Economica LSE

Economica (2022) **89**, 830–848
doi:10.1111/ecca.12438

Rising Stars: Expert Reviews and Reputational Yardsticks in the Research Excellence Framework

By Erich Battistin* and Marco Ovidi†

*University of Maryland, CEPR, FBK-IRVAPP and IZA †Università Cattolica del Sacro Cuore

Final version received 1 July 2022.

We use the UK's 2014 Research Excellence Framework (REF) to study the attributes of top-scoring (four-star) publications in Economics and Econometrics. Although official documents contain aggregate scores for each institution, we show how these aggregates can be used to infer the score awarded by REF panellists to each publication. We demonstrate that this score responds to journal prestige as measured by the Thomson Reuters Article Influence Score. Several econometric analyses confirm the limited contribution of other publication attributes, such as the citation counts, to the awarded score, and publications in the top generalist and top five economics journals are awarded four stars unambiguously. We conclude that in large-scale evaluations such as the REF, peer reviews and bibliometrics should be viewed as complementary modes of assessment; the time-consuming task of peer reviews would be more cost-effective if targeted at publications whose quality cannot be classified unambiguously using bibliometrics.

Introduction

Performance-based research funding has been adopted in most European countries to encourage and reward excellence in universities, which leads to a surge in the number of national assessments conducted on a regular basis (Zacharewicz *et al*. 2019). Research quality serves as the yardstick by which, commonly, productivity and reputation are assessed and compared. The extent to which assessments are based on quantitative indicators of research impact (i.e. bibliometrics) or reviews by academic experts varies across countries.

We consider the Research Excellence Framework (REF), whose latest evaluation round across higher education institutions of the UK has recently been completed. Approximately 20% of higher education budget in the country relies on public funding (Carpentier 2021), allocated using the REF outcomes. Assessments have been conducted in the UK since 1986, and the most recent results were published at the end of 2014.[1]

One of the accountability pillars of the REF is the quality assessment of research outputs, which is the focus of our work. Research quality is assessed by peer reviews following general guidelines regarding originality, significance and rigour. However, the contribution of each output to the awarded quality of an institution is not disclosed. For each institution, official documents report only the share of outputs by the number of awarded *stars*, which ranges from one to four. This lack of full disclosure has spurred the discussion on how to filter top-scoring (four-star) works in future submissions. Therefore, understanding how the quality of a publication is assessed is essential for the incentive structure faced by institutions.

We consider REF2014, and use all research outputs submitted to the Economics and Econometrics sub-panel to infer the determinants of research quality awarded by experts. We complement this information with output-level data on citations, bibliometric indicators of journal impact, and information about the authors as of the REF2014 submission date. These variables are the closest approximations to the information available to panellists at the time of their assessments.

Although REF statistics are aggregated by institutions, we show in Section III that they remain informative on the classification of single outputs. We use this result to develop an econometric model to estimate the contribution of journal impact and other publication characteristics to the classification of experts of the individual papers. Specifically, we show that the number of stars awarded to publications is predicted by the influence of a journal's articles, which we measure with the Article Influence Score (AIS—see Thomson Reuters 2014). Citation counts vary considerably across publications in the same journal. However, we reveal that statistically, citations and other output characteristics such as the h-index of authors add very little to the AIS in explaining how research quality is awarded by experts.[2] The limited contribution of citation counts is particularly evident for publications in high-impact journals. In particular, the results of our analyses indicate that publications in top generalist and top five economics journals are unambiguously awarded four stars.

We mark important differences from previous research. The study by Bertocchi *et al*. (2015) is the closest to ours, although they consider the Italian equivalent of the REF with a methodology tailored to their case study.[3] The model in Pitt and Yan (2017) is similar in spirit to ours, although—other than using different statistical assumptions—their analysis does not rely on output-level information. In contrast to Hole (2017), who develops an algorithm to classify journals, our analysis does not restrict all publications in one journal to contribute equally to the classification of REF outputs. In addition, Hole (2017) considers only journals with the most submissions in REF2014; instead, we use all research outputs, including those not published in academic journals. Our findings mirror the correlations between citations and REF scores in Traag and Watman (2019) across several disciplines. Their analysis is based on an indicator of excellence derived from citations. Instead, we estimate the correlation of peer reviews with journal-level metrics (AIS) and output-level characteristics (such as citations) to draw more nuanced recommendations for policy.

The predictive power of AIS documented here mirrors the high correlation in Higher Education Funding Council for England (2015) between stars awarded in the REF and the SCImago Journal Rank index (an indicator similar to AIS). The Council's analyses are mostly univariate, employ anonymized data that are not available publicly, and are limited to works published in 2008, which covers approximately 10% of the REF submissions for economics and econometrics. Our work provides a method to overcome these limitations using data that can be retrieved from official documents, and opens the door to replicability in fields other than economics. We select the AIS as the indicator of journal influence due to multidisciplinary research. The AIS builds on the same algorithm used by Google to rank websites (Brin and Page 1998). The optimality of the AIS for measuring intellectual influence is demonstrated using the axiomatic approach in Palacios-Huerta and Volij (2004). Moreover, the AIS predicts expert-based evaluations of research quality (Hudson 2013). In particular, the REF2014 data show that the AIS predicts what institutions value when deciding about submissions, as more than half of their outputs are concentrated in few journals (23 of 283 in total) with above-average AIS, as we show below.

Our results are of interest to policymakers who study the regulatory framework of future assessments. The evaluation of research outputs in REF2021 will follow the same REF2014 regulation. We do not take a normative approach on how the reviews should be conducted, which is the focus of Regibeau and Rockett (2016), among others. However, our study is informative of how peer reviews and bibliometrics can be combined in the interests of time and public spending (see also Moed 2007). Specifically, by outlining the path to REF2021, Lord Stern's review (Stern 2016) recommends the responsible use of bibliometrics (Wilsdon *et al*. 2015).[4] Our findings suggest that for economics and econometrics, peer reviews and bibliometrics in a large-scale evaluation such as the REF may be viewed as complementary modes of assessment to identify unambiguously top-scoring journals and review outputs outside this tier only. Our policy recommendation is that the time and resources needed for peer reviews should be devoted to finding hidden four-star gems in academic outlets with lower bibliometric indicators of impact rather than overrated outputs in top-scoring outlets.

The remainder of this paper is organized as follows. Section I presents the institutional background. Section II describes how we integrate the information in REF official documents with bibliometrics data. Section III documents how aggregate statistics on REF performance by institutions are informative on the correlation between the AIS and the classification of single outputs. Our econometric model is presented in Section IV. Section V presents the empirical results, and Section VI concludes the paper.

I. Background and Context

The Research Excellence Framework

The UK is considered a world leader in higher education, which contributes £73 billion annually to the national economy and has been linked to 20% of GDP growth between 1982 and 2005 (Universities UK 2015). The quality of research produced by institutions is assessed on an approximate six-year cycle by the Research Excellence Framework (REF), which was commissioned by the four UK higher education funding bodies.[5] The REF provides accountability for investment in research, with implications for the allocation of public funding and the reputation of institutions.

We use data from REF2014, completed at the end of 2014. Special panels assessed the productivity of 154 universities between 2008 and 2013, and reviewed 190,000 research outputs by 52,000 academics. In the 2020/21 fiscal year, Research England distributed approximately £1.6 billion of research funding using REF2014 outcomes (Research England 2020). This source of funding represents 20% of universities' budgets and is their second largest source of income (Carpentier 2021).

We consider the evaluation of research outputs, which constitute the most important component of the REF. Compared to its predecessors, REF2014 assessed research impact in addition to research quality.[6] Research quality accounted for 65% of the profile of each institution, 20% was awarded for impact, and an additional 15% was for the research environment (e.g. infrastructure and income through research activities). The performance on research quality is what matters most in the decision to hire or expand (De Fraja *et al*. 2019).

