I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11198

Rising Stars

Erich Battistin
Marco Ovidi

DECEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11198

Rising Stars

Erich Battistin
Queen Mary University of London, CEPR, FBK-IRVAPP and IZA

Marco Ovidi
Queen Mary University of London

DECEMBER 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Rising Stars*

We use the UK's 2014 Research Excellence Framework (REF) to study which attributes characterize a top-scoring (four-star) publication in Economics and Econometrics. We frame the analysis as a classification problem and, using information in official documents, derive conditions to infer the unobservable score that panellists awarded to each publication. Juxtaposing institutions' submissions with REF outcomes provides information on the latent pass-marks used for assigning quality levels, which respond to journal prestige measured by the Thomson Reuters Article Influence Score. We explore this statistical feature in the econometric analysis, which reveals the limited contribution to awarded quality made by other publication attributes, possibly unobservable to us, conditional on the Article Influence Score. We conclude that, in large-scale and costly evaluations such as the REF, the time-consuming task of peer reviews should be devoted to publications not in academic outlets with unambiguously top-scoring bibliometric indicators of journal impact. Our model also predicts a ranking of academic journals consistent with the classification of REF panellists.

JEL Classification: H52 , H83 , I23 , I28

Keywords: education policy, higher education, journal ranking, research funding

Corresponding author:
Erich Battistin
School of Economics and Finance
Queen Mary University of London
Mile End Road
London E1 4NS
United Kingdom

E-mail: e.battistin@qmul.ac.uk

* Our thanks to Graziella Bertocchi, Francesco Fasani, Andrew Oswald, Barbara Petrongolo and James Wilsdon for helpful discussions and comments, and to seminar participants at the XVI Brucchi Luchino Workshop. Any responsibility for the views expressed in the article rests solely with the authors.

1 Introduction

A number of European countries have created agencies charged with evaluating higher education institutions and research organizations. Following standards and guidelines set by the ministers of the European Higher Education Area, the last decade has witnessed a surge in the number of national assessments conducted on a regular basis.[1] The *High Council for Evaluation of Research and Higher Education* (HCERES) in France, the *National Agency for the Evaluation of the University and Research Systems* (ANVUR) in Italy, and the *Higher Education Funding Council for England* (HEFCE) are examples of independent authorities holding higher education institutions accountable for their performance. Among the various dimensions considered, research quality provides the yardstick by which productivity and reputation are often assessed and compared. In the UK, the case study considered here, assessments have been conducted since 1986. The most recent results from the *Research Excellence Framework* (REF) were published at the end of 2014 to inform the allocation of public funding across the country's institutions. Approximately 40% of total funding currently depends on REF-measured performance (HESA, 2017).

One of the REF's accountability pillars is the quality assessment of research outputs, which is what we consider here. Institutions can submit multiple outputs (the vast majority of which are publications in academic journals) whose quality is assessed by external experts. The assessment follows general guidelines for originality, significance and rigour that are known to institutions at the time of submission. However, somewhat surprisingly and in contrast with other countries, the contribution of each output to an institution's awarded quality is not disclosed. Official documents report, for each institution, only the share of outputs classified in five mutually exclusive tiers, from "below the national standards" to "world leading". The number of stars awarded to each output, ranging from one to four, is commonly used as shorthand for this classification. This lack of full disclosure has spurred the proliferation of internal REFs to filter top-scoring (four-star) work in future submissions. Understanding how a publication's quality is assessed is therefore essential for the incentive structure faced by institutions in conducting their research.

We use REF 2014 official documents for the Economics and Econometrics sub-panel and

[1] See European Association for Quality Assurance in Higher Education (2005).

develop a statistical model to infer the latent "value" of publications awarded by panellists. Our population consists of micro-data for 2,600 outputs in 283 journals, authored by 2,597 scholars from 28 institutions in the UK. Multiple indexes of citations, bibliometric indicators of impact and information about authors at the time of the evaluation are used as objective proxies for the quality of each output. These proxies represent the closest approximation to the information made available to panellists to inform assessments in the evaluation process, as we explain below. Importantly, our analysis does not restrict publications in one journal to contribute equally to the final count of four-star outputs, marking an important difference from previous research (including but not exclusively on REF; see Hudson, 2013, Hole, 2017, and Pitt and Yan, 2017).

We start by showing that the final classification made by the panellists is strongly correlated with a journal's impact factor (which we measure with the Thomson Reuters Article Influence Score (AIS); see Clarivate Analytics, 2014). This statistical property is revealed from publicly available documents and does not rely on assumptions. For example, if the number of submissions in one journal exceeds the number of "world leading" outputs for at least one institution, then it must be that a publication in that journal is not consistently awarded a "world leading" ranking. Using this idea, we document a strong gradient in the relationship between the impact factor and output quality attributed by panellists, despite the latter being unobservable. This analysis also reveals that publications in top-field journals (the most prestigious outlets for articles in specialized fields of economics) need not be consistently awarded four stars. Imposing restrictions on the classification process further strengthens this finding. For example, we show that by assigning four stars to all outputs in top-five Economics journals, an assumption motivated by the discussion in Heckman (2017), top-field publications in econometrics, economic theory and health economics would be placed at the border between three and four stars. This finding is consistent with the role given to REF panellists, who should read and evaluate all submissions to identify overrated outputs and hidden four-star gems regardless of the publication outlet.

Building on this evidence, we use micro-data on all submissions and derive the identifying conditions to retrieve the latent classification of outputs. We develop a statistical model in which the output-specific contribution to institution quality is inferred from institution-level counts of the number of stars awarded. This marks an important departure from previous

studies on REF and is a contribution of this research.[2] REF panellists know the authors and publication outlet, which means that they are giving informed as opposed to double-blind reviews. The operationalization of guidelines to assess quality leaves the door open to the signalling effects of outputs published in top journals, and this raises an important practical question: do ranks from informed experts differ significantly from those that would have been obtained using widely accessible bibliometric indicators of influence?

The answer to this question, at least for the REF Economics and Econometrics sub-panel, is no. The estimation results from our model show that, in a large-scale and costly exercise such as the REF, panellists produce assessments broadly consistent with the Thomson Reuters AIS indicator of journal impact. Interestingly, AIS ranks journals with the same algorithm used to list the most relevant websites in a Google search (Brin and Page, 1998) and is similar to the optimal algorithm resulting from an axiomatic approach to determine intellectual influence based on citations (Palacios-Huerta and Volij, 2004). The role of bibliometrics is strengthened by the high correlation (75%) between the Scimago Journal Rank, an indicator similar to the AIS, and the number of stars awarded computed by HEFCE (2015) using output-level anonymized data for work published in 2008 in Economics and Econometrics.

We find that output attributes other than AIS, such as citation counts, the h-index of authors or the extent to which research focuses on the economics discipline, are not leading variables for predicting the quality awarded. For example, citation counts and h-index averaged by institution explain 57% of the variability in the REF quality score awarded to institutions. When the average AIS of outputs submitted is added to this regression, citation counts and h-index become statistically not significant while the explained variance increases to 89%. The analysis on microdata shows that citations matter only for outputs that are most likely to be on the border between one and two stars. These findings suggest that peer review may be a cost-effective assessment for publications that are not in unambiguously four-star journals, a fact consistent with the conclusions in Hudson (2013).

[2]The related study by Hole (2017) develops an algorithm to classify journals. We conduct a richer analysis using bibliometric indicators of journal impact and citations and state the assumptions required for the validity of our classification exercise. In addition, Hole (2017) considers only publications in journals with the most submissions in REF 2014. We instead use all research outputs, including those not in academic journals. A recent paper by Pitt and Yan (2017) on REF data addresses the problem of inferring the impact of unit-level observations (outputs) when the outcome of interest (number of stars awarded) is measured at a coarser level (the institution).

We document a lack of systematic differences between the REF indicators of institutional performance and the same indicators predicted using our model. These differences are uncorrelated (by construction) with a number of bibliometric variables that we control for in the analysis but also with indicators of research performance in the 2008 Research Assessment Exercise (the REF's predecessor), other characteristics of higher education institutions evaluated by the REF (non-academic impact and research environment), and measures of ties between panellists and institutions that are considered by Zinovyeva and Bagues (2015) and Colussi (2017). This finding prompts the use of our model to predict a ranking of academic outlets by the estimated probability of scoring four stars. The results for the most frequent journals in REF submissions, presented in Table 3 below, show that top generalist journals in Economics, including all top-five outlets, are unambiguously awarded four stars. In line with the descriptive evidence, the predicted likelihood of a "world leading" rating is substantially lower for a number of specialized journals regarded as top-field. These findings mirror those in Pitt and Yan (2017), showing that the main conclusions from our analysis cannot merely be an artefact of the statistical model we employ.

Our estimation faces empirical challenges that we address by imposing some important restrictions. First, we maintain the assumption that the quality awarded to each output is independent of the (observable and unobservable) attributes of other outputs submitted by the institution and their number. In other words, we assume that each output is classified only on the basis of its characteristics and independently of the bundle submitted by the institution. We see this as a mild assumption, which is consistent with the idea that panellists should perform an independent assessment of all 2,600 outputs submitted (see par. 126 in REF, 2012). We show that this criterion implies an exclusion restriction in our model, which we use for identification.

Second, our measure of a publication's quality is spanned using the number of citations, journal impact and characteristics of the authors at the time of REF evaluation. This makes it possible that our results are driven by output rigour and originality as assessed by experts, beyond bibliometrics. Our conclusions survive a series of sensitivity checks designed to address this problem. We maintain the assumption that a publication with "many" citations (i.e., a broad measurement of its influence in the profession) is not of lower quality than any other publication in the same journal with "few" citations. This imposes a monotonicity con-

dition between unobserved quality and citations that we use in the estimation to account for unobservables. Building on this idea, we show that our conclusions cannot be the mechanical consequence of omitted unobservables related to the quality of the outputs submitted.

Finally, REF documents (and Table 1 below) demonstrate that only a subset of the array of publications submitted by institutions must contribute significantly to the share of top-scoring outputs. It follows that the relationship between the share of four-star outputs and the array of submissions is sparse (Hastie et al., 2015), in the sense that most elements of the array must contribute with zero weight. Sparsity, combined with the small number of institutions involved, calls for some parametric assumptions that we discuss in the empirical section.

The fact that bibliometrics explain REF rankings should be of interest to policy makers studying the regulatory framework of future evaluations, in the UK and other countries. Outlining the path to the next REF, Lord Stern's review (Stern, 2015) suggests a responsible use of bibliometrics: the tide of quantitative indicators should be handled with care and should not replace peer-reviewed evaluations (Wilsdon et al., 2015).[3] However, Lord Stern's review also reports, "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" (see page 21). Our findings suggest that, in Economics and Econometrics, peer reviews and bibliometrics should be viewed as complementary modes of assessment to identify unambiguously top-scoring journals and review only outputs outside this tier. The time and resources 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. In the Italian ANVUR national assessment, for example, the Economics and Econometrics panel relies on bibliometrics for articles in scientific journals and assesses all other outputs by peer review, with full disclosure of the final outcomes to the author of the submitted output (see Bertocchi et al., 2015).

