78	-	12.37	0.00100
Economics Letters	18.73	3.20	3.86	0.01574
Economics of Education Review	0.35	-	2.16	0.00614
Energy Economics	0.03	-	-	0.00868
Environmental and Resource Economics	-	-	1.72	0.00650
European Economic Review	23.76	13.30	8.53	0.01271
Experimental Economics	-	-	-	0.00874
Games and Economic Behavior	35.49	33.40	21.24	0.01679
Health Economics	0.20	-	3.90	0.01064
Insurance: Mathematics and Economics	0.16	-	2.45	0.00702
International Economic Review	23.04	16.00	39.44	0.01271
International Journal of Game Theory	6.09	13.20	2.72	0.00399
International Journal of Industrial Organization	4.26	-	4.07	0.00766
Journal of Accounting and Economics	0.76	-	16.38	0.01281
Journal of Accounting Research	-	-	12.29	0.01041
Journal of Applied Econometrics	16.59	13.30	8.56	0.01062
Journal of Banking and Finance	2.62	-	2.49	0.01428
Journal of Business and Economic Statistics	38.41	15.20	17.66	0.00989
Journal of Corporate Finance	-	-	6.14	0.00546
Journal of Development Economics	5.50	-	7.65	0.01357
Journal of Econometrics	54.91	21.70	25.99	0.03767
Journal of Economic Behavior and Organization	7.05	5.30	8.55	0.01514
Journal of Economic Dynamics and Control	14.54	10.80	11.16	0.01077
Journal of Economic Growth	-	-	29.45	0.00407
Journal of Economic Literature	18.78	80.60	-	0.01483
Journal of Economic Perspectives	34.26	31.80	-	0.02436
Journal of Economic Theory	58.76	35.30	34.58	0.02574
Journal of Economics and Management Strategy	1.38	-	8.06	0.00610
Journal of Environmental Economics and Management	11.85	12.50	7.78	0.00752
Journal of Finance	-	-	38.33	0.06137
Journal of Financial and Quantitative Analysis	-	-	12.12	0.00927
Journal of Financial Economics	9.89	15.40	30.97	0.05343
Journal of Financial Intermediation	-	-	11.35	0.00386
Journal of Health Economics	1.60	-	8.67	0.01269
Journal of Human Resources	21.34	17.40	9.25	0.01034

Journal of Industrial Economics	3.85	-	6.03	0.00620
Journal of International Economics	7.84	11.70	22.87	0.02049
Journal of International Money and Finance	-	-	2.11	0.00625
Journal of Labor Economics	12.76	17.80	19.21	0.01222
Journal of Law and Economics	3.90	-	11.24	0.00649
Journal of Mathematical Economics	7.64	10.30	10.63	0.00391
Journal of Monetary Economics	36.41	47.30	37.91	0.02699
Journal of Money, Credit and Banking	-	-	14.87	0.01401
Journal of Political Economy	65.19	68.60	51.34	0.03635
Journal of Public Economics	19.77	16.70	17.10	0.02492
Journal of Risk and Uncertainty	5.58	16.20	16.92	0.00329
Journal of the European Economic Association	-	-	-	0.01763
Journal of Urban Economics	4.37	-	6.07	0.00988
Management Science	-	-	6.65	0.03400
Marketing Science	-	-	14.81	0.01085
Oxford Bulletin of Economics and Statistics	8.35	2.70	5.16	0.00469
Pharmacoeconomics		-	-	0.00755
Public Choice	4.95	-	3.30	0.00770
Quarterly Journal of Economics	58.11	101.40	72.41	0.04757
RAND Journal of Economics	11.44	20.60	14.11	0.01507
Review of Accounting Studies	-	-	12.83	0.00310
Review of Economic Dynamics	-	-	10.71	0.00844
Review of Economic Studies	45.15	66.00	53.02	0.04750
Review of Economics and Statistics	28.02	16.70	20.11	0.02885
Review of Financial Studies	-	-	30.39	0.04750
Scandinavian Journal of Economics	10.66	4.30	5.26	0.00407
Social Choice and Welfare	6.89	12.80	10.22	0.00588
Value in Health	-	-	-	0.00921
World Development	3.22	-	2.02	0.01541

Notes: In column (1) the value assigned to each journal is taken from the "impact, age and self-citation adjusted by number of pages" figure listed in column (5) of Table 1 in Kalaitzidakis et al (2003). In column (2) the value assigned to each journal is taken from the ratio of the number of impact-weighted citations received by a journal relative to those obtained by the best journal in the sample derived in Table 1 in Palacio-Huerta and Volij (2004) using what they describe as the "invariant method". Column (3) reports the updated journal ratings presented in Table 1 of Ritzberger (2009) which uses the 2006 Social Science Edition of the Journal Citation Reports published by the Institute for Scientific Information, as well as a number of journals in related disciplines such as Finance. Each of these three sources contains more information on the precise methodology used. Column (4) lists Eigenfactors from the Thomson Reuters ISI Web of Knowledge 2010 Social Sciences Database (subcategory Economics). 75 journals were selected by taking journals linked with Economics, with an Eigenfactor at or above 0.006 recorded in the ISI Web of Knowledge (2010 edition) plus any other journal from the top 35 in the other 3 ranking exercises. The journals are ordered alphabetically. This list should not be taken to imply that we, as authors, believe that these are, in some unambiguous sense, the best seventy-five journals. Other approaches to rankings are discussed in Hudson (2012).