The latest assessment, REF2021, has been implemented similarly to REF2014. The most substantial changes concern the selection of faculty, which is not a dimension considered in this study. Specifically, the representativeness of research was not guaranteed in REF2014 because institutions could decide to submit outputs for selected faculty members.[7] Moreover, the portability of research outputs provided the incentives to hire productive researchers from other institutions close to the REF census date. REF2021 introduces some changes to address these issues, which Stern (2016) describes as sources of 'gaming' behaviour.[8]

Assessment of research outputs

We look at the 2600 research outputs submitted to the REF2014 Economics and Econometrics sub-panel. This unit of assessment considered the outputs from 28 departments in the UK (see Table G.1 of the Online Appendix). Guidance and criteria for the evaluation process were disclosed well before the submission deadline. Institutions were invited to submit outputs authored by the employed staff and published between 2008 and 2013.

The evaluation relied entirely on peer reviews by the panellists. The Economics and Econometrics sub-panel consisted of 18 national and international academics who read all submissions to identify the overrated outputs and hidden four-star gems regardless of the publication outlet. This process implies an average workload of approximately 140 outputs per panellist. Each listed output in a submission was assessed based on its originality, significance and rigour. Since this definition allows for subjective judgement, the panellists used bibliometrics to inform assessments 'when considered appropriate' (REF 2012). According to official documents, bibliometrics affected the quality awarded in 'very few cases' (REF 2015, p. 51).[9] The submission of interdisciplinary research was encouraged, and publications falling outside the expertise of the Economics and Econometrics sub-panel were cross-referred to and assessed by experts in panels of other disciplines. However, nearly all outputs in our data (98.8%) were assessed by the Economics and Econometrics sub-panel.[10]

Each output was assigned to one of five mutually exclusive tiers, but the output-level classification was not disclosed. The quality depended on the number of awarded stars to distinguish among 'world leading' (four-star), 'internationally excellent' (three-star), 'internationally recognized' (two-star) and 'nationally recognized' (one-star) research. Submissions that fell short of national standards were flagged as 'unclassified quality'. The statistic used to allocate funding and rank institutions was the share of outputs listed in a submission that were assigned to each quality level. The lack of transparency regarding the classification of single outputs has fuelled discussion on how to identify four-star work.

The research quality in economics and econometrics was found to be outstanding, with more than two-thirds of submitted outputs being at least 'internationally excellent'. However, the recognized excellence exhibited substantial variation across institutions in the proportion of four-star outputs (as demonstrated in Table G.1 of the Online Appendix).

Related literature

Several studies have investigated past research assessments in the UK (for a review, see Traag and Watman 2019). Johnes *et al*. (1993) suggest that research ratings improve with the size and reputation of an institution. Clerides *et al*. (2011) conclude that departments might benefit from having members on the evaluation panel. Consistent with the latter study, De Fraja *et al*. (2019) suggest that institutions represented on the panel were awarded higher scores on the REF. They also show that the portability of outputs in REF2014 induced institutions to attract more-productive researchers by offering them higher salaries.

The REF generated intense debate over its regulatory framework and incentive structure. The mix of peer review and bibliometrics is often central to the discussion. Sgroi and Oswald (2013) show that research excellence could be predicted using the journal rankings and citations. Regibeau and Rockett (2016) argue that the journal impact and citations could identify the quality of economics departments without relying on reviews from experts. Peer review leaves the door open for subjective bias that may affect assessments. Hudson (2013) shows that, conditional on various proxies for research quality, experts prefer theory journals and outlets with a strong focus on economics.

The merits and limitations of using citations to rank journals in economics, and the influence on hiring and promotions in academia, have been discussed at length (e.g. Liebowitz and Palmer 1984; Laband and Piette 1994; Kalaitzidakis *et al*. 2003; Varin *et al*. 2016; Hamermesh 2018). A model-based approach to rank scientific outlets is shown in Bartolucci *et al*. (2015). In their work, the journal quality is unobserved, and indicators such as the AIS are used to proxy for this latent factor. We also address the unobserved quality awarded to each research output, which we infer from the total number of stars assigned to institutions.

Palacios-Huerta and Volij (2004) adopt an axiomatic approach to rank academic journals and demonstrate the optimality of the PageRank algorithm, which is also employed by Google to rank websites (Brin and Page 1998). We find that the AIS, which is based on the same methodology, is the strongest predictor of the research quality awarded by the REF panel.

II. Data and Bibliometrics

We accessed submissions for all institutions through the REF website. The data consist of 2600 outputs in Economics and Econometrics, most of which are journal articles (2388) and working papers (168). Starting from this information, we assigned to the corresponding journal all working papers that were flagged as forthcoming or published by August 2015.[11] Panel A of Table G.2 of the Online Appendix presents the breakdown by publication type that resulted from this selection. We collected citations for the journal and authors of each output to characterize the research influence and prestige. The Economics and Econometrics sub-panel had access to the citation counts for each publication made available from Elsevier's Scopus database in early 2014, and contextual data on the distribution of citations in the field and year of publication of the output. These files were deemed confidential and deleted at the end of the REF process, which makes replicating the results impossible. Therefore we approximated the bibliometric indicators available to panellists with the most similar indicators obtained from the web.

We characterized each journal by its AIS. This choice was motivated by a study that demonstrated that the AIS predicted expert-based evaluations of research quality (Hudson 2013).[12] We considered the AIS for 2013, which was the latest release at the time of the REF evaluation. We standardized the AIS to have zero mean and unit variance by the Thomson Journal Citation Reports (JCR) field to adjust for differences in citation behaviour across disciplines. Interdisciplinary research was encouraged warmly (REF 2011, p. 15); although 94% of submissions appeared in economics journals, the remaining outputs spanned fields such as psychology, mathematics and physics.[13] This finding suggests that economics has extramural influence on many other disciplines, which is consistent with the conclusions in Angrist *et al*. (2020).

The JCR database does not have universal coverage of journals submitted to the REF. In addition, working papers that are not forthcoming, and other research outputs (e.g. books or book chapters), cannot be attributed an AIS value. The distribution of outputs for which the AIS was retrieved, which is 91% of REF submissions, is shown in Figure G.1 of the Online Appendix. This distribution presents a long upper tail driven by high-impact outlets (such as the top five journals in economics) and spikes across the entire support. Table 1 reveals the origin of these spikes and lists all journals with at least 30 submissions in our sample (see columns (1) and (2)).

The citation count for each publication was obtained using Elsevier's Scopus, which is the source available to the REF panellists. We measured the citations at the end of 2013 (i.e. as of the REF submission date) and retrieved information for 2441 outputs (94% of the sample). Additionally, we considered Google Scholar because of its much larger array of publishing formats, although our conclusions are robust to the source of information employed.[14] This information was completed with the h-index of all authors at the time of the REF submission, which was computed from the Scopus database, and their reported affiliation in each output. Descriptive statistics for all bibliometrics are presented in panel B of Table G.2 of the Online Appendix. The citation counts and h-index in our analyses were

TABLE 1
ACADEMIC JOURNALS MOST FREQUENTLY SUBMITTED

			Estimated probability	
Journal	Frequency (1)	AIS (2)	4 stars (3)	3 stars (4)
4 stars				
Quarterly Journal of Economics	30	7.05	1.000	0.000
Econometrica	70	4.48	0.936	0.064
Review of Economic Studies	63	3.44	0.904	0.096
American Economic Review	115	2.64	0.889	0.111
Probable 4 stars				
Journal of the European Economic Association	73	1.48	0.621	0.379
Review of Economics and Statistics	59	2.16	0.613	0.387
Possible 4 stars				
Economic Journal	106	1.11	0.478	0.522
Journal of Econometrics	95	0.96	0.374	0.626
3 stars				
Journal of Monetary Economics	42	1.14	0.299	0.701
Journal of International Economics	37	0.90	0.273	0.727
Journal of Public Economics	57	0.77	0.175	0.825
International Economic Review	30	0.76	0.166	0.834
Journal of Economic Theory	84	0.78	0.149	0.851
Journal of Development Economics	50	0.66	0.112	0.888
Econometric Theory	35	0.59	0.103	0.897
Journal of Health Economics	33	0.40	0.001	0.999
Games and Economic Behaviour	83	0.32	0.000	0.915
European Economic Review	52	0.22	0.000	0.848
Journal of Money Credit and Banking	34	0.14	0.000	0.771
Probable 3 stars				
Journal of Economic Behavior & Organization	42	−0.07	0.000	0.529
Possible 3 stars				
Journal of Economic Dynamics and Control	45	−0.13	0.000	0.389
Economic Theory	49	−0.07	0.000	0.378
2 stars				
Economics Letters	63	−0.37	0.000	0.114

Notes
The table lists, in columns (1) and (2), journals with at least 30 submissions in economics and econometrics, together with their standardized AIS. Journal names are sorted by the estimated probability of scoring four stars, reported in column (3). Column (4) reports the estimated probability of scoring three stars. Journals are grouped by number of stars using the ranking methodology from Hudson (2013). See Section V for details.

adjusted for the publication year and field of study. Specifically, they are residuals from regressions on a full set of year of publication and field of publication dummies.