The remainder of this paper is organized as follows. Section 2 presents the institutional background. Section 3 describes how we integrated information from REF official documents with data on bibliometrics. Following a brief graphical analysis, Section 4 documents the

[3]Similar recommendations on the use of bibliometric indicators are in the Leiden Manifesto (Hicks et al., 2015) and in the San Francisco Declaration on research assessments (see http://www.ascb.org/files/SFDeclarationFINAL.pdf).

correlation between the REF results and Thomson Reuters AIS. The implications of this finding for the empirical model are presented in Section 5, where the identifying restrictions are also discussed. Section 6 presents the main results, while various sensitivity analyses and a proposed classification of journals consistent with REF outcomes appear in Section 7. Section 8 concludes the paper.

2 Background and Context

The Research Excellence Framework

The United Kingdom is regarded as a world leader in higher education, which contributes £73 billion per year to the national economy and has been linked to 20% of GDP growth between 1982 and 2005 (Universities UK, 2015). The quality of institutions has been monitored and evaluated since 1986, with the aim of informing the allocation of public funding for targeted investments in research. The latest university research audit, known as the *REF*, was a costly and comprehensive exercise conducted in 2014 and commissioned by the four UK higher education funding bodies.[4]

Special panels across disciplines assessed the productivity of 154 universities between 2008 and 2013 and reviewed a total of 190,000 research submissions by 52,000 academic staff. Published at the end of 2014, the results are used to form league tables that have important funding and reputational consequences for the institutions involved. A department with poor performance can be closed, while a top rating implies steady funding. The HEFCE distributed approximately £1.5 billion in research funds in 2016-2017, two-thirds of which reflected the quality profiles in the REF (HEFCE, 2016). Every institution wants to describe itself as a top-ranked research university, and the next evaluation round, planned for 2021, is well under way.[5]

The REF evaluation was re-designed compared to its predecessors and included research impact outside academia.[6] As a result of this change, the scientific quality of research output

[4]These are the HEFCE, the Scottish Funding Council (SFC), the Higher Education Funding Council for Wales (HEFCW) and the Department for Employment and Learning, Northern Ireland (DEL).

[5]Initial decisions on the regulation of REF 2021 are consistent with the framework of the past REF (REF, 2017).

[6]Impact was 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), and was assessed through

counted for 65% of each institution's profile, 20% was awarded for impact and an additional 15% for the research environment at the institution (e.g., infrastructure, Ph.D. students and income generated through research activities). Although REF results present the quality breakdown of institutions by category, only the classification of research outputs is considered in what follows. Performance on research is the most important determinant of the allocation of public funding and, as shown in De Fraja et al. (2016), is what matters most in the decision to hire or expand.

Guidance and criteria

Our analysis is limited to submissions in the Economics and Econometrics sub-panel, yielding a total of 28 institutions (departments) in the UK. Guidance and criteria for the evaluation process were disclosed before the submission deadline in November 2013 (REF, 2011).

Research outputs, mostly journal articles and working papers resulting from "investigation leading to new insights", must be authored by staff at the submitting institution and published between 2008 and 2013. The representativeness of research at institutions is not guaranteed by design: institutions could choose how many of their academic staff should be considered and submit at most four outputs for each scholar (public funding, however, increases in the number of academics submitted). Importantly, outputs can be considered regardless of the institution where they were generated, conditional on membership at the time of REF submission. It is well known that many institutions poached researchers from one another to improve performance using portability of outputs (De Fraja et al., 2016).

The evaluation relied entirely on in-house assessment from panellists (2,600 outputs). The panel consisted of 18 national and international academics coordinated by a chair, which ensured consistent criteria across the social sciences. Each output listed in a submission was assessed on the basis of its originality, significance and rigour. This definition allows for subjective judgement, although the Economics and Econometrics panel (and few others) used bibliometric indicators to inform assessments "when considered appropriate" (REF, 2012). Official documents state that citations affected the quality awarded in "very few cases", a fact consistent with our results below.[7] Each output was assigned to one of five

case studies.

[7]See www.ref.ac.uk for a description of rules, outcomes and the identity of panellists.

mutually exclusive tiers. Quality depends on the number of stars awarded, distinguishing between "world leading" (four-star), "internationally excellent" (three-star), "internationally recognized" (two-star) and "nationally recognized" (one-star) research. Submissions falling short of national standards, or not meeting the eligibility criteria, were flagged as "unclassified quality". The information used for accountability purposes is the share of outputs listed in a submission that were assigned to each quality level. The allocation of each output to a quality tier is not revealed, which has fuelled discussion on how to filter four-star work and influenced hiring decisions.

Research quality was found to be outstanding, with more than two-thirds of submitted publications being at least "internationally excellent". However, recognized excellence exhibits substantial variation in the proportion of four-star outputs. This can be seen from Table 1, which reports results from official documents. Columns (2) to (6) show the breakdown by quality across institutions. Using this information, a number of summary indicators of research quality were obtained and widely disseminated through media and research policy outlets. The Grade Point Average (GPA), in column (9), is calculated by multiplying the percentage of research in each group by its rating, adding them all together and dividing by 100. The Quality Index (QI) in column (10) is a weighted average reflecting the current funding allocation formula, which depends on the incidence of top-quality outputs (80% and 20% for four- and three-star research, respectively, and no contribution from the remaining outputs; see HEFCE, 2016). Additional indicators were developed to measure the "research power" of institutions – see column (11) – after adjusting the quality for the fraction of full-time equivalent faculty members submitted.[8] As we are not interested in modelling the selection of staff in submissions, in what follows, we only consider the ranking of institutions based on research quality. Appendix Figure A.1 shows that the share of staff submitted by institutions does not predict the REF score from research outputs.

Related Literature

Several studies have investigated past research assessments in the UK. Early analyses by Johnes et al. (1993) suggest that research ratings improve with an institution's size and

[8]The QI and power rating were developed by Research Fortnight. Here, we present our own calculations based on the current funding formula and on the outputs sub-profile only.

reputation, while Clerides et al. (2011) conclude that departments may benefit from having members on the RAE panel (RAE being the predecessor to the REF). In line with the latter study, De Fraja et al. (2016) suggest that institutions represented on the panel were awarded higher scores on the REF; they also show that the portability of outputs induced institutions to attract more-productive researchers by offering them higher salaries.

As in past evaluation rounds, the REF generated intense debate over its regulatory framework and incentive structure. The optimal mix between peer review and bibliometrics is often central to the discussion. Sgroi and Oswald (2013) show that outstanding research impact can be predicted using journal rankings and citations. Regibeau and Rockett (2016) argue that journal impact and citations can identify the quality of economic 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. Analysing the Australian research assessment, Tombazos and Dobra (2014) suggest that experts overstate the quality of journals where they have published or in their field of expertise.

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 large (examples are Pinski and Narin, 1976, Liebowitz and Palmer, 1984, Sauer, 1988, Laband and Piette, 1994, Kalaitzidakis et al., 2003, and Varin et al., 2016). A model-based approach for ranking scientific outlets is in Bartolucci et al. (2015). In their work journal quality is unobserved, and indicators such as AIS are used to proxy such latent factor. We also deal with unobserved quality awarded to each research output, which we infer from the total number for stars assigned to institutions by panellists. Palacios-Huerta and Volij (2004) take an axiomatic approach to the ranking of 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 research quality awarded by the REF panel. The classification model in Pitt and Yan (2017) is similar in spirit to ours, although – in addition to using different statistical assumptions – their analysis doesn't rely on output-level characteristics.

A related line of inquiry considers the publication process. There is mounting evidence that statistically significant results are disproportionally more frequent, suggesting some bias

in the judgment of what makes a good paper (Doucouliagos and Stanley 2013 and Brodeur et al. 2016). Ellison (2002) documents a slowdown in the publication process for top-journals in economics which is driven by increasingly extensive revisions. Exploiting exceptional data on submissions to top journals, Card and Della Vigna (2017) find that editors and referees set higher quality thresholds for publishing the work of established authors.

Finally, our work connects with a growing body of research on social ties in academia. For example, economists are more likely to publish in top journals when colleagues or Ph.D. supervisors are on the editorial board (Colussi, 2017). By exploiting random assignment to evaluators, Zinovyeva and Bagues (2015) show that acquainted candidates are more likely to be selected for academic promotions.

3 Data and Bibliometrics

The complete list of submissions for all institutions is available through the REF website. It consists of 2,600 outputs in Economics and Econometrics, mostly journal articles (2,388) and working papers (168). Starting from this information, we assigned to the corresponding journal all working papers flagged as forthcoming or published by August 2015.[9] Panel A of Table 2 presents the breakdown by publication type resulting from this selection. We collected citations for the journal and authors of each output to characterize research influence and prestige. The Economics and Econometrics sub-panel had access to citation counts for each publication made available from Elsevier's Scopus database in early 2014, together with contextual data on citations distribution in the output's field and year of publication. These files were deemed confidential and deleted at the end of the REF process. We therefore approximated the bibliometric indicators available to panellists with the most similar indicators obtained from the web.

We started by characterizing each journal by its AIS, which represents the average number of citations of its articles from other journals over the first five years from publication. In short, we let a journal's prestige depend on the average influence of its articles, adjusted for self-citations using the Thomson Reuters Journal Citation Reports (JCR) database. This

[9]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.

choice is motivated by past research demonstrating that the AIS predicts expert-based evaluations of research quality (Hudson, 2013).[10] We considered the AIS computed for 2013, the latest release at the time of the REF evaluation, and standardized it to have zero mean and unit variance by field to adjust for differences in citation behaviour across disciplines. Interdisciplinary research was warmly encouraged: although 94% of submissions in Economics and Econometrics appeared in economics journals, the remaining outputs span across fields such as psychology, mathematics and physics.[11] This finding suggests that economics has extramural influence on a number of other disciplines, in line with conclusions in Angrist et al. (2017) and Hudson (2017).

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, 91% of REF submissions, is shown in Figure 1. This presents a long upper tail driven by high-impact outlets (such as the top-five journals in Economics), and spikes across the whole support. Table 3 sheds more light on the origin of these spikes and lists all journals with at least 30 submissions in our sample.

The citation count for each publication is obtained using Elsevier's Scopus, the same source made available to REF panellists. We measured citations at the end of 2013, retrieving information for 2,441 outputs (94% of the sample). We additionally considered Google Scholar because of its much larger array of publishing formats, although our conclusions are robust to the source of information employed.[12] Citation counts in the analysis below are always standardized by year of publication and field. The information above is completed with the h-index (Hirsch, 2005) of all authors, which was computed from the Scopus database, and their affiliation as reported in each output. Descriptive statistics for all bibliometrics are

[10]The Italian ANVUR national assessments use the Scimago Journal Rank (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%.

[11]When a journal is assigned multiple fields (e.g., economics and statistics), the standardization is done using mean and standard deviation across all economics journals. In those rare cases in which all fields are outside economics, we considered the average of field-specific standardized scores. The classification of fields made available to REF panellists uses Elsevier's ASJC categories. To standardize the AIS, however, we use Thomson's JCR categories. These alternative classifications are, for submissions in Economics and Econometrics, substantially equivalent.

[12]The Scopus and Google Scholar archives do not cover the same population of journals and publishers. The correlation between the two citation measurements, computed from 2,441 outputs, is 86%.

presented in Panel B of Table 2.

4 Graphical Analysis

Impact factor predicts classification of research outputs

We start by characterizing the statistical property of the classification process revealed by the data. Consider, for example, outputs published in the Economic Journal (EJ). If REF submissions for this journal exceed the number of four-star outputs for at least one institution, then it must be that publications in the EJ are not deterministically awarded four stars. Similarly if the number of EJ submissions exceeds the number of outputs awarded one or two stars for at least one institution, then at least one publication in this journal must have been awarded three stars or more. We use this idea to investigate the existence of such critical cases and study their relationship with the AIS.

The likelihood of four-star outputs increases with a journal's impact factor. This can be seen in Panel A of Figure 2, which presents the share of journals at or above a given value of the AIS and that may not have four stars. This quantity is not monotone in AIS by construction, which explains the saw-tooth pattern in the figure. The data do not allow us to reject the hypothesis that outputs in journals with impact as large as that of the American Economic Review (AER) or the EJ are always awarded four stars. A critical threshold emerges around the Journal of Health Economics (JHE), suggesting an increased likelihood of borderline journals below this point. Panel B shows the share of journals at or above a given value of the AIS that may have more than two stars. Consistent with Panel A, the likelihood of a top-scoring publication increases with the journal impact factor. Similar conclusions can be obtained when considering combinations (e.g., pairs or triplets) of journals submitted to, instead of one journal at the time. This can be seen from the two bottom panels of Figure 2, which use journal pairs to compute the likelihood of submissions with fewer than four stars (Panel C) or more than two stars (Panel D).[13]

The number of journals that cannot be consistently awarded four stars increases when restrictions are imposed on the classification process. The example we consider here is the assumption that all submissions in top-five Economics journals are awarded four stars. These

[13]We document in the Appendix how these profiles were computed.

are general-interest outlets with standardized AIS values between 2.6 (AER) and 7.1 (the Quarterly Journal of Economics), which are often viewed as a "curse" because of the incentives that they generate in the profession (see the discussion in Heckman, 2017). The assumption here is that there are no "bad" AER articles that would deserve fewer than four stars. By maintaining this assumption, we find that submissions in the Journal of Econometrics (with a standardized AIS equal to 0.96) or in the Journal of Economic Theory (0.78), two of the most frequent outlets in Table 3, exceed the number of outputs awarded four stars by REF. Figure 3 offers a visual summary of this analysis and replicates Panel A of Figure 2, showing a steeper profile with respect to AIS. The message here is that top-field outputs may not have been awarded four stars despite the value assigned to these journals by the profession (see the discussion in Hudson, 2013).

The strong predictive power of the AIS emerging from the micro-data spills over to the correlation between an institution's average AIS and REF score. Figure 4 plots the AIS average using all publications submitted by an institution against its QI score (the GPA score conveys a similar message). Superimposed are predictions from a regression on linear and quadratic terms in AIS. The small departures from a deterministic trend suggest that the AIS provides a fair approximation of the classification criteria followed by the panel (see also findings in HEFCE, 2015, to corroborate this statement). Obviously an improved fit may follow by adding additional output attributes such as citation counts, an empirical question that we address in the next section.

Building on this evidence, we derive in Figure 5 the critical thresholds for awarding three or four stars if classification were based only on the AIS.[14] The critical journals resulting from this analysis are remarkably similar to those in Figure 2, again suggesting that the EJ is a borderline case for awarding four stars and the JHE lies comfortably in a three-star area. Panel A of Figure 5 shows the share of institutions that have reached the four-star threshold by values of the AIS. High-impact journals such as the AER are well above the threshold for awarding four stars. Moving to the left of the impact distribution, the EJ represents the

[14]We ranked publications by the value of the AIS and defined critical values using the classification of outputs made by panellists. For example, 20.2% of the publications submitted by Queen Mary University of London, 19 in total, were awarded four stars (see Table 1). We rank all submissions in academic journals by values of the AIS and define the critical threshold by considering the AIS of the 20th publication, the first that would be awarded three stars. The remaining thresholds are determined in a similar way. The analysis here excludes other outputs such as working papers, books and book chapters. Their classification will be addressed in Section 5 below.

cut-off for a "world leading" publication for many institutions (12 out of 28). The line in Panel B, defined by analogy for the three-star threshold, shows that the JHE is critical for scoring "internationally excellent". Sharp discontinuities in these plots reveal critical values of journal impact that are similar across institutions, again suggesting a strong gradient in the relationship between the bibliometric indicator and REF classification.

Citation data

Citations received by a publication increase with the journal impact factor, as expected, but present significant differences across outputs published in the same journal. This can be seen from column (1) of Table 4, where we report results from an output-level regression of Scopus citations on AIS, controlling for research field and publication year. Others have documented substantial variation in the number of citations for publications in the same journal (Starbuck, 2005 and Heckman, 2017 are examples). We find that a one-standard-deviation (hereafter, σ) increase in AIS is associated with a 0.164σ increase in citations, and the coefficient is highly significant. Controlling for the average h-index of the authors, which we use as additional proxy for a publication's quality, the size of the coefficient on AIS is 0.132σ and still significant – see column (3). The residual variability after netting out journal and author impact – the R^2 in column (3) is 28.5% – suggests that publication quality may have an effect on citations beyond the variables considered. Since the distribution of citations is heavily skewed, we investigate whether our conclusions are mechanically driven by the linear fit. Following Card and Della Vigna (2017), we further estimate a model for the inverse hyperbolic sine of citations in columns (2) and (4). Although the R^2 rises substantially (57%), the bibliometric indexes are still far from explaining the variability in citations. Columns (5)-(8) present estimates using Google Scholar citations, yielding a similar conclusion.

Citation counts aggregated by institution predict the institutional quality awarded by REF panellists. However, after controlling for AIS, citations are no longer significant. We find that a regression of QI on the average number of citations and average h-index of outputs submitted yields an R^2 of approximately 57%, with a statistically zero coefficient on the latter variable. The R^2 rises substantially to 89% after adding to this regression the average AIS by institution and the share of top-five economic journals submitted. In the

latter specification, only the coefficient on AIS is statistically significant, suggesting that citations are uncorrelated with REF classification once journal impact is taken into account. This expectation is borne out by Figure 6, which plots the residual QI of institutions from a regression on h-index and impact factor against residual citations after controlling for the same variables. Panel A and Panel B are obtained using Scopus and Google Scholar citations, respectively, and suggest no effect of citations on REF rankings conditional on other variables. A similar figure is obtained considering the average by institution of the authors' h-index for all outputs submitted.

5 Empirical Specifications

General formulation of the problem

Let J denote the number of journals where REF outputs where published (283, in our data). We assume here that all outputs are articles published in journals with AIS values and show in the Appendix how the equations below can be adjusted to account for other outputs such as books or book chapters. The number of submissions from institution i in journal j is X_{ij}, where $i = 1, \ldots, 28$ and $j = 1, \ldots, J$. Let $\mathbf{X}_i'$ be the $J \times 1$ vector for institution i, whose terms are the X_{ij}s. The variable D_{jk} denotes the number of stars awarded to output k in journal j, where conventionally we set $D_{jk} = 0$ for "unclassified" outputs. As the last category is non-empty in very few cases, we will omit it in what follows and consider only four tiers.[15] The quantity $\mathbf{Z}_i'$ represents the vector of attributes (e.g., citations or the h-index of authors) of all publications submitted by institution i. The elements of this vector are Z_{jk}. Finally, let:

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

be the number of publications from institution i awarded d stars by the REF panel, where $d = 1, \ldots, 4$. The measurements Y_i^d are contained in the 4×1 vector $\mathbf{Y}_i'$. Official publications combined with citation data reveal $(\boldsymbol{Y}_i, \mathbf{X}_i, \boldsymbol{Z}_i)$. However the classification of each research output is unknown, and thus, is the index $\mathbb{1}(D_{jk} = d)$.

The following exclusion restriction is maintained throughout:

[15] The number of unclassified outputs is 16, corresponding to 0.6% of submissions.

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

implying that the classification of output k depends solely on its characteristics Z_{jk} and is independent of the bundle submitted by the institution. This is consistent with the idea that quality should be assessed independently for all outputs, as explained in REF documents. The assumption implies that two publications in journal j sharing the same attributes Z_{jk} should be awarded equal quality regardless of the submitting institutions and the composition of their REF submissions. The right-hand side of equation (2) is almost non-parametric (indeed, it is fully non-parametric with one binary attribute) in that it imposes a mild restriction between output attributes and the classification probability. Of course, some of the attributes used by panellists in taking their decisions may be unobservable to us, a problem that we address in Section 7 below. Alternative specifications for the conditional probabilities in (2), for example those in Bartolucci et al. (2015), are possible but would make the estimation strategy below less straightforward.

Combined with the exclusion restriction, equation (1) implies the following:

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

for $d=1,\ldots,4$. The following constraints ensure the adding up condition for the probabilities in (2):

$$\sum_{d=1}^{4}\alpha_j^d=1,\qquad\sum_{d=1}^{4}\gamma_j^d=0. \tag{4}$$

With one attribute Z_{jk}, the relationship in (3) and the constraints (4) define a system of equations in $J\times 6$ unknowns, implying that sample size is too small to allow for estimation. If the true model is sparse, for example because only a few journals matter for determining the total number of four-star outputs, estimation using lasso and related methods may offer a solution (Hastie et al., 2015). The dimensionality problem can be addressed at the cost of imposing parametric assumptions, for example as in the MCMC algorithm by Pitt and Yan (2017). We take a different approach here, which reduces the dimensionality by imposing restrictions guided by the graphical analysis in the previous section and has the advantage of a much simpler estimation strategy.

Parameterization adopted

We group journals into three mutually exclusive tiers depending on AIS. Moving from Figure 2, we use the EJ as the lower limit for a top tier (Tier 1 in what follows) that will include four-star outputs with high probability. A middle tier is then defined, Tier 2 in what follows, spanning over a grey area that comprises a mix of four- and three-star publications. Finally, a bottom tier is defined as the complement to all journals included above. To fix ideas and building on the graphical evidence in Section 4, we use the JHE as the threshold to define the latter two tiers. We will address the problem of selecting alternative thresholds for the definition of tiers in the empirical section. As we shall see, however, our conclusions are robust to a number of alternative choices.

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 journal characteristics (such as AIS) denoted by W_j. In addition, we impose constant effects of publication attributes within a tier. These two assumptions imply the following restrictions:

$$\alpha_j^d = \alpha_{0\tau}^d + \alpha_{1\tau}^d W_j, \qquad \gamma_j^d = \gamma_\tau^d, \tag{5}$$

for all js in tier τ.[16] By substituting into (3), we have:

$$\begin{aligned} E\left[Y_i^d | \mathbf{X}_i, \mathbf{Z}_i\right] &= \sum_\tau \alpha_{0\tau}^d \left(\sum_{j\in\tau} X_{ij}\right) + \sum_\tau \alpha_{1\tau}^d \left(\sum_{j\in\tau} W_j X_{ij}\right) \\ &+ \sum_\tau \gamma_\tau^d \left(\sum_{j\in\tau} \sum_{k=1}^{X_{ij}} Z_{jk}\right), \end{aligned} \tag{6}$$

for $d = 0, \ldots, 4$. In the parameterization adopted, the constraints in (4) are implied by:

$$\sum_d \alpha_{0\tau}^d = 1, \qquad \sum_d \alpha_{1\tau}^d = 0, \qquad \sum_d \gamma_\tau^d = 0. \tag{7}$$

[16] For example, the probability that publication k in journal j is awarded four stars:

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

will depend on a journal's attributes W_j and characteristics Z_{jk} (such as citations or the h-index of authors).

Restrictions on the classification probabilities

Two additional sets of restrictions are imposed to reduce the dimensionality of the problem. First, we assume that outputs in Tier 1 and Tier 2 are always awarded at least three stars. We also impose that outputs in Tier 3 can never be awarded four stars. More formally, the constraints:

$$E\left[\mathbb{1}(D_{jk} \leq 2)|\mathbf{X}_i, \mathbf{Z}_i\right] = 0, \quad j \in \{Tier\,1,\, Tier\,2\}\,,$$

$$E\left[\mathbb{1}(D_{jk} = 4)|\mathbf{X}_i, \mathbf{Z}_i\right] = 0, \quad j \in \{Tier\,3\}$$

are imposed in addition to the adding-up restrictions (7). The system of equations derived from (6) is reported in the Appendix and is estimated using seemingly unrelated regressions when imposing all constraints.

The restrictions above are consistent with the graphical analysis in Section 4 and allow for errors in the classification of outputs of at most one star (this is an assumption frequently made in empirical work on misclassification; see, for example, Battistin and Sianesi, 2011 and references therein). These restrictions imply that a "bad" publication in the AER is never worth fewer than three stars. They also allow for 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.

More generally, these restrictions on the classification probabilities mirror results from past research on the informational content of international lists used for journal rankings. Comparing a number of bibliometric indicators with the views of experts, Hudson (2013) concludes that some journals can be unambiguously clustered with respect to the number of stars awarded. This classification is fuzzy ("probable" and "possible" is his narrative) in other cases. Appendix Table A.1 shows that our definition of Tier 1 and Tier 2 coincides with that of unambiguously three-star or higher outputs in Hudson (2013) and that the bulk of outputs in Tier 3 is expected to have a two-star classification. Similar conclusions emerge by considering other research on the REF (Hole, 2017).

Classification of other research outputs

A choice has to be made on the classification of outputs for which an AIS is not available. These represent approximately 8.6% of the population considered, as shown in Table 2. In our baseline specification books from international editors (e.g., Princeton University Press), the most frequent example, are assigned to Tier 1. Other books, including edited collections of chapters, are assigned to Tier 2. For chapters in edited books we follow a similar rule, downgrading by one tier the output classification: book chapters from international editors are assigned to Tier 2, and to Tier 3 otherwise. Classification in tiers of books and book chapters is shown in columns (1) to (3) of Appendix Table A.3. As a sensitivity check, we also consider a classification that assigns all book chapters to Tier 3, and books from international editors to Tier 2. This classification is described in columns (4) to (6) of the same table. Finally, the few remaining outputs (working papers not published by August 2015, datasets and reports) are assigned to Tier 3. Our estimation results are robust to these classification criteria.

6 Results

Baseline specifications

We start from a baseline specification that controls for AIS only through stratification on tiers. In other words, we estimate (6) while imposing $\gamma_\tau^d = 0$ and $\alpha_{1\tau}^d = 0$ for all τs and for all ds. We find a strong polarization of outputs in Tier 1 and Tier 2 with respect to the probability of scoring four and three stars, respectively (90% or above). Tier 3 is instead more heterogeneous in quality, with outputs almost equally split between two and three stars. This can be seen from columns (1) to (3) in Panel A of Table 5 where, following the graphical analysis in Section 4, Tier 2 is defined as the AIS interval from JHE to EJ. Here and below, outputs not in academic journals are assigned to tiers as explained in the last section.

Columns (4) to (6) show estimates when imposing $\gamma_\tau^d = 0$ for all τ's but controlling for a quadratic polynomial in AIS ($\alpha_{1\tau}^d \neq 0$). The estimation here adjusts for within-tier heterogeneity in the AIS, and the values shown are the average probabilities in a tier implied

by our model.[17] The conclusions after adjusting for AIS are similar to those from the baseline model reported in columns (1) to (3). The relationship between estimated probabilities and the AIS is presented in Appendix Figure A.2, which shows more pronounced within-tier heterogeneity at the bottom end of the impact distribution. In line with the graphical analysis in Section 4, our model predicts that outputs in journals with an AIS as large as that of the EJ are almost deterministically awarded four stars.

Robustness to the definition of tiers

Our main definition of tier cut-offs is data driven but somewhat arbitrary. For example, a number of journals with AIS values similar to that of the EJ would belong to a lower tier only because of this small difference. We address this concern here by exploring alternative definitions of Tier 2. The results in Panel A of Table 5 suggest that outputs to the left of JHE have a nearly equal chance of being awarded three or two stars. There is also fuzziness in the classification of Tier 1 and Tier 2 outputs, with a 6.4% probability of scoring four stars in Tier 2 and 11% probability of scoring three stars in Tier 1 – see columns (1) and (2).

Panel B of Table 5 presents estimates from an alternative definition of tiers that is obtained by minimizing the probability of four-star outputs in Tier 2. Combined with the restrictions on the classification probabilities discussed in Section 5, this alternative definition is expected to maximize the share of four-star outputs in Tier 1 and of one- and two-star outputs in Tier 3. We determine the optimal width of Tier 2 by means of a grid search over 60×60 possible choices obtained by varying the upper and lower limits. We start by setting the lower limit on Tier 2 at JHE. We then select 60 journals with AIS in a window centred on the EJ and use them to define alternative upper limits on Tier 2. This defines a range of 60 possible intervals between the JHE and the new upper limit, which we use iteratively to estimate our model. We then select the definition of Tier 2 that yields estimates at the minimum distance from the following constraints:

$$p_2(3) = 1, \qquad p_3(3) = 0,$$

where $p_\tau(d)$ is the probability of a d-starred publication in tier τ. In words, Tier 2 is defined

[17] In addition, we impose continuity of the classification probabilities for journals at the boundaries between tiers. For example, a publication in the EJ (which marks the lower end of Tier 1) has a probability of being awarded four stars equal to that of a publication in the journal with the highest AIS in Tier 2.

to maximize between-tier distance in classification probabilities while ensuring within-tier homogeneity. This procedure is replicated by replacing JHE with 60 journals falling in a window around its AIS. Appendix Figure A.3 shows that the distance from the constraints is minimized when the upper limit on Tier 2 is the Journal of Econometrics, below the EJ, and the lower limit on Tier 2 is the Journal of Economic Dynamics and Control, below the JHE.[18] We therefore present results when Tier 2 spans the interval from the Journal of Economic Dynamics and Control to the Journal of Econometrics.

Consistent with expectations, columns (1) to (3) in Panel B of Table 5 show that outputs in Tier 3 have little chance of being awarded three stars. The estimated probabilities in column (1) across the two panels also suggest that publications in journals to the left of the EJ may not be consistently awarded four stars, as the probability of top-scoring outputs in Panel B drops to 80%. Columns (4) to (6) of the table show substantially similar conclusions after controlling for AIS. The profile of classification probabilities with respect to AIS, not reported here, conveys the same message as Appendix Figure A.2.

The sensitivity of our conclusions to the choice of Tier 2 is further explored with the aid of a graphical analysis. The definition of Tier 2 here is obtained by varying the upper limit while leaving the lower limit at the JHE. Presented in Figure 7 are the estimated probabilities of four-star outputs in Tier 1 and Tier 2. The value zero on the horizontal axis corresponds to the definition of Tier 1 in Panel A of Table 5, and the sensitivity of results to deviations below (negative values) and above the EJ (positive values) is investigated. The value -5 corresponds to the definition of Tier 1 in Panel B of Table 5 (from the Journal of Econometrics). We find that the probability of four-star outputs in Tier 1 quickly grows to one by moving to the right of the EJ. This probability also grows in Tier 2, which implies that a positive density of top-scoring outputs exists at or above the EJ.

A closer examination of REF submissions in the light of the above results reveals that publications in top generalist outlets are unambiguously regarded as "world leading" (four-

[18] Calculations available on request show that the Journal of Econometrics is always the optimal upper limit for all lower limits in our grid. Appendix Figure A.3 reports values of the following quantity:

$$\left\{[p_2(3)-1]^2 + Var\left[p_2(3)\right]\right\} + \left\{[p_3(3)]^2 + Var\left[p_3(3)\right]\right\},$$

where the $p_\tau(d)$s are estimated from (6) when imposing $\gamma_\tau^d = 0$ and $\alpha_{1\tau}^d = 0$ for all τs and for all ds. Panel A shows the average value of this quantity over the 60 possible choices for the lower limit on Tier 2. Panel B reports the value of this quantity when the upper limit on Tier 2 is the Journal of Econometrics.

star), and that the classification of specialized outlets is less uncontroversial. Estimates of the classification probabilities are strikingly robust to tier definitions, which shows that the strong gradient between journal impact and REF classification is not driven by our choice of cut-offs. We conclude that, while the probability of four-star outputs is high for publications in journals to the right of the EJ, there exists a grey area where the classification is more ambiguous. This area in our data includes a number of top-field journals, notably the Journal of Econometrics and the Journal of Economic Theory.

The effect of publication characteristics

We consider specifications of (6) that adjust for the effect of output's citations, the average h-index of authors and field when $\gamma_\tau^d \neq 0$. The estimation results are shown in Appendix Table A.2 using the optimized definition of Tier 2 and controlling for within-tier heterogeneity in AIS. The results are remarkably similar to the baseline estimates in Table 5.

We find that output characteristics do not affect the probability of being awarded four stars, conditional on tier membership. The fact that citations are marginally associated with the assessments of panellists is consistent with official documents (for example, see REF, 2015, p.51). The contribution of citations and h-index is discussed here with the aid of simple graphs reporting how a change from the tenth percentile in tier to the ninetieth percentile of these variables affects the probability of scoring four stars. Figure 8 shows that the contribution to the probability of scoring four stars is not statistically significant in Tier 1 and Tier 2. The effects are larger in Tier 3, which is arguably the most heterogeneous group in terms of quality, but still imprecisely estimated.[19]

[19] All specifications include controls for books, book chapters and other outputs as explained in the Appendix. Estimates not reported here suggest that books published by international editors (see Appendix Table A.3) were most likely awarded four stars, with book chapters most likely receiving three stars. Our analysis does not reveal any clear pattern for the field coefficients, perhaps reflecting the fact that 94% of submissions were in economics journals.

7 Specification Tests and Journal Ranking

Results are not driven by unobserved quality

The REF panel assessed traits of research quality, such as significance and rigour, which are unobservable to us. We adjust our estimates for such unobservables maintaining the assumption that, within the same journal, a publication with many citations cannot be of lower quality than a publication with few citations. This implies that unobservable quality is non-decreasing in citations. The analysis below shows that our conclusions are robust to alternative specifications for the relationship between unobservable quality and our indicators of citations included in Z_{jk}. It follows that what we have learnt on the classification process is not driven by unobservable quality of research outputs.

Consider the following version of equation (2):

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

where $\mathbf{U}_i'$ represents the vector of attributes, unobservable to us, of all publications submitted by institution i. The elements of this vector are U_{jk}. The equation implies that the classification of output k depends on attributes we can observe (e.g., citations) Z_{jk} and a latent indicator of quality U_{jk} assessed by panellists. An exclusion restriction is maintained, stating that each output is classified independently of the bundle submitted by the institution. We are interested in understanding to what extent our conclusions are affected by the omitted variable U_{jk}.

If the latter variable is independent of other outputs conditional on Z_{jk}, one can write:

$$E\left[\mathbb{1}(D_{jk}=d)|\mathbf{X}_i,\mathbf{Z}_i\right]=\alpha_j^d+\gamma_j^d Z_{jk}+\delta_j^d E\left[U_{jk}|Z_{jk}\right], \tag{8}$$

which clarifies the effects of omitted variables in equation (2). The important message from equation (8) is that estimates in Table A.2 are robust to output's unobserved quality if the latter increases linearly with citations. More generally, the unobservable component U_{jk} will introduce non-linearities into the relationship between Z_{jk} and the classification probabilities. To see this, consider the following parametric relationship between unobserved quality and citations:

$$E\left[U_{jk}|Z_{jk}\right]=\rho_{0j}+\rho_{1j}Z_{jk}+\rho_{2j}Z_{jk}^2. \tag{9}$$

When this function is linear, substituting into (8) and re-arranging terms yields new intercept and slope parameters that are a combination of α_j^d, γ_j^d, $\delta_j^d\rho_{0j}$ and $\delta_j^d\rho_{1j}$. As we are not interested in disentangling the values of these parameters, the case of unobservables linear in Z_{jk} is already embedded in the specifications considered above. It also follows that (8) will contain a quadratic term in Z_{jk} when $\rho_{2j} \neq 0$.

Appendix Table A.4 shows that our baseline estimates of the classification probabilities are unchanged after adding a quadratic term in citations to the estimating equations. A test for the joint significance of the linear and quadratic terms in citations does not allow us to reject the null for Tier 2 and Tier 3 outputs. Instead, the inclusion of the quadratic term makes citations marginally significant in Tier 1, although the conclusions drawn for outputs in this tier are unaffected. We find that a move from an output with an average number of citations in Tier 1 to an output in the ninetieth percentile of the citation distribution increases the probability of scoring four stars from approximately 70% — see column (1) of Table A.4 — to 90%.

Model predicts institutions' performance

We compute differences between the number of outputs awarded d stars and the same number predicted by our model. These differences must be uncorrelated, by construction, with all bibliometric indicators included in the analysis. Here, we study the correlation with other measures of research excellence at an institution to corroborate the validity of the assumptions underlying our analysis. We predict for the 28 institutions the number outputs awarded d stars, $\hat{Y}_i^d$, and subtract it from the number observed, Y_i^d, for $d = 1, \ldots, 4$. We then consider the following equation defined from 28×4 observations:

$$Y_i^d - \hat{Y_i^d} = \mu_d + \xi S_i + \eta_i,$$

where S_i is a vector of institution-level characteristics. All specifications include dummies for the number of stars awarded, the indicator of research quality available from the previous national evaluation round (RAE 2008), and the REF scores of non-academic impact and research environment. The value $\hat{Y}_i^d$ is computed by considering the model specification in the previous section – see columns (4) to (6) of Appendix Table A.4. Standard errors are clustered on institutions, and the results are reported in Appendix Table A.5.

We also consider the correlation with indicators of academic ties between REF panellists and institutions. Past research has demonstrated that the professional network plays an important role in academic publications and promotions (see Zinovyeva and Bagues, 2015 and Colussi, 2017 for examples). We therefore collected data on the current and past affiliations of REF panellists and their co-authors and computed two proxies for the connections between panellists and the institution. Column (2) of Appendix Table A.5 controls for a dummy equal to one if at least one panellist was ever employed at the institution, interacting this variable with dummies for the number of stars awarded. Column (3) considers the share of panellists or panellists' co-authors ever employed at the institution (see the Appendix for details on the construction of these variables).[20]

The predictions of our models do not depart systematically from the actual number of outputs at any level of awarded quality. The differences between observed and predicted REF scores are uncorrelated with the indicators of research performance and institutional ties with the panel. Coefficients in the table are small and fairly precise zeros, with only one coefficient in column (3) being marginally significant.

Journal ranking

Given the proliferation of rankings and their role in personnel decisions, we report in Table 3 the predicted probabilities of scoring three or four stars for the journals with the most submissions to the REF. To ensure comparability with previous studies (Hudson, 2013 and Hole, 2017), journals are grouped depending on values of predicted probabilities. Unambiguously four-star journals are those with a probability of having a "world leading" classification at least equal to 65%. 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. The classification at lower values of the AIS becomes less clear-cut, and the EJ and the Journal of Econometrics are

[20]Pitt and Yan (2017) find that outputs published in the same journal contribute to the final count of four-star publications depending on the submitting institution. The variables on networks here are used to model this channel, although the findings in Pitt and Yan (2017) may be driven by output attributes unobservable to them that are controlled for in our analysis.

examples of possible four-star journals. Top-field publications in the Journal of Economic Theory and the Journal of Public Economics are most likely awarded three stars. The classification algorithm in Pitt and Yan (2017) yields a very similar ranking of academic outlets, despite the different statistical conditions imposed.

8 Summary and Directions for Further Work

The strong correlation between subjective assessments of research quality and bibliometric indicators is often used to argue against peer reviews. In many countries, disciplinary panels of peers are provided with information on bibliometric indicators to inform their assessments. The REF 2014 Economics and Econometrics sub-panel, which we considered here, is one notable example. The practical question then arises of whether the ranking of journals by informed experts mirrors objective indicators of journal impact that could be attained at much lower costs and on a more frequent basis.

We used the REF as a natural experiment to re-consider this issue, in light of the strong resistance to bibliometric assessments from the academic community in the months preceding the national evaluation (Wilsdon et al., 2015). This exercise is not straightforward because the classification of outputs made by panellists is not disclosed. Official documents report, for each institution, only the share of outputs submitted by the number of stars awarded. This has fuelled discussions in the national academic community on the determinants of top-scoring (four-star) output beyond those indicators of impact and citations that are widely available.

Our analysis shows that a classification of outputs based on Thomson Reuters AIS approximates fairly well the final ranking by the experts. The AIS is a well-known predictor of experts' valuation of research quality (Hudson, 2013), it ranks journals following the same formula in a Google search (Brin and Page, 1998), and shares similarities with the optimal algorithm from an axiomatic approach to determine intellectual influence (Palacios-Huerta and Volij, 2004). The role of AIS for REF outputs is revealed non-parametrically by the data, and investigated through a statistical model of classification. Under the assumption that rigour and originality assessed through peer reviews are non-decreasing in a publication's number of citations, our conclusions cannot be the mechanical consequence of omitted un-

observables related to quality of the outputs submitted. The correlation between subjective assessments of research quality and bibliometric indicators does not come as surprise, and was also shown by the *Higher Education Funding Council for England* using anonymous data (HEFCE, 2015). A similar relationship has been documented in past research on large-scale assessments for the UK (Clerides et al., 2011, Taylor, 2011, and Hudson, 2013) and other countries (see, for example, Bertocchi et al., 2015). This result is often used to advocate metric-based evaluations as opposed to peer reviews to reduce administrative burden and the risk of bias (see, for example, Laband, 2013).

Our estimates also show that outputs in academic journals with a large impact factor are always awarded four stars, independent of other output attributes such as citations. Well-regarded generalist journals in Economics, including all top-five outlets, are above the threshold and unambiguously considered "world leading", in line with previous research (Hudson, 2013). However, our results imply that the classification of a number of specialized journals often regarded as top-field is more ambiguous. A direct policy implication from our analysis is that costly and time-consuming peer reviews such as REF should focus on those research outputs for which the assessment of quality is more controversial. Bibliometrics could be used for accountability purposes and continuous monitoring between large-scale assessment exercises.

References

Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2017). Inside job or deep impact? Using extramural citations to assess economic scholarship. *NBER working paper No. 23698*.

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)*, 178:1025–1049.

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):451–466.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN System*, 30(1-7):107–117.

Brodeur, A., Le, M., and Zylberberg, M. S. Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Card, D. and Della Vigna, S. (2017). What do editors maximize? Evidence from four Economics journals. *NBER working paper No 23282*.

Clarivate Analytics (2014). 2013 Journal Citations Report. Technical report.

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*, 78:565–583.

Colussi, T. (2017). Social ties in academia: a friend is a treasure. *Review of Economics and Statistics, forthcoming*.

De Fraja, G., Facchini, G., and Gathergood, J. (2016). How much is that star in the window? Professorial salaries and research performance in UK universities. *NICEP Working Paper: 2016-08*.

Doucouliagos, C. and Stanley, T. D. (2013). Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys*, 27(2):316–339.

Ellison, G. (2002). The slowdown of the Economics publishing process. *Journal of Political Economy*, 110(5):947–993.

European Association for Quality Assurance in Higher Education (2005). Standards and guidelines for quality assurance in the European Higher Education Area.

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*.

Hastie, T., Wainwright, M., and Tibshirani, R. (2015). *Statistical learning with sparsity*. CRC Press.

Heckman, J. J. (2017). Publishing and promotion in Economics: The curse of the Top Five. *2017 AEA Annual Meeting*.

HEFCE (2015). The metric tide: Correlation analysis of REF 2014 scores and metrics (supplementary report ii to the independent review of the role of metrics in research).

HEFCE (2016). Guide to funding 2016-17.

HESA (2017). Higher education statistics for the United Kingdom 2015/16 - finance data.

Hicks, D., Wouters, P., Waltman, L., de Rijke, S., and Rafols, I. (2015). The leiden manifesto for research metrics. *Nature*, 520:429–431.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Hole, A. R. (2017). Ranking Economics journals using data from a national research evaluation exercise. *Oxford Bulletin of Economics and Statistics*, 79(5):621–636.

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

Hudson, J. (2017). Identifying Economics place amongst academic disciplines: a science or a social science? *Scientometrics*, 113:735–750.

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)*, 156:271–286.

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):1346–1366.

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

Laband, D. N. and Piette, M. J. (1994). The relative impacts of Economics journals: 1970-1990. *Journal of Economic Literature*, 32(2):640–666.

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

Palacios-Huerta, I. and Volij, O. (2004). The measurement of intellectual influence. *Econometrica*, 72(3):963–977.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: theory, with application to the literature of Physics. *Information processing and management*, 12:297–312.

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.

REF (2011). Assessment framework and guidance on submissions.

REF (2012). Panel criteria and working methods.

REF (2015). Research Excellence Framework 2014: Overview report by main panel C and sub-panels 16 to 26.

REF (2017). Initial decisions on the Research Excellence Framework 2021.

Regibeau, P. and Rockett, K. E. (2016). Research assessment and recognized excellence: Simple bibliometrics for more efficient academic research evaluations. *Economic Policy*, 31(88):611–652.

Sauer, R. D. (1988). Estimates of the returns to quality and coauthorship in economic academia. *Journal of Political Economy*, 96(4):855–866.

Sgroi, D. and Oswald, A. J. (2013). How should peer-review panel behave? *Economic Journal*, 123:F255–F278.

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

Stern, N. (2015). Building on success and learning from experience, an independent review of the Research Excellence Framework.

Taylor, J. (2011). The assessment of research quality in UK universities: Peer review or metrics? *British Journal of Management*, 22:202–217.

Tombazos, C. G. and Dobra, M. (2014). Formulating research policy on expert advice. *European Economic Review*, 72:166–181.

Universities UK (2015). The economic role of UK universities.

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.

Zinovyeva, N. and Bagues, M. (2015). The role of connections in academic promotions. *American Economic Journal: Applied Economics*, 7(2):264–292.

Table 1: Research Excellence Framework (REF) outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Institution	**Outputs**	**% 4***	**% 3***	**% 2***	**% 1***	**% n.c.**	**FTE staff**	**% FTE submitted**	**GPA**	**QI**	**PR**
University College London	142	69.7	28.2	2.1	0	0	36.9	82.0	3.68	76.75	62.92
London School of Economics and Political Science	183	56.3	33.3	4.9	0.6	4.9	51.4	91.8	3.36	64.63	73.80
University of Cambridge	99	54.5	39.4	5.1	1	0	27	71.1	3.47	64.35	38.60
University of Warwick	136	42.6	50.8	6.6	0	0	41.6	80.0	3.36	55.30	51.11
University of Oxford	242	42.6	44.2	11.1	2.1	0	83.9	86.5	3.27	53.65	100.00
Royal Holloway, University of London	51	35.3	45.1	15.7	1.9	2	14.4	62.6	3.10	46.58	14.90
University of Edinburgh	55	30.9	54.6	12.7	1.8	0	17.5	62.5	3.15	44.55	17.32
University of Essex	113	29.2	60.2	10.6	0	0	33.33	83.3	3.19	44.25	32.77
University of Surrey	71	26.8	52.1	21.1	0	0	20.65	82.6	3.06	39.83	18.27
University of East Anglia	49	20.4	71.4	8.2	0	0	14	63.6	3.12	38.25	11.90
University of St Andrews	51	23.5	58.9	17.6	0	0	20.5	66.1	3.06	38.23	17.41
University of Bristol	63	22.2	58.8	19	0	0	18.6	74.4	3.03	36.90	15.25
University of Nottingham	127	19.7	65.3	14.2	0	0.8	35	76.1	3.03	36.03	28.01
Queen Mary University of London	94	20.2	62.8	15.9	1.1	0	24.45	78.9	3.02	35.90	19.50
University of Glasgow	83	18.1	61.4	18.1	2.4	0	23.75	79.2	2.95	33.45	17.65
University of Southampton	82	22	37.8	35.3	3.7	1.2	21.8	77.9	2.76	31.45	15.23
University of Leicester	80	18.8	50	28.7	0	2.5	22.4	77.2	2.83	31.30	15.58
University of York	104	14.4	59.6	24.1	1.9	0	28.07	61.0	2.87	29.30	18.27
University of Exeter	83	13.3	57.8	19.3	9.6	0	24.5	79.0	2.75	27.75	15.10
University of Sussex	54	14.8	46.3	35.2	1.8	1.9	17.4	72.5	2.70	26.38	10.20
City University London	54	16.7	37	29.6	16.7	0	13.7	52.7	2.54	25.95	7.90
University of Manchester	114	11.4	53.5	30.7	4.4	0	33.2	73.8	2.72	24.78	18.27
University of Birmingham	79	7.6	58.2	32.9	1.3	0	24.2	89.6	2.72	22.15	11.91
Birkbeck College	97	10.3	47.4	37.1	4.2	1	25.15	78.6	2.62	22.15	12.38
University of Sheffield	50	8	56	36	0	0	14.9	57.3	2.72	22.00	7.28
University of Aberdeen	63	4.8	50.8	30.1	14.3	0	19.25	80.2	2.46	17.50	7.48
University of Kent	79	2.5	43.1	37.9	16.5	0	21.9	84.2	2.32	13.28	6.46
Brunel University London	102	2	22.5	63.7	11.8	0	26.2	90.3	2.15	7.63	4.44

Note. The table reports selected results from official REF publications for the Economics and Econometrics sub-panel (see http://www.ref.ac.uk/). Column (1) reports the number of research outputs submitted. Columns (2) to (6) show the quality breakdown of submissions for the 28 institutions involved (by the number of stars awarded and outputs not classified). Columns (7) and (8) present number and share of full-time equivalent (FTE) staff members submitted, respectively (source: Higher Education Statistic Agency). Columns (9) to (11) show the Grade Point Average (GPA), the Quality Index (QI) and the Power Rating (PR), respectively (see Section 2 for definitions).

Table 2: Descriptive statistics for research outputs

	(1) Mean	(2) Std. Dev.
Panel A. Characteristics of Outputs Submitted		
Publication Type:		
Journal	0.9185	0.2737
Book Chapter or Proceedings	0.0046	0.0678
Book	0.0115	0.1068
Other	0.0654	0.2473
Field:		
Economics	0.9378	0.2415
Statistics	0.0941	0.2920
Finance	0.0787	0.2694
Mathematics	0.0472	0.2122
Missing	0.0719	0.2584
Authors:		
Number	2.2794	0.9319
H-index (at submission)	9.1165	5.9447
Panel B. Bibliometrics (December 2013)		
From Thomson Reuter's Journal Citation Reports:		
Article Influence Score *	3.0800	2.9037
Missing	0.0858	0.2801
Citations:		
From Elsevier's Scopus *	6.5834	14.9900
From Google Scholar *	29.1400	63.0600
Not in Elsevier's Scopus	0.0612	0.2397
Not in Google Scholar	0.0012	0.0340
Total number of submissions	2,600	

Note. The table presents descriptive statistics for all submissions in Economics and Econometrics. Panel A shows the breakdown by type, field and number of authors. Panel B reports the bibliometric indicators considered for the analysis: Article Influence Score, citation counts from Elsevier's Scopus, citation counts from Google Scholar and the h-index of authors. See Section 3 for details and definitions. *: conditional on non-missing data.

Table 3: Academic journals most frequently submitted

Journal	(1) Frequency	(2) AIS	(3) Estimated Probability: 4 Stars	(4) Estimated Probability: 3 Stars
4*				
Quartely Journal of Economics	30	7.05	1.000	0.000
Econometrica	70	4.48	1.000	0.000
Review of Economic Studies	63	3.44	0.891	0.109
American Economic Review	115	2.64	0.762	0.238
Review of Economics and Statistics	59	2.16	0.669	0.331
Probable 4*				
Journal of the European Economic Association	73	1.48	0.512	0.488
Possible 4*				
Journal of Monetary Economics	42	1.14	0.423	0.577
Economic Journal	106	1.11	0.415	0.585
Journal of Econometrics	95	0.96	0.378	0.622
3*				
Journal of International Economics	37	0.90	0.326	0.674
Journal of Economic Theory	84	0.78	0.234	0.766
Journal of Public Economics	57	0.77	0.226	0.774
International Economic Review	30	0.76	0.221	0.779
Journal of Development Economics	50	0.66	0.155	0.845
Econometric Theory	35	0.59	0.115	0.885
Journal of Health Economics	33	0.40	0.032	0.968
Games and Economic Behaviour	83	0.32	0.009	0.991
European Economic Review	52	0.22	0.000	1.000
Journal of Money Credit and Banking	34	0.14	0.000	1.000
Journal of Economic Behavior & Organization	42	-0.07	0.000	1.000
Economic Theory	49	-0.07	0.000	1.000
Journal of Economic Dynamics and Control	45	-0.13	0.000	1.000
2*				
Economics Letters	63	-0.37	0.000	0.268

Note. The table lists, in columns (1) and (2), journals with at least 30 submissions in Economics and Econometrics, together with their standardized Article Influence Score (AIS). Journal names are sorted by the estimated probability of scoring four stars, reported in column (3) and the presumed number of stars using the classification in Hudson (2013), obtained as described in Section 6. 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 6 for details.

Table 4: Relationship between citation count and Article Influence Score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Elsevier`s Scopus Citations				Google Scholar Citations			
	Citations	Asinh(citations)	Citations	Asinh(citations)	Citations	Asinh(citations)	Citations	Asinh(citations)
AIS	0.164***	0.269***	0.132***	0.227***	0.207***	0.434***	0.176***	0.386***
	(0.030)	(0.047)	(0.026)	(0.045)	(0.038)	(0.076)	(0.035)	(0.075)
H-index			0.206***	0.271***			0.201***	0.308***
			(0.044)	(0.018)			(0.044)	(0.025)
Constant	0.755	3.001***	0.789	3.045***	0.045	3.932***	0.081	3.987***
	(0.735)	(0.592)	(0.806)	(0.679)	(0.290)	(0.405)	(0.358)	(0.501)
Field Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,349	2,349	2,349	2,349	2,377	2,377	2,376	2,376
R-squared	0.249	0.536	0.285	0.570	0.174	0.435	0.207	0.468
Method	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS

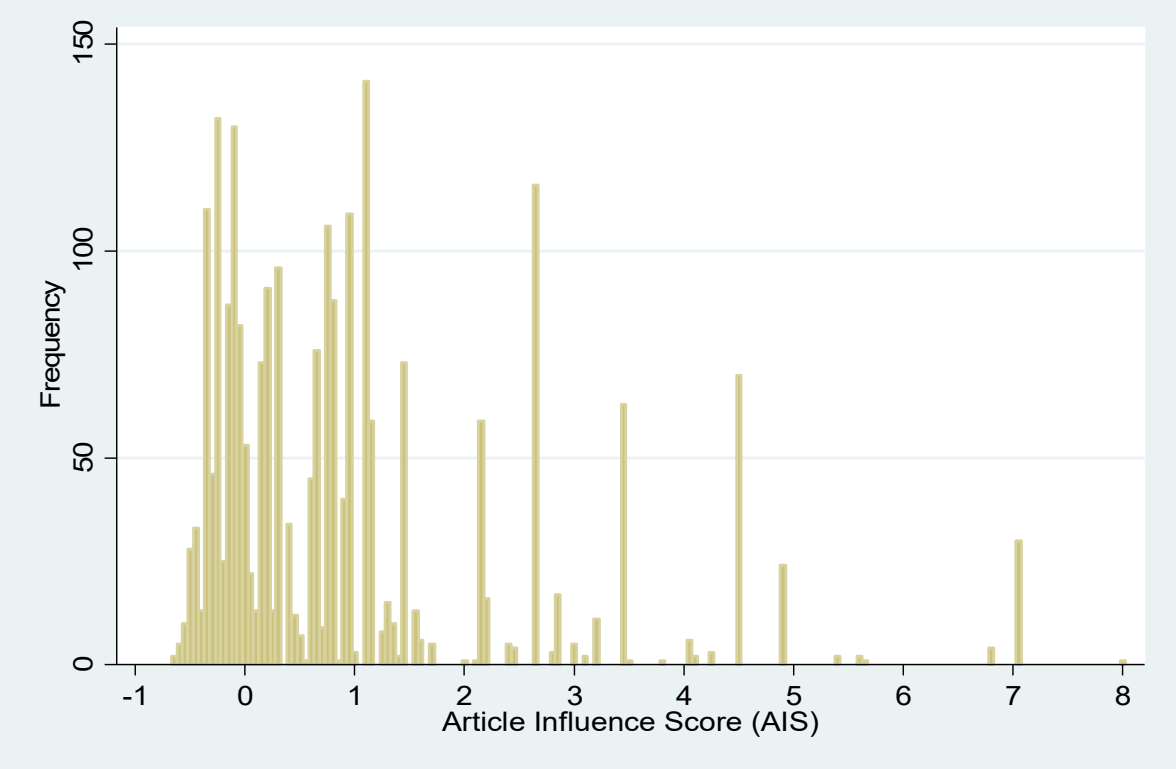
Note. The table shows regressions from publication-level data that control for field and year of publication. All equations consider citations as the left-hand-side variable, using Elsevier's Scopus in columns (1)-(4) and Google Scholar in columns (5)-(8). In columns (1), (3), (5) and (7), citations are standardized to have zero mean and unit variance in the sample. In columns (2), (4), (6) and (8), the dependent variable is the inverse hyperbolic sine of citations, as in Card and Della Vigna (2017). Columns (1), (2), (5) and (6) consider specifications that control for Article Influence Score, standardized by field. Columns (3), (4), (7) and (8) add the average h-index of authors at the time of REF submission, standardized to have zero mean and unit variance in the sample. Standard errors are clustered on the journal. See Section 4 for details. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Estimation results from the baseline model

	(1)	(2)	(3)	(4)	(5)	(6)
	without adjustment			with adjustment (AIS)		
	Tier			Tier		
	1	2	3	1	2	3
	Panel A. Tier 2 from JHE to EJ					
4* ("world leading")	0.890***	0.064		0.801***	0.161***	
	(0.043)	(0.082)		(0.037)	(0.06)	
3* ("internationally excellent")	0.110**	0.936***	0.491***	0.199***	0.839***	0.507***
	(0.043)	(0.082)	(0.033)	(0.037)	(0.06)	(0.022)
2* ("internationally recognised")			0.436***			0.422***
			(0.027)			(0.018)
1* ("nationally recognised")			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	1,265	784	551	1,265
Mean of standardized AIS in tier	2.57	0.76	-0.09	2.57	0.76	-0.09
	Panel B. Optimized Tier 2					
4* ("world leading")	0.799***	0.037		0.714***	0.086***	
	(0.029)	(0.035)		(0.024)	(0.024)	
3* ("internationally excellent")	0.201***	0.963***	0.044	0.286***	0.914***	0.098***
	(0.029)	(0.035)	(0.041)	(0.024)	(0.024)	(0.034)
2* ("internationally recognised")			0.811***			0.770***
			(0.037)			(0.029)
1* ("nationally recognised")			0.145***			0.132***
			(0.019)			(0.021)
Number of journals in tier	53	88	143	53	88	143
Number of publications in tier	888	1,067	645	888	1,067	645
Mean of standardized AIS in tier	2.37	0.33	-0.31	2.37	0.33	-0.31

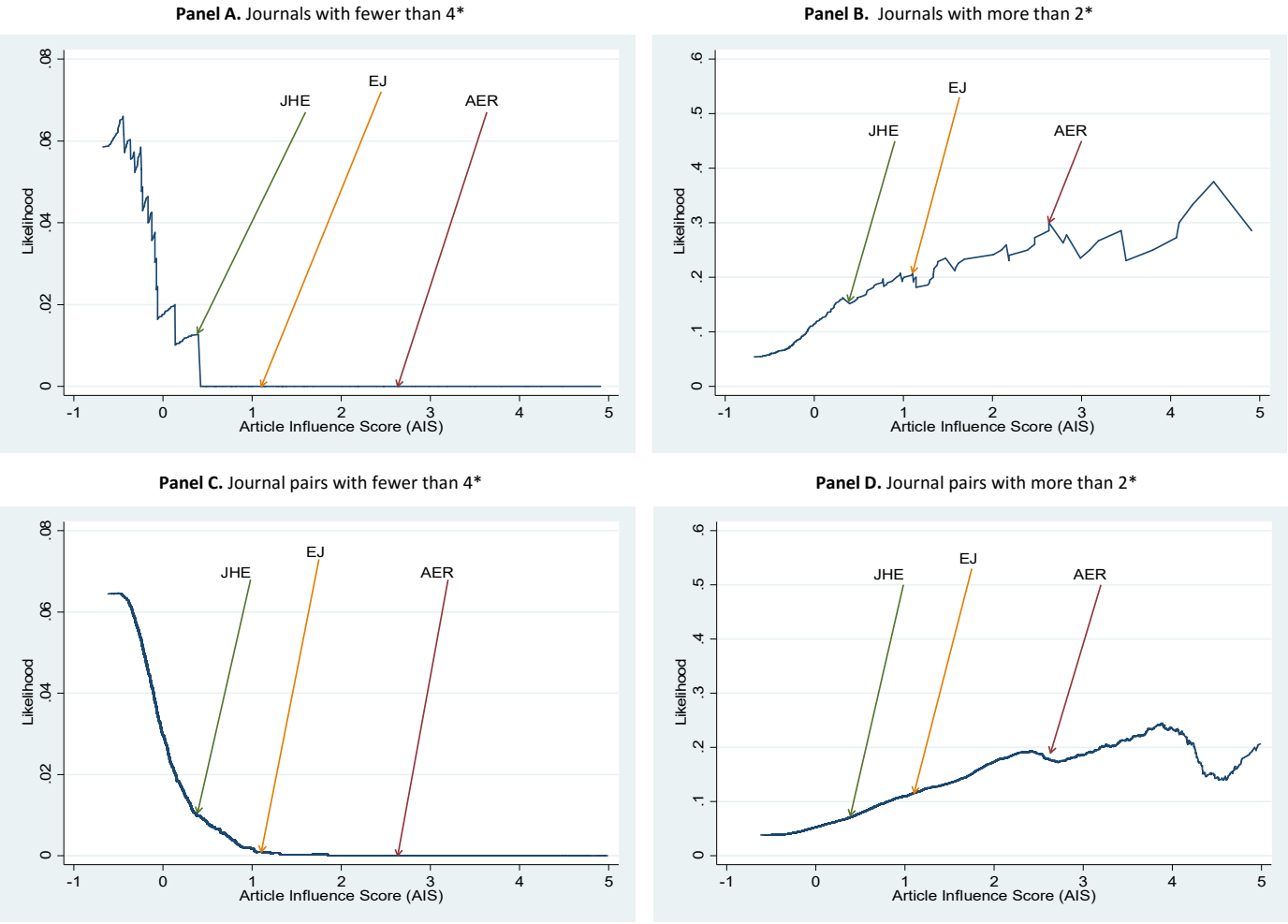
Note. Note. Columns (1) to (3) show results from a baseline specification that allows for tier-specific intercepts. Columns (4) to (6) show results from regressions that control for within-tier heterogeneity using a quadratic polynomial in Article Influence Score (AIS). Tier 2 in Panel A is defined as in Section 5, from the Journal of Health Economics (JHE) to the Economic Journal (EJ). Tier 2 in Panel B is defined to maximize the probability of including three-star journals (see Section 6 for details). 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 5. *** p<0.01, ** p<0.05, * p<0.1.

Figure 1: Article Influence Score distribution



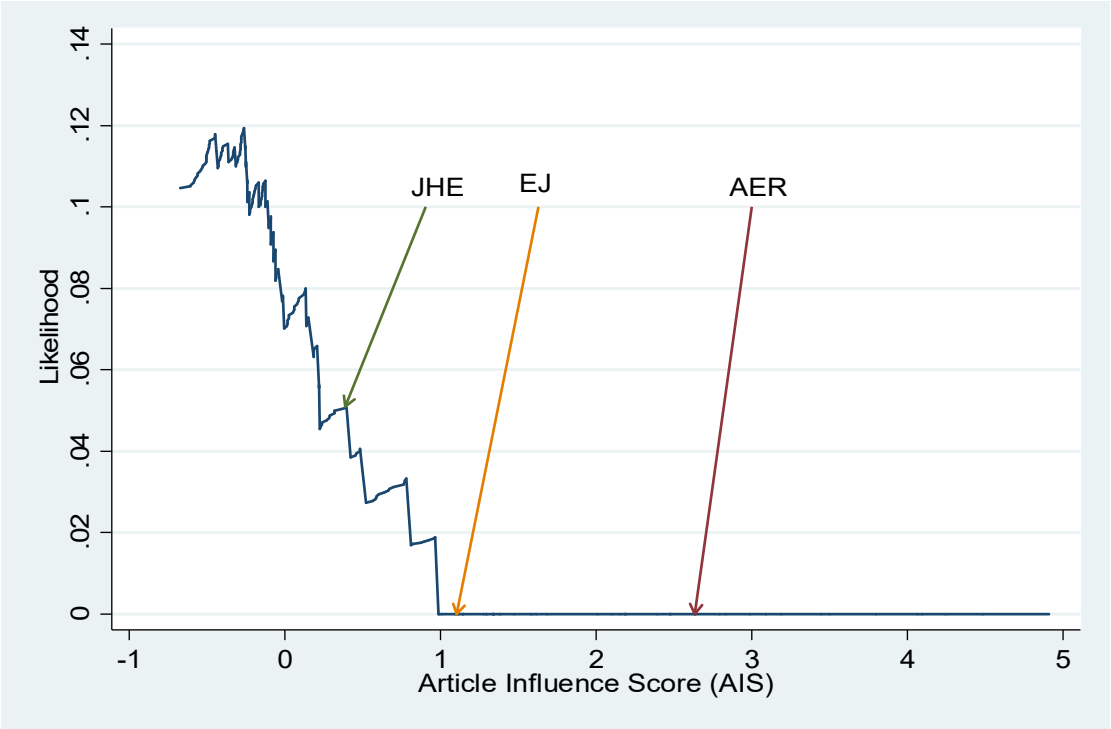
Note. The figure shows the distribution of submissions by the value of the Article Index Score (AIS) from the 2013 Thomson Reuters Journal Citation Reports database. Only outputs published in academic journals are considered. AIS is standardized to have zero mean and unit variance by research field. See Section 3 for details.

Figure 2: REF classification of outputs



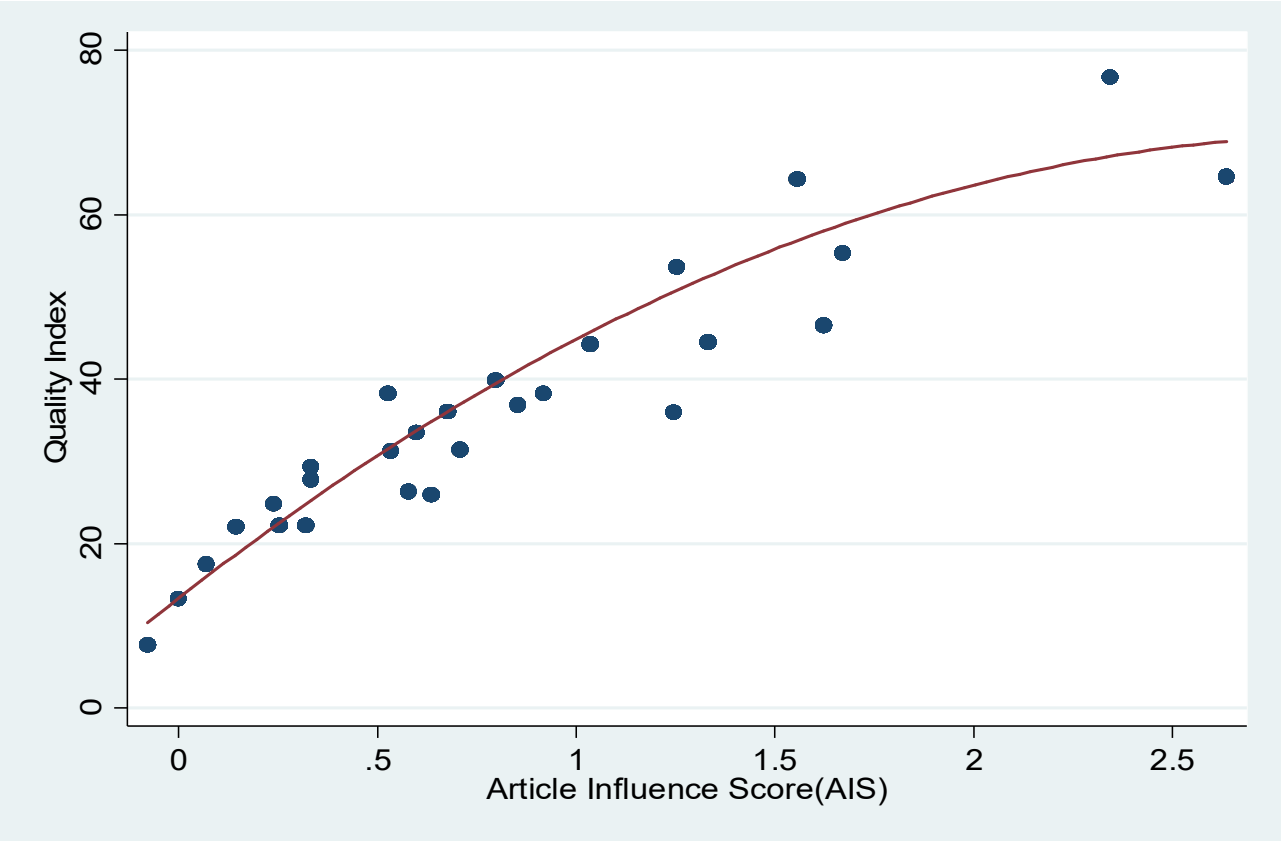
Note. This figure considers submissions by journal (in the top panels) and journal pairs (in the bottom panels) as explained in Section 4. Panel A shows the likelihood of being awarded less than four stars for publications in journals with standardized Article Influence Score (AIS) values at or above 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 American Economic Review (AER) or the Economic Journal (EJ) are awarded four stars. The Journal of Health Economics (JHE) represents a critical threshold. Panel B shows the likelihood of being awarded at least three stars for publications in journals with AIS values at or above a certain value. **Panel C** shows the likelihood of being awarded fewer than four stars for publications in journal pairs with AIS values at or above a certain value. **Panel D** shows the likelihood of being awarded at least three stars for publications in journal pairs 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.

Figure 3: Restrictions on the REF classification of outputs



Note. The figure replicates Panel A of Figure 2 when assuming that all publications in the "top-five" Economics journals (the American Economic Review, Econometrica, the Journal of Political Economy, the Review of Economic Studies and the Quarterly Journal of Economics) are awarded four stars. It shows the likelihood of being awarded fewer than four stars for publications in journals with standardized Article Influence Score (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 that of the Economic Journal (EJ) are awarded four stars. The support of the standardized AIS distribution is truncated at five because of the low number of journals above this value.

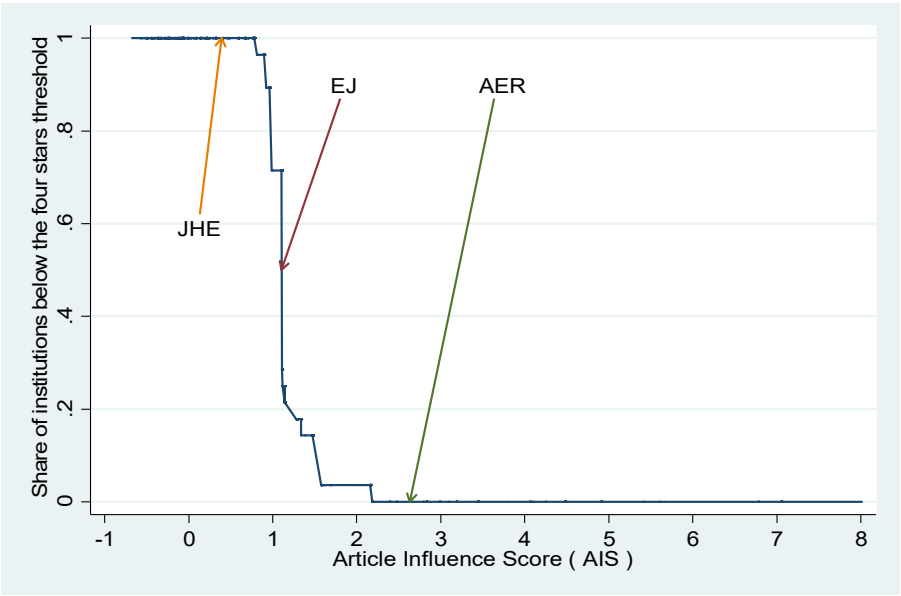
Figure 4: REF Quality Index and Article Influence Score



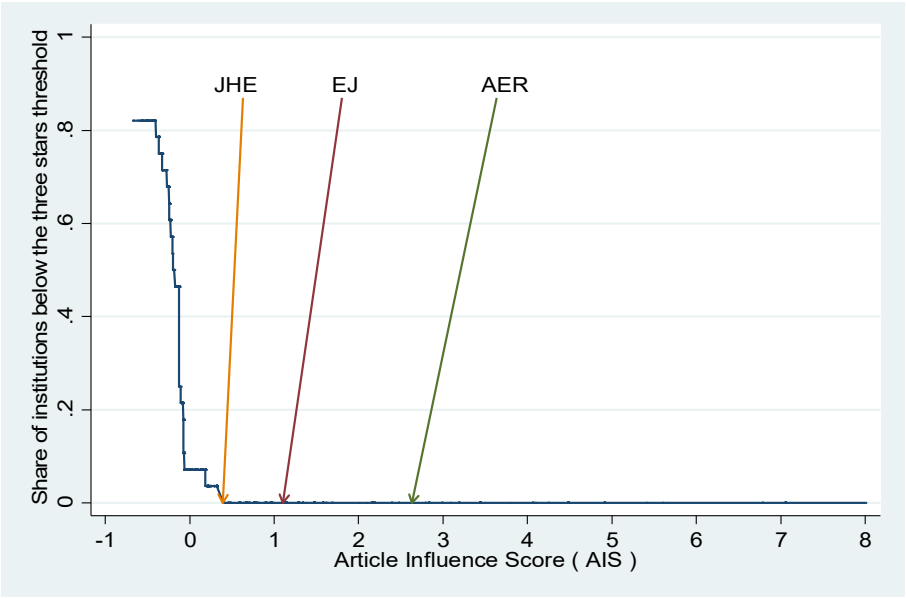
Note. The figure reports the scatterplot of an institution's Quality Index, on the vertical axis, against the average Article Influence Score (AIS) of all outputs submitted by the institution. Superimposed are predictions from a regression on linear and quadratic terms in AIS weighted by the number of outputs submitted. The Quality Index 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). See Section 2 for details and definitions.

Figure 5: Counterfactual classification based on Article Influence Score

Panel A. Threshold between three and four star publications



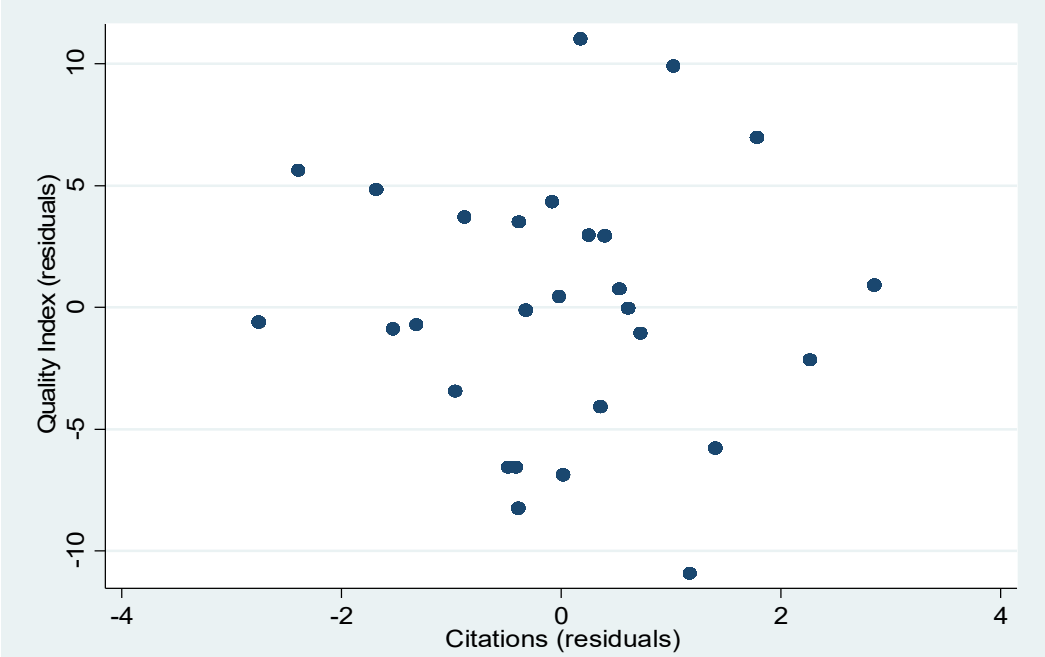
Panel B. Threshold between two and three star publications