Table 2: Possible Bayesian Priors for Journal Quality

Journal	(1)	(2)	(3)	(4)	(5)
Accounting Review	-	-	0.17	0.15	0.16
AER Papers and Proceedings	-	0.17	-	-	0.17
American Economic Review	0.95	0.73	0.38	0.95	0.75
American Journal of Agricultural Economics	0.11	-	0.07	0.11	0.10
Applied Economics	0.07	-	0.05	0.11	0.08
Ecological Economics	0.06	-	0.05	0.26	0.12
Econometric Theory	0.46	0.19	0.16	0.13	0.24
Econometrica	0.92	0.95	0.95	0.46	0.82
Economic Inquiry	0.10	0.11	0.12	0.10	0.11
Economic Journal	0.24	0.16	0.20	0.24	0.21
Economic Theory	0.25	0.21	0.19	0.15	0.20
Economics and Philosophy	0.06	-	0.16	0.06	0.09
Economics Letters	0.22	0.08	0.08	0.19	0.14
Economics of Education Review	0.05	-	0.07	0.10	0.08
Energy Economics	0.05	-	-	0.13	0.09
Environmental and Resource Economics	-	-	0.07	0.11	0.09
European Economic Review	0.26	0.17	0.13	0.16	0.18
Experimental Economics	-	-	-	0.13	0.13
Games and Economic Behavior	0.37	0.34	0.24	0.20	0.29
Health Economics	0.05	-	0.09	0.14	0.09
Insurance: Mathematics and Economics	0.05	-	0.07	0.11	0.08
International Economic Review	0.26	0.19	0.40	0.16	0.25
International Journal of Game Theory	0.10	0.17	0.07	0.09	0.11
International Journal of Industrial Organization	0.09	-	0.09	0.12	0.10
Journal of Accounting and Economics	0.06	-	0.20	0.16	0.14
Journal of Accounting Research	-	-	0.16	0.14	0.15
Journal of Applied Econometrics	0.20	0.17	0.13	0.14	0.16
Journal of Banking and Finance	0.07	-	0.07	0.18	0.11
Journal of Business and Economic Statistics	0.40	0.18	0.21	0.14	0.23
Journal of Corporate Finance	-	-	0.11	0.10	0.10
Journal of Development Economics	0.10	-	0.12	0.17	0.13
Journal of Econometrics	0.54	0.24	0.28	0.38	0.36
Journal of Economic Behavior and Organization	0.11	0.10	0.13	0.18	0.13
Journal of Economic Dynamics and Control	0.18	0.14	0.15	0.15	0.16
Journal of Economic Growth	-	-	0.32	0.09	0.20
Journal of Economic Literature	0.22	0.76	-	0.18	0.39
Journal of Economic Perspectives	0.36	0.33	-	0.27	0.32
Journal of Economic Theory	0.58	0.36	0.36	0.28	0.39
Journal of Economics and Management Strategy	0.06	-	0.12	0.10	0.10
Journal of Environmental Economics and Management	0.16	0.16	0.12	0.12	0.14
Journal of Finance	-	-	0.39	0.59	0.49
Journal of Financial and Quantitative Analysis	-	-	0.16	0.13	0.15
Journal of Financial Economics	0.14	0.19	0.33	0.52	0.29

Journal of Financial Intermediation	-	-	0.15	0.08	0.12
Journal of Health Economics	0.06	-	0.13	0.16	0.12
Journal of Human Resources	0.24	0.20	0.13	0.14	0.18
Journal of Industrial Economics	0.08	-	0.10	0.11	0.10
Journal of International Economics	0.12	0.15	0.26	0.23	0.19
Journal of International Money and Finance	-	-	0.07	0.11	0.09
Journal of Labor Economics	0.16	0.21	0.22	0.16	0.19
Journal of Law and Economics	0.09	-	0.15	0.11	0.11
Journal of Mathematical Economics	0.12	0.14	0.15	0.08	0.12
Journal of Monetary Economics	0.38	0.46	0.39	0.29	0.38
Journal of Money, Credit and Banking	-	-	0.18	0.17	0.18
Journal of Political Economy	0.64	0.65	0.51	0.37	0.54
Journal of Public Economics	0.23	0.20	0.20	0.27	0.22
Journal of Risk and Uncertainty	0.10	0.19	0.20	0.08	0.14
Journal of the European Economic Association	-	-	-	0.21	0.21
Journal of Urban Economics	0.09	-	0.10	0.14	0.11
Management Science	-	-	0.11	0.35	0.23
Marketing Science	-	-	0.18	0.15	0.16
Oxford Bulletin of Economics and Statistics	0.13	0.07	0.10	0.09	0.10
Pharmacoeconomics	0.05	-	-	0.12	0.08
Public Choice	0.09	-	0.08	0.12	0.10
Quarterly Journal of Economics	0.57	0.94	0.70	0.47	0.67
RAND Journal of Economics	0.15	0.23	0.18	0.18	0.19
Review of Accounting Studies	-	-	0.17	0.08	0.12
Review of Economic Dynamics	-	-	0.15	0.12	0.14
Review of Economic Studies	0.46	0.63	0.53	0.47	0.52
Review of Economics and Statistics	0.30	0.20	0.23	0.31	0.26
Review of Financial Studies	-	-	0.32	0.47	0.40
Scandinavian Journal of Economics	0.15	0.09	0.10	0.09	0.10
Social Choice and Welfare	0.11	0.16	0.14	0.10	0.13
Value in Health	-	-	-	0.13	0.13
World Development	0.08	-	0.07	0.19	0.11