III. GRAPHICAL ANALYSIS

Although REF statistics are aggregated by institution, we show that they are informative on the classification of single outputs. We use this insight to show that the AIS of a journal predicts the number of stars awarded to individual outputs, and publications in high-impact journals appear to have solid four stars. We also show that although citation counts vary considerably across publications in the same journal, they are less effective than the AIS at predicting the REF performance of an institution.

AIS predicts the classification of research outputs

To fix ideas, we consider publications in the *Economic Journal* (*EJ*). If REF submissions in this journal exceeded the number of four-star outputs in at least one institution, then *EJ* publications must not have always been awarded four stars. Similarly, if the number of *EJ* submissions exceeded the number of outputs awarded one or two stars, then some *EJ* publications must have been awarded three stars or more. We use this idea to demonstrate that the likelihood of a top-scoring publication increases with the AIS of its journal.

Panel A of Figure 1 shows a value 1.1 for the AIS of the *EJ* (standardized). Among journals with AIS at or above 1.1, the vertical axis presents the percentage of publications that are not always awarded four stars. Since this value is zero, the data do not reject the claim that outputs in journals with AIS at least as large as that of the *EJ* may be always awarded four stars. For example, these journals include the *American Economic Review* (*AER*). Panel A replicates this analysis for all AIS values, and shows that a critical threshold emerges around the *Journal of Health Economics* (*JHE*). Specifically, the number of submitted *JHE* publications to the REF exceeds the number of four-star outputs in at least one institution, which implies that some *JHE* publications must have been awarded fewer than four stars. Panel A shows that the latter pattern is more likely for journals with AIS below that of the *JHE*. Panel B of Figure 1 strengthens this conclusion. For example, among journals with AIS at or above 2.6 (the value for the *AER*), approximately 30% of the publications were awarded three stars or more. The number is approximately 10% when the (standardized) AIS is zero.[15]

The relationship between output classification and AIS of the journal can be refined with assumptions on the quality of submitted outputs. For example, commonly economists assume that publications in the top five journals are always awarded four stars. These journals are highly respected and have standardized AIS values between 2.6 (*AER*) and 7.1 (*Quarterly*

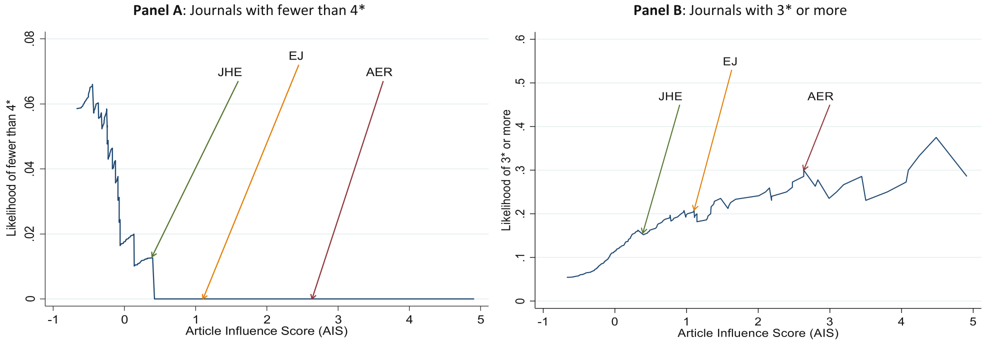

FIGURE 1. REF classification of outputs.
Notes: This figure considers submissions by journal as explained in Section III. Panel A shows the likelihood of being awarded fewer than four stars for publications in journals with standardized AIS values at or above a certain value. For example, the data do not reject the hypothesis that publications in journals with AIS values at least equal to those of the *AER* or the *EJ* are awarded four stars. The *JHE* represents a critical threshold. Panel B shows the likelihood of being awarded three stars or more for publications in journals with AIS values at or above a certain value. The support of the standardized AIS distribution is truncated at five because of the low number of journals above this value.

Journal of Economics), which are often considered a 'curse' because of their impact on visibility and career (Heckman and Moktan 2020). If this assumption is correct, then not all outputs published in top-field economics journals were awarded four stars. Figure G.2 of the Online Appendix, which replicates panel A of Figure 1, shows that submissions in the *Journal of Econometrics* (AIS 0.96) or *Journal of Economic Theory* (AIS 0.78) cannot be awarded four stars consistently.

What would be the critical AIS values to award stars if the outputs were classified entirely based on this bibliometric indicator? Figure 2 shows that these critical values are remarkably close to those emerging in Figure 1. For example, 20.2% (or 19) of the publications submitted by Queen Mary University of London were awarded four stars (see Table G.1 of the Online Appendix). We ranked all submissions in academic journals from Queen Mary University of London by AIS values, and defined the critical cut-off by considering the AIS of the 20th publication, which is the first that would be awarded three stars. The remaining critical cut-offs were determined similarly. By repeating the analysis for all institutions, panel A of Figure 2 shows the share of institutions that have reached the four-star threshold based on AIS values. We observe that publications with AIS at least as high as that of the *AER* would always score four stars. Publications with AIS lower than that of the *JHE* would always fall short of the critical cut-off to receive four stars. The critical journals that result from this analysis have AIS around that of the *EJ*, which represents the cut-off for a 'world leading' publication for 43% of institutions (12 of 28). The line in panel B, which is defined analogously for the three-star cut-off, shows that the *JHE* is critical for scoring 'internationally excellent'.

The strong predictive power of the AIS spills over to the correlation between the average AIS and the REF score of an institution. This correlation has also been documented in other studies (e.g. Traag and Watman 2019). Figure 3 plots the average AIS of all publications submitted by an institution against the Quality Index (QI) of the institution awarded by the REF (see the Notes on the figure for definitions). The predictions from a regression on a quadratic polynomial in AIS are superimposed, which suggests that the AIS provides a fair

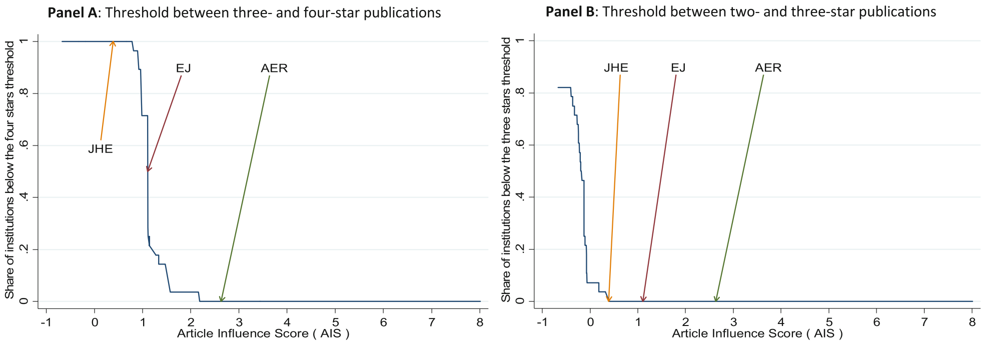

FIGURE 2. Counterfactual classification based on AIS.
Notes: Both panels are derived by ordering outputs by their AIS, from the highest to the lowest. A classification of outputs based solely on AIS is maintained throughout. For each institution, outputs with the highest AIS are assigned four stars proportionately to the REF classification in column (2) of Table E.1 of the Online Appendix. The remaining outputs are classified using columns (3), (4) and (5) of Table E.1 to assign three, two and one stars, respectively. Panel A shows the proportion of institutions that have passed the threshold for awarding four stars by value of AIS. For example, if classification were based solely on AIS, then all publications in journals at least equal to the *AER* would be awarded four stars. In 43% of institutions, the *EJ* would represent the pass mark between three and four stars. The *JHE* would determine publications below four stars in all institutions. Panel B shows the proportion of institutions that have passed the threshold for awarding three stars by value of AIS. For example, if classification were based on AIS, then the *JHE* would be the critical threshold. See Section III for details.

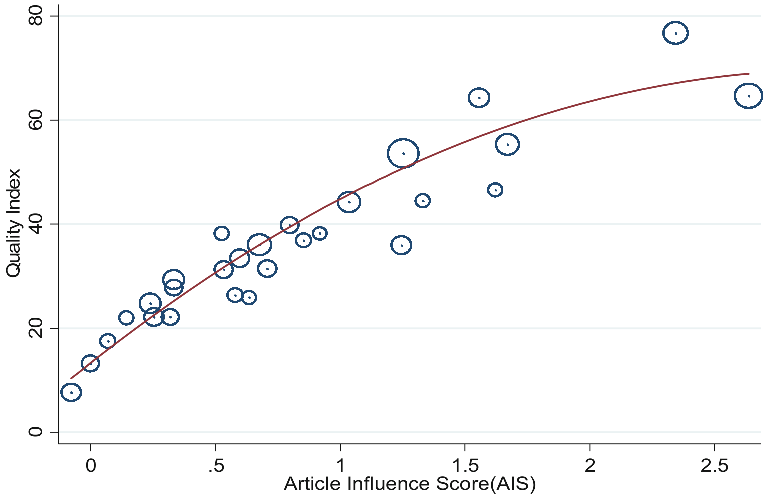

FIGURE 3. REF QI and AIS.
Notes: The figure reports the scatterplot of an institution's QI, on the vertical axis, against the average AIS of all outputs submitted by the institution. Bubbles are proportional to the number of outputs submitted. Superimposed are predictions from a regression on linear and quadratic terms in AIS weighted by the number of outputs submitted. The QI is computed using the current funding allocation formula, which depends on the incidence of top-quality outputs (80% and 20% to four- and three-star research, respectively, and no contribution of remaining outputs). The QI is computed considering only the research output element. See Section III for details.

approximation of the classification criteria followed by the panel. Obviously, it is important to understand whether the fit improves significantly after including additional output attributes such as citation counts. We address this empirical question in the next subsection.

The AIS predicts the final ranking from panellists, and correlates with how institutions made their strategic decisions about submissions. Specifically, we find that 55% of all submissions appeared in 23 journals with AIS much higher than the average. Thus the departments had similar expectations about what constitutes a good output, and that perceived quality is strongly correlated with the AIS. Specifically, panel A of Figure 4 shows that publications in journals with more than 30 entries in REF submissions account for 55% of the total (these journals are listed in Table 1). Panel B shows that these publications are concentrated in 23 journals (out of 283 in the REF), with average standardized AIS 1.32 (see panel C). Remarkably, we show in Section B of the Online Appendix that the number of publications in 7 of these 23 journals predicts almost perfectly the number of four-star outputs awarded to institutions (with R^2 at 98.7%).[16]

Citations do not predict the classification of research outputs

The citation counts increase with the journal AIS but present significant differences within the same journal (as found in Starbuck 2005; Anauati *et al*. 2020; Heckman and Moktan 2020). This trend can be observed from column (1) of Table G.3 of the Online Appendix, which shows the results from an output-level regression of Scopus citations on AIS, where we controlled for research field and publication year. We find that a one-standard-deviation (σ) increase in AIS is associated with a 0.164σ increase in citations, and the coefficient is highly significant. We find similar results when controlling for the average h-index of the authors (see column (3)). However, with R^2 at 28.5%, the AIS alone cannot explain most of the variability in citations.[17]

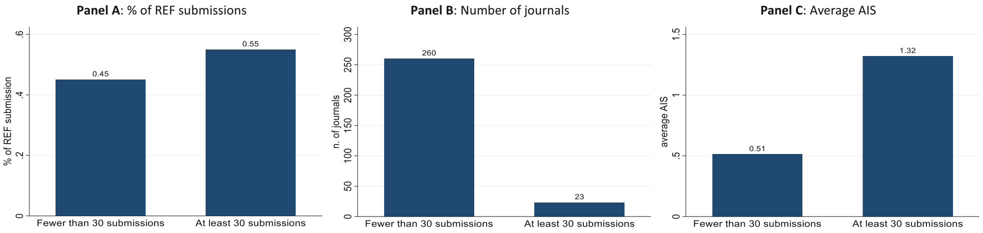

FIGURE 4. Concentration of REF submissions.
Notes: The figure reports statistics on academic journals in REF submissions. Bars depict separate results for journals with fewer than 30 outputs or with at least 30 outputs submitted to the REF. Panel A reports the share of total research outputs submitted. Panels B and C report the number of academic journals and their standardized AIS, respectively. The average is reported on top of the bars. See Section III for details.

Are citation counts more informative than the AIS in predicting the overall performance of institutions in the REF? We find that the answer to this question is no. Specifically, a regression (Table G.4 of the Online Appendix) of the QI of an institution on the average number of citations and average h-index of the authors of submitted outputs yields an R^2 value of 57%. In this regression, only the coefficient on citations is statistically significant. After including the average AIS of submitted outputs, R^2 increases to 89%. However, in the latter specification, only the coefficient on AIS remains statistically significant, which suggests that citations are uncorrelated with the REF classification once the journal impact is controlled for.

IV. EMPIRICAL SPECIFICATIONS

We now motivate the empirical specifications in our analysis. These specifications lay out the relationship between variables in public-use data and unknown parameters in equations that describe how stars are awarded to outputs. The takeaways from these specifications are summarized at the end, and more technical details are in Sections C and D of the Online Appendix.

Relationship between data and unobserved quantities

The REF outputs were published in 283 journals.[18] The number of submissions from institution i in journal j is X_{ij}, where $i = 1, \dots, 28$ and $j = 1, \dots, 283$. The terms X_{ij} for institution i are contained in the 283×1 vector $\mathbf{X}'_i$. This vector is retrieved from REF publications. Additionally, we collected data on the attributes of the submitted outputs (e.g. citations, field, and h-index of the authors). Let $\mathbf{Z}'_{jk}$ be the vector that contains the attributes of output k published in journal j; $\mathbf{Z}'_i$ contains the attributes of all submissions by institution i.

Variable D_{jk} denotes the number of stars awarded to output k in journal j.[19] Variable D_{jk} is *not* observed in REF data. However, we can still write

$$Y_i^d = \sum_{j=1}^{283} \sum_{k=1}^{X_{ij}} \mathbb{1}(D_{jk} = d), \tag{1}$$

which is the number of publications awarded d stars by the REF panel for institution i, where $d = 1, \dots, 4$. The measurements Y_i^d are contained in the 4×1 vector $\mathbf{Y}'_i$, which can be retrieved from official REF publications. Our investigation uses information on variables $\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i$ to infer the stars awarded to single outputs.

General formulation of the problem

Equation (3) below is the starting point of our investigation and follows from the REF criteria. Official documents state that the output quality was assessed independently for all outputs. Thus the probability of awarding d stars to output k must depend on its attributes Z_{jk} and not on other outputs submitted by the institution. In econometric terms, this reasoning leads to the equation

$$E\left[\mathbb{1}(D_{jk}=d)|\mathbf{X}_i,\mathbf{Z}_i\right]=\alpha_j^d+\gamma_j^d Z_{jk}. \tag{2}$$

Using equation (2) in equation (1) yields the following system of equations for $d=1,\ldots,4$:

$$E\left[Y_i^d|\mathbf{X}_i,\mathbf{Z}_i\right]=\sum_{j=1}^{283}\alpha_j^d X_{ij}+\sum_{j=1}^{283}\gamma_j^d\sum_{k=1}^{X_{ij}}Z_{jk}. \tag{3}$$

The estimation is challenged by the large number of unknown parameters. We address this problem by imposing restrictions guided by the graphical analysis in the previous section.

Adopted parametrization

We group 283 journals into three mutually exclusive tiers depending on the AIS. We use the *EJ* as the lower limit for a top tier (Tier 1) that will include four-star outputs with high probability. Then a middle tier is defined, Tier 2, which spans a grey area comprising a mix of four- and three-star publications. Finally, a bottom tier is defined, Tier 3, as the complement to all included journals above. Building on the graphical analysis in Section III, we use the *JHE* as the threshold to define the latter two tiers. We study the sensitivity of our conclusions to the definition of tiers in Section V.

Because three tiers are now used instead of 283 journals, equation (3) simplifies to the system of equations (5) below as follows. Specifically, journals are grouped by tier τ, where $\tau=1,2,3$. We assume that the deviation of α_j^d from the tier average $\alpha_{0\tau}^d$ depends on the journal characteristics (such as the AIS), which we denote by W_j. In addition, we impose constant effects of publication attributes within a tier. These two assumptions imply

$$\alpha_j^d=\alpha_{0\tau}^d+\alpha_{1\tau}^d W_j, \quad \gamma_j^d=\gamma_\tau^d, \tag{4}$$

for all j in tier τ. By substituting equation (4) into equation (3), we have

$$\begin{aligned}E\left[Y_i^d|\mathbf{X}_i,\mathbf{Z}_i\right]=&\sum_\tau\alpha_{0\tau}^d\left(\sum_{j\in\tau}X_{ij}\right)+\sum_\tau\alpha_{1\tau}^d\left(\sum_{j\in\tau}W_jX_{ij}\right)\\&+\sum_\tau\gamma_\tau^d\left(\sum_{j\in\tau}\sum_{k=1}^{X_{ij}}Z_{jk}\right).\end{aligned} \tag{5}$$

Two restrictions are imposed to reduce further the number of unknown parameters. First, we assume that the outputs in Tiers 1 and 2 are always awarded at least three stars. We also impose that the outputs in Tier 3 can never be awarded four stars. These restrictions are consistent with the graphical analysis in Section III, and allow for errors in the classification of outputs of at most one star.[20]

TABLE 2
CLASSIFICATION RESTRICTIONS

	Classification allowed			
	4*	3*	2*	1*
Tier 1	×	×		
Tier 2	×	×		
Tier 3		×	×	×

Notes
This table summarizes the restrictions imposed in estimation. For example, we assume that publications in Tier 1 journals are awarded four or three stars. See Section IV for details and definitions of journal tiers.

Summary of the estimation approach

We use all outputs to estimate the system of equations (5).[21] The four equations in this system are regressions of Y_i^d ($d = 1, \ldots, 4$) on the number of submitted outputs in each tier $\left(\sum_{j\in\tau} X_{ij}\right)$, a term that involves the interactions between submitted publications and journal characteristics (e.g. AIS) in each tier $\left(\sum_{j\in\tau} W_j X_{ij}\right)$, and a term that represents the attributes (e.g. citations) of the submitted publications in each tier $\left(\sum_{j\in\tau}\sum_{k=1}^{X_{ij}} Z_{jk}\right)$. Estimation is performed using seemingly unrelated regressions to account for the correlation between dependent variables. This approach yields the estimates of the parameters $\alpha_{0\tau}^d$, $\alpha_{1\tau}^d$ and γ_τ^d in equation (5), which we use to compute the probabilities of awarding stars. For example, the probability that publication k in journal j is awarded four stars is

$$E\left[\mathbb{1}(D_{ijk} = 4) \mid \mathbf{X}_i, \mathbf{Z}_i\right] = \alpha_{0\tau}^4 + \alpha_{1\tau}^4 W_j + \gamma_\tau^4 Z_{jk},$$

which depends on the characteristics W_j (e.g. AIS) of a journal and attributes Z_{jk} (e.g. citations) of a publication.

We constrain the probabilities of four, three, two and one stars across tiers as explained in the previous section. These constraints are summarized in Table 2, where cells without crosses indicate that the probabilities are set to zero. For example, these constraints imply that a 'bad' publication in the *AER* (Tier 1) is never worth fewer than three stars. They also permit the presence of four-star gems in the grey area identified by Tier 2 (which comprises many top-field journals). Two stars is, instead, the expected valuation for outputs in Tier 3, although 'bad' publications and hidden gems in this lowest tier may revise expectations either way by at most one star. In the econometric analysis, the constraints in Table 2 are obtained by constraining the parameters $\alpha_{0\tau}^d$, $\alpha_{1\tau}^d$ and γ_τ^d in the estimation of equation (5).

V. RESULTS

The econometric analysis demonstrates that the journal tier based on the AIS strongly predicts the REF classification. Adding output characteristics such as citations and field does not change this conclusion. Additionally, we obtain a ranking of the most submitted journals implied by our analysis.

Baseline specifications

We start from a specification that controls for the AIS only by stratifying on tiers. Specifically, we estimate equation (5), which imposes $\gamma_\tau^d = 0$ and $\alpha_{1\tau}^d = 0$ for all τ and all d. Using this

TABLE 3
ESTIMATION RESULTS FROM THE BASELINE MODEL

	Without adjustment—tiers only			With adjustment (AIS)		
	Tier 1 (1)	Tier 2 (2)	Tier 3 (3)	Tier 1 (4)	Tier 2 (5)	Tier 3 (6)
4* (world leading)	0.890***	0.064		0.801***	0.161***	
	(0.043)	(0.082)		(0.037)	(0.060)	
3* (internationally excellent)	0.110**	0.936***	0.491***	0.199***	0.839***	0.507***
	(0.043)	(0.082)	(0.033)	(0.037)	(0.060)	(0.022)
2* (internationally recognized)			0.436***			0.422***
			(0.027)			(0.018)
1* (nationally recognized)			0.074***			0.071***
			(0.012)			(0.012)
Number of journals in tier	48	31	205	48	31	205
Number of publications in tier	784	551	1265	784	551	1265
Mean of standardized AIS in tier	2.57	0.76	−0.09	2.57	0.76	−0.09

Notes
Columns (1)–(3) show results from a baseline specification that allows for tier-specific intercepts. Columns (4)–(6) show results from regressions that control for within-tier heterogeneity using a quadratic polynomial in AIS. Tier 2 is defined as in Section IV, from the *JHE* to the *EJ*. Columns (1) and (4) show the estimated probabilities that a publication in Tier 1 is awarded four or three stars. Columns (2) and (5) show the estimated probabilities that a publication in Tier 2 is awarded four or three stars. Columns (3) and (6) show the estimated probabilities that a publication in Tier 3 is awarded three stars, two stars or one star. The estimating equations are discussed in Section IV, and Section D of the Online Appendix.
***, **, * indicate $p < 0.01$, $p < 0.05$, $p < 0.1$, respectively.

specification, columns (1)–(3) of Table 3 show the estimated probabilities of the number of stars in each tier. Tier 2 outputs are defined as those with AIS values between those of the *JHE* and the *EJ* (the corresponding interval is 0.4–1.1).

We find that Tier 1 and Tier 2 outputs are awarded four and three stars, respectively, with very high probability (89% and 93.6%). We cannot reject the hypothesis that Tier 2 publications are never awarded four stars, since the value 6.4% in column (2) of Table 3 is not statistically different from zero. The classification of Tier 3 outputs is more nuanced, with similar chances of scoring two and three stars (see column (3)). Similar conclusions are obtained after adjusting for within-tier differences in AIS across outputs. Specifically, columns (4)–(6) of Table 3 show the estimated probabilities from equation (5), which imposes $\gamma_\tau^d = 0$ and enables $\alpha_{1\tau}^d \neq 0$ when W_j includes a quadratic polynomial in AIS. The numbers in columns (4)–(6) are probabilities for the outputs with average AIS value in tier.[22] We conclude that Tier 1 and Tier 2 outputs are solid four and three stars, respectively.

Several checks documented in Section E of the Online Appendix study the robustness of this conclusion to the definition of tiers. For example, Figure E.2 of the Online Appendix shows the change in estimated probabilities of scoring four stars in Tiers 1 and 2 when the upper limit of Tier 2 is moved around the *EJ*. In this figure, any number n on the horizontal axis indicates that the upper limit of Tier 2 is defined using a journal n ranks apart from the rank of the *EJ*. The top line of Figure E.2 suggests that restricting Tier 1 to journals with AIS higher than that of the *EJ* would reinforce our conclusions on the probability of scoring four stars in this tier. Tier 2 would also be more likely to include four-star publications, as shown in the bottom line of the figure. Meanwhile, including journals with AIS lower than that of the *EJ* in Tier 1 would result in lower probabilities of four-star publications in this tier, but the probability remains above 80%. The inclusion of these journals in Tier 1 would affect negatively the probability of four-star publications in Tier 2.

We conclude that publications in journals with AIS higher than that of the *EJ* are classified unambiguously as 'world leading' (four-star). On the other hand, there exists a grey area for journals with AIS lower than that of the *EJ*, where the classification is more ambiguous. As we will see, this area includes a number of top-field journals, notably the *Journal of Econometrics* and the *Journal of Economic Theory*.

Effect of the output characteristics

What is the role of output characteristics such as citations and field of publication? To answer this question, we estimate equation (5) with $\alpha^d_{1\tau} \neq 0$ as shown in columns (4)–(6) of Table 3. In addition, $\gamma^d_\tau \neq 0$ when Z_{jk} includes the outputs citations, average h-index of authors, and field. The predicted probabilities for outputs with the average value of Z_{jk} in tier are remarkably similar to the baseline estimates, as shown in Table G.6 of the Online Appendix. For example, the probability that outputs in Tier 1 are awarded four stars is 80.1% in column (4) of Table 3. The corresponding probability is 77.5% after controlling for citations and h-index of authors in column (7) of Table G.6. A similar pattern emerges for all remaining probabilities.

We find that the output characteristics affect only marginally the probability of being awarded more stars, which is conditional on tier membership. Specifically, Figure 5 shows the effects of a 10% increase in number of citations and h-index compared to their tier averages. We show the effects and confidence intervals on the probability of being awarded the highest score: four stars in Tiers 1 and 2, three stars in Tier 3. For example, more citations are not associated with a better chance of scoring four stars in Tier 1; publications in this tier with authors whose h-index is 10% higher than the average have approximately 1% additional chance of being awarded four stars (from a baseline of 77.5%). The marginal

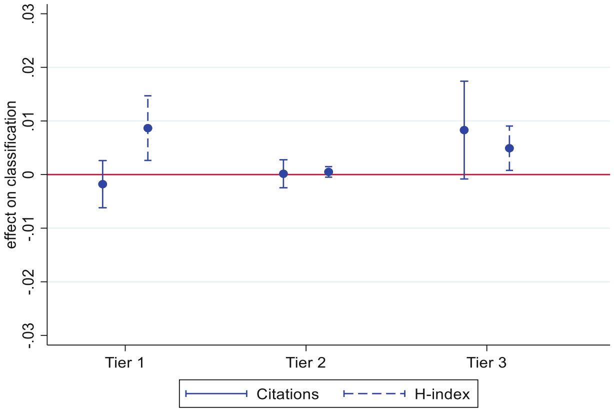

FIGURE 5. Citations and h-index effects.

Notes: Reported are the estimated effects on the probability of scoring four stars (in Tiers 1 and 2) or three stars (in Tier 3) from a change in citations (solid lines) or h-index (dashed lines). The effects correspond to an increase of 10% from the tier-specific average of citations count or h-index. The 95% confidence intervals are obtained from the specifications in columns (7)–(9) of Table G.6 of the Online Appendix. Citation count and h-index are measured as residuals from a publication-level regression on year and field fixed effects. The h-index of a publication is the highest value among all co-authors. See Section V for details.

association between citations and assessments is consistent with the fact that panellists used this information in only a limited number of cases (see REF 2015, p. 51).

Figure 5 also shows that lower AIS values are associated with larger heterogeneity in the effects of other output attributes that contribute to the final classification by panellists. This finding suggests that peer reviews may be a cost-effective assessment for publications that are not in journals with unambiguously good bibliometrics. For example, the role of citations is relatively more important in Tier 3. However, these effects are estimated imprecisely at approximately 1% (from a baseline of 55.1% in column (7) of Table G.6 of the Online Appendix).

Our conclusions are robust to further checks, as discussed at length in Section F of the Online Appendix. First, REF panellists assess the traits of research quality, such as significance and rigour, which are unobservable to us. We show that the results of our analyses are unlikely to be driven by such unobservable traits. Second, we show that the errors committed by using our model to predict the outcomes in equation (5), $Y_i^d - \hat{Y}_i^d$, are unrelated to the indicators of research performance in the 2008 Research Assessment Exercise (the predecessor of the REF), characteristics of higher education institutions evaluated by the REF (e.g. non-academic impact and research environment), and measures of ties between panellists and institutions.

Journal ranking

Given the proliferation of rankings and their role in personnel decisions, we report in Table 1 the predicted probabilities of scoring three or four stars for the journals with the most submissions to the REF. We use all submitted outputs to estimate equation (5), and report the probabilities only for journals with at least 30 submissions. Predictions are reported for an output at the mean citation count of each journal. To ensure comparability with previous studies (Hudson 2013; Hole 2017), the journals are grouped depending on values of predicted probabilities. Unambiguously four-star journals are those with at least a 65% probability of having a 'world leading' classification. For probable and possible four-star journals, this probability must be larger than 50% and 35%, respectively. The same definitions are used to rank three-star and two-star journals.

Our results suggest that there is little space for 'bad' outputs in top five or generalist journals, for which our classification is unambiguously four-star (see columns (3) and (4) of Table 1). The classification at lower values of AIS becomes less clear-cut, and the *EJ* and *Journal of Econometrics* are examples of possible four-star journals. Publications in top-field outlets such as the *Journal of Economic Theory* and *Journal of Public Economics* are most likely awarded three stars.

VI. Conclusion

In many countries, disciplinary panels of experts have access to bibliometrics that can be used to inform their assessments. The following practical question emerges: does the ranking of journals by informed experts mirror objective indicators of journal influence that can be attained more frequently and at much lower costs?

We have used the REF to reconsider this issue given the resistance to bibliometric assessments by the academic community in the months preceding national evaluations (see Wilsdon *et al*. 2015; UK Forum for Responsible Research Metrics 2018). Unfortunately, this exercise is not straightforward because the classification of outputs in the REF is not disclosed. This lack of disclosure has fuelled discussion in the UK academic community regarding the determinants of top-scoring (four-star) outputs beyond these indicators of influence and citations, which are widely available.

We have shown that a thorough but expensive national assessment like the REF yields a classification of economics and econometrics outputs that can be replicated, at least partially, using 'crude but cheap' assessments based on bibliometric indicators of journal impact. Specifically, we have shown that publications in journals with high Thomson Reuters Article Influence Score (AIS) are almost certainly awarded four stars by REF panellists. There are 1335 such publications, which we grouped in Tier 1 and Tier 2 journals in Section V, representing 51% of submissions to REF2014.[23] On the other hand, we have shown that the quality awarded by REF experts to publications in journals with relatively lower impact is more difficult to predict and depends on other publication attributes such as citations.

Our findings suggest that revisions of national assessments are possible, in fields like economics and econometrics, without altering the overall quality of the REF. Our view is that peer reviews and bibliometrics should be complementary modes of assessment. For example, at least part of REF submissions could be classified automatically using statistics like the AIS, while costly and time-consuming peer reviews could be targeted more efficiently at publications in journals with relatively lower impact and in interdisciplinary journals. A slimmed-down version of the current national assessments could reduce total costs substantially, which are estimated at £246 million for REF2014 and about 15% of the research funding distributed annually (see Stern 2016; Research England 2020).

Our findings raise a number of additional questions, including the choice of the most appropriate metrics to identify unambiguously top-scoring journals and to review outputs outside this tier only. The external validity of our findings in fields different from economics and econometrics should be investigated as well. It is also important to consider the non-monetary costs associated with any changes to how the REF is conducted currently, for example because these changes could exacerbate measurement problems in the assessment of how research quality has changed over time. We hope to address some of these challenging questions in future work.

ACKNOWLEDGMENTS

Our thanks go to Graziella Bertocchi, Daniele Checchi, Francesco Fasani, Andrew Oswald, Barbara Petrongolo, Ludovic Renou and James Wilsdon for helpful discussions and comments, to two referees for insightful comments, and to seminar participants at the XVI Brucchi Luchino Workshop, the 2018 Annual Conference of the Royal Economic Society, and the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR). Any responsibility for the views expressed in the article rests solely with the authors.

NOTES

1. The results of the 2021 assessment were published in May 2022; see https://results2021.ref.ac.uk (accessed 8 July 2022).
2. Citation counts depend naturally on field of study and time elapsed since publication. In our analysis, we use a measure of citations that adjusts for the publication year and field of study.
3. In the Italian ANVUR national assessment, the Economics and Econometrics sub-panel relies on bibliometrics to assess articles in scientific journals, and assesses all other outputs by peer review, with a full disclosure of the final outcomes to the author of the submitted output. Checchi *et al*. (2019) find that the number of three- and four-star outputs in the REF can be predicted almost perfectly by the bibliometric algorithm of ANVUR. However, their work considers the institutional-level score and does not disentangle the contribution of the journal impact and output-level citations to the REF classification.
4. Stern (2016, p. 21) also reports that 'bibliometric evidence could be useful to panels in determining whether there is a significant discrepancy between the grade profile for outputs . . . as determined by peer review, and citation data'. Similar recommendations are in the Leiden Manifesto (Hicks *et al*. 2015) and the San Francisco

Declaration on Research Assessment (see https://www.ascb.org/files/SFDeclarationFINAL.pdf, accessed 6 July 2022).

5. These bodies are Research England, the Scottish Funding Council, the Higher Education Funding Council for Wales and the Department for the Economy, Northern Ireland. See https://www.ref.ac.uk/about/whatref (accessed 6 July 2022) for a description of the REF.
6. Impact is defined as 'an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia' (REF 2011).
7. However, the amount of public funding depends on the number of academics submitted. One additional complication that affects the representativeness of research is that institutions can submit staff strategically to different units of assessment. For example, within economics departments, staff may be submitted to the Business and Management Studies sub-panel, where the subject matter overlaps, and this panel would likely rank the output more highly.
8. All staff with significant responsibilities for research must be submitted to REF2021. Institutions may return the outputs of staff employed previously if publications were generated during the period of employment (REF 2017, pp. 50–1). The weight assigned to research outputs in REF2021 remains predominant (60%) but is lower than that in REF2014. The contribution of research impact increases to 25%.
9. See https://www.ref.ac.uk/2014 (accessed 6 July 2022) for a description of the rules, outcomes and identity of the panellists.
10. REF reports are accessible at https://www.ref.ac.uk/2014/panels/paneloverviewreports (accessed 6 July 2022). They show that cross-reference to other panels was limited to 2.7% of outputs. The outputs cross-referred to the Economics and Econometrics sub-panel were excluded from our analysis.
11. These are outputs for which we retrieved the publication status at the end of 2016. The implicit assumption here is that a working paper published by August 2015 must have been accepted for publication while the REF panel was at work. Our choice explains the small differences in submission counts by journal in Table 1 with respect to the official counts, which can be found in Hole (2017), among others.
12. A ranking of economics journals by the AIS can be accessed at bit.ly/2Q00Eld (accessed 6 July 2022). The Italian ANVUR national assessments use the SCImago Journal Rank index (Gonzalez-Pereira *et al*. 2010), which is constructed similarly to the AIS. In our sample, the Spearman correlation between the two indicators of journal impact is 94%.
13. When a journal is assigned multiple fields (e.g. economics and statistics), standardization was performed using the mean and standard deviation across all economics journals. In rare cases where all fields are outside economics, we considered the average of the field-specific standardized scores. The classification of fields available to REF panellists uses the Elsevier All Science Journal Classification categories. However, to standardize the AIS, we used the Thomson JCR categories. These alternative classifications are equivalent substantially for submissions in economics and econometrics.
14. The Scopus and Google Scholar archives do not cover the same population of journals and publishers. The correlation between the two citation measurements is 86%, as computed from 2441 outputs.
15. The quantities in Figure 1 need not be monotone in AIS, which explains the sawtooth pattern. As we explain in Section A of the Online Appendix, similar conclusions are obtained by considering combinations (e.g. pairs or triplets) of journals submitted to the REF instead of considering individual journals; see Figure A.1 of the Online Appendix.
16. These journals are (ordered from the largest AIS): *Econometrica*, *Review of Economic Studies*, *American Economic Review*, *Journal of Monetary Economics*, *Economic Journal*, *Journal of Development Economics and Econometric Theory*.
17. Since the distribution of citations is heavily skewed, we investigate whether our conclusions are driven mechanically by the linear fit. Following Card and Della Vigna (2020), we further estimate a model for the inverse hyperbolic sine of citations in columns (2) and (4) of Table G.3 of the Online Appendix. Although R^2 increases substantially (to 57%), a large part of variability in citations remains unexplained. Columns (5)–(8) of Table G.3 present the estimates using Google Scholar citations, which yield a similar conclusion.
18. In this section, we work as if all outputs were articles in journals with AIS. We show in Section C of the Online Appendix how we account for other outputs such as books or book chapters.
19. We set $D_{jk} = 0$ for 'unclassified' outputs. The number of unclassified outputs is 16, which corresponds to 0.6% of submissions. Because of this low number, we consider only four tiers $D_{jk} \in \{1, 2, 3, 4\}$ hereafter.
20. This assumption is made frequently in empirical work on misclassification; e.g. Battistin and Sianesi (2011) and references therein. More generally, these restrictions on the classification probabilities mirror the results from past research on the informational content of international lists for journal rankings. By comparing various bibliometric indicators with the views of experts, Hudson (2013) concludes that some journals could be unambiguously clustered with respect to the number of awarded stars. This classification is fuzzy ('probable' and 'possible' is his narrative) in other cases. Table G.5 of the Online Appendix shows that our definition of Tiers 1 and 2 coincides with that of unambiguously three-star or higher outputs in Hudson (2013), and most outputs in Tier 3 are expected to have a two-star classification. Similar conclusions emerge by considering other research on the REF (Hole 2017).
21. We assign the outputs for which AIS is not available, such as book chapters, to tiers by considering the publishing editor, as detailed in Section C of the Online Appendix. Our conclusions are robust to various alternative choices for assigning outputs without AIS to tiers.

22. The relationship between estimated probabilities and AIS is shown in Figure G.3 of the Online Appendix, which shows a more pronounced within-tier heterogeneity at the bottom end of the distribution. In addition, our specification imposes a continuity of the classification probabilities for journals at the boundaries between tiers. For example, we impose that a publication in the *EJ* (which marks the lower end of Tier 1) has a probability of being awarded four stars equal to that for a publication in the journal with the highest AIS in Tier 2. All specifications include controls for books, book chapters and other outputs as explained in Section C of the Online Appendix. The estimates that are not reported here suggest that the books published by international editors (see Table C.1 of the Online Appendix) were most likely awarded four stars, and book chapters most likely receive three stars. Our analysis does not reveal any clear pattern for the field coefficients, which perhaps reflects that 94% of submissions were in economics journals.
23. The correlation between subjective assessments of research quality and bibliometric indicators has been documented in past research that investigates large-scale assessments in the UK (Clerides *et al*. 2011; Taylor 2011; Traag and Watman 2019) and other countries (see, for example, Bertocchi *et al*. 2015). This result is often used to advocate for metric-based evaluations as opposed to peer reviews, to reduce the administrative burden and risk of bias (see, for example, Laband 2013).

REFERENCES

ANAUATI, M. V., GALIANI, S. and GALVEZ, R. H. (2020). Differences in citation patterns across journal tiers in economics. *Economic Inquiry*, **58**(3), 1217–32.

ANGRIST, J., AZOULAY, P., ELLISON, G., HILL, R. and LU, S. F. (2020). Inside job or deep impact? Using extramural citations to assess economic scholarship. *Journal of Economic Literature*, **58**(1), 3–52.

BARTOLUCCI, F., DARDANONI, V. and PERACCHI, F. (2015). Ranking scientific journals via latent class models for polytomous item response data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **178**, 1025–49.

BATTISTIN, E. and SIANESI, B. (2011). Misclassified treatment status and treatment effects: an application to returns to education in the United Kingdom. *Review of Economics and Statistics*, **93**(2), 495–509.

BERTOCCHI, G., GAMBARDELLA, A., JAPPELLI, T., NAPPI, C. A. and PERACCHI, F. (2015). Bibliometric evaluation vs. informed peer review: evidence from Italy. *Research Policy*, **44**(2), 451–66.

BRIN, S. and PAGE, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, **30**(1–7), 107–17.

CARD, D. and DELLA VIGNA, S. (2020). What do editors maximize? Evidence from four economics journals. *Review of Economics and Statistics*, **102**(1), 195–217.

CARPENTIER, V. (2021). Three stories of institutional differentiation: resource, mission and social inequalities in higher education. *Policy Reviews in Higher Education*, **5**(2), 197–241.

CHECCHI, D., CIOLFI, A., DE FRAJA, G., MAZZOTTA, I. and VERZILLO, S. (2019). *Have you read this? An empirical comparison of the British REF peer review and the Italian VQR bibliometric algorithm. CEPR Discussion Paper no. 13521*.

CLERIDES, S., PASHARDES, P. and POLYCARPOU, A. (2011). Peer review vs metric-based assessment: testing for bias in the RAE ratings of UK Economics departments. *Economica*, **78**, 565–83.

DE FRAJA, G., FACCHINI, G. and GATHERGOOD, J. (2019). Academic salaries and public evaluation of university research: evidence from the UK Research Excellence Framework. *Economic Policy*, **34**(99), 523–83.

GONZALEZ-PEREIRA, B., GUERRERO-BOTE, V. P. and MOYA-ANEGON, F. (2010). A new approach to the metric of journals scientific prestige: the SJR indicator. *Journal of Informetrics*, **4**(3), 379–91.

HAMERMESH, D. S. (2018). Citations in economics: measurement, uses, and impacts. *Journal of Economic Literature*, **56**(1), 115–56.

HECKMAN, J. J. and MOKTAN, S. (2020). Publishing and promotion in economics: the tyranny of the top five. *Journal of Economic Literature*, **58**(2), 419–70.

HICKS, D., WOUTERS, P., WALTMAN, L., DE RIJKE, S. and RAFOLS, I. (2015). The Leiden Manifesto for research metrics. *Nature*, **520**, 429–31.

HIGHER EDUCATION FUNDING COUNCIL FOR ENGLAND (2015). *The Metric Tide: Correlation Analysis of REF2014 Scores and Metrics*. Bristol: HEFCE.

HOLE, A. R. (2017). Ranking economics journals using data from a national research evaluation exercise. *Oxford Bulletin of Economics and Statistics*, **79**(5), 621–36.

HUDSON, J. (2013). Ranking journals. *Economic Journal*, **123**, F202–F222.

JOHNES, J., TAYLOR, J. and FRANCIS, B. (1993). The research performance of UK universities: a statistical analysis of the results of the 1989 Research Selectivity Exercise. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **156**, 271–86.

KALAITZIDAKIS, P., STENGOS, T. and MAMUNEAS, T. P. (2003). Rankings of academic journals and institutions in economics. *Journal of the European Economic Association*, **1**(6), 1346–66.

LABAND, D. (2013). On the use and abuse of economics journal rankings. *Economic Journal*, **123**, F223–F254.

——— and PIETTE, M. J. (1994). The relative impacts of economics journals: 1970–1990. *Journal of Economic Literature*, **32**(2), 640–66.

LIEBOWITZ, S. and PALMER, J. (1984). Assessing the relative impacts of economics journals. *Journal of Economic Literature*, **22**(1), 77–88.

MOED, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, **34**(8), 575–83.

PALACIOS-HUERTA, I. and VOLIJ, O. (2004). The measurement of intellectual influence. *Econometrica*, **72**(3), 963–77.

PITT, M. and YAN, Z. (2017). *How does the REF panel perceive journals? A new approach to estimating ordinal response model with censored outcomes*. Unpublished manuscript, University of Warwick.

REGIBEAU, P. and ROCKETT, K. E. (2016). Research assessment and recognized excellence: simple bibliometrics for more efficient academic research evaluations. *Economic Policy*, **31**(88), 611–52.

RESEARCH ENGLAND (2020). Research and knowledge exchange funding for 2020 to 2021; available online at https://re.ukri.org/finance/annual-funding-allocations/annual-funding-allocations-2020-21 (accessed 6 July 2022).

RESEARCH EXCELLENCE FRAMEWORK (REF) (2011). Assessment framework and guidance on submissions; available online at https://www.ref.ac.uk/2014/pubs/2011-02 (accessed 6 July 2022).

——— (2012). Panel criteria and working methods; available online at https://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf (accessed 6 July 2022).

——— (2015). REF2014: Overview report by main panel C and sub-panels 16 to 26; available online at https://ref.ac.uk/2014/panels/paneloverviewreports (accessed 6 July 2022).

——— (2017). Initial decisions on the Research Excellence Framework 2021; available online at https://www.ref.ac.uk/publications/2017/initialdecisionsontheresearchexcellenceframework2021.html (accessed 6 July 2022).

SGROI, D. and OSWALD, A. J. (2013). How should peer-review panels behave? *Economic Journal*, **123**, F255–F278.

STARBUCK, W. H. (2005). How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science*, **16**(2), 180–200.

STERN, N. (2016). *Building on Success and Learning from Experience: An Independent Review of the Research Excellence Framework*. London: Department for Business, Energy and Industrial Strategy.

TAYLOR, J. (2011). The assessment of research quality in UK universities: peer review or metrics? *British Journal of Management*, **22**, 202–17.

Thomson REUTERS (2014). *2013 Journal Citation Reports*. London: Clarivate Analytics.

TRAAG, V. A. and WATMAN, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications*, **5**, article 29.

UK FORUM FOR RESPONSIBLE RESEARCH METRICS (2018). UK progress towards the use of metrics responsibly: three years on from The Metric Tide report; available online at https://dera.ioe.ac.uk//31945 (accessed 6 July 2022).

UNIVERSITIES UK (2015). The economic role of UK universities; available online at https://issuu.com/universitiesuk/docs/theeconomicroleofukuniversities_c8b621395591ab (accessed 6 July 2022).

VARIN, C., CATTELAN, M. and FIRTH, D. (2016). Statistical modelling of citation exchange between statistics journals. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **179**, 1–63.

WILSDON, J., ALLEN, L. and BELFIORE, E. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management; available online at* https://doi.org/10.13140/RG.2.1.4929.1363 (accessed 6 July 2022).

ZACHAREWICZ, T., LEPORI, B., REALE, E. and JONKERS, K. (2019). Performance-based research funding in EU member states—a comparative assessment. *Science and Public Policy*, **46**(1), 105–15.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1: Supplementary Information