Notes: The contents of Table 2 are direct transformation of the contents of Table 1 on a column-by-column basis. The sources for the raw data used in Table 2 are therefore the same as for Table 1. The transformation used is to select the highest-rated journal and use this as a benchmark. Every journal's rating is then expressed relative to the benchmark journal, but with a simple transformation to shrink the range to (0.05, 0.95) and thereby ensure that there exists noise in the prior even at the extremes of the distribution. The precise formula is described in the main text. The journals are ordered alphabetically.

Appendix B: Background Information on the REF

This Appendix summarizes and reproduces some key parts of the recent REF consultation document, "Consultation on draft panel criteria and working methods. Part 2C: Draft statement of Main Panel C" (which includes the Economics and Econometrics sub-panel). It is important to note that this is a consultation document and not yet a firm set of rules. Nevertheless it provides an indication of the thinking of the REF panel as of August 2011.

Appendix B.1: Journal Ratings

From page 10 of the REF document we see a clear distinction between four star submissions and the rest. In particular, for the social sciences, a four star submission makes an "outstanding contribution to setting agendas" whereas lower rated submissions merely advance or contribute to the field. The precise descriptions from page 10 are as follows:

Submission Rating	***Description***
Four star	*Work that makes an outstanding contribution to setting agendas for work in the field or has the potential to do so*
Three star	*Work that very considerably advances the field*
Two star	*Work that considerably advances the field*
One star	*Work that contributes to the field*
Unclassified	*Work that fails to contribute to the field, or does not meet the published definition of research for the purposes of this assessment*

Appendix B.2: Use of Citations

From page 14 of the REF consultation document, the numbered points 65 to 70 are especially relevant. To summarize, no sub-panel will make use of journal Impact Factors or hierarchies, the "Economics and Econometrics" sub-panel will receive and make use of citations data but will be careful when examining citations especially for very recent outputs. The specific points are reproduced below:

65. Sub-panels 16, 19, 20, 21, 22, 23, 24, 25 and 26 will neither receive nor make use of citation data, or any other form of bibliometric analysis.

66. No sub-panel within Main Panel C will use journal Impact Factors or any hierarchy of journals in their assessment of outputs.

67. Sub-panels 17 (Geography, Environmental Studies and Archaeology) and 18 (Economics and Econometrics) will receive and may make use of citation data, where they are available and considered appropriate. Sub-panel 17 may make use of citation data for some areas of physical geography and environmental studies, consistent with the practice in UOA 7 (Earth Systems and Environmental Sciences). It will not use citation in respect of the archaeology outputs that it assesses, nor for human geography.

68. Where such data are available, the REF team will provide citation counts for those outputs where this is possible (by a pre-determined date and from a pre-specified and consistent set of sources), as additional information. The absence of citation data for any individual output will have no bearing whatsoever on its assessment.

69. Sub-panels 17 (Geography, Environmental Studies and Archaeology) and 18 (Economics and Econometrics) will be mindful that for some forms of output (for example research monographs, or forms relating to applied research), and especially for very recent outputs, citation data may be unavailable or a particularly unreliable indicator.

70. They will also be aware of the analysis of the REF bibliometrics pilot exercise in relation to equality implications of using citation data, and will be alert to any potential bias that might arise from using citations data.

71. Citation data will not be used as a primary tool in the assessment, but only as supplementary information, where this is deemed helpful, about the academic significance of an output. Sub-panels will make rounded judgments about the quality of outputs, taking into account the full range of assessment criteria (originality, significance and rigour). The sub-panels will only use citation data provided by the REF team and will not refer to any additional sources of bibliometric analysis, including journal Impact Factors.

Appendix B.3: Calibration

Page 28 of the REF document also discusses calibration, particularly with reference to international comparability. The specific numbered points are reproduced below:

135. To ensure the consistent application of assessment standards, each sub-panel will undertake calibration exercises with respect to outputs and impact at an early stage in the assessment phase (or, in the case of research outputs, possibly immediately prior to it).

136. The calibration exercise for research outputs will involve all those members of the sub-panel who will subsequently be involved in assessing outputs and, as far as practicable, academic assessors of the sub-panels.

International members of the main panel will also participate in the calibration exercises to assist in benchmarking judgments against levels of international quality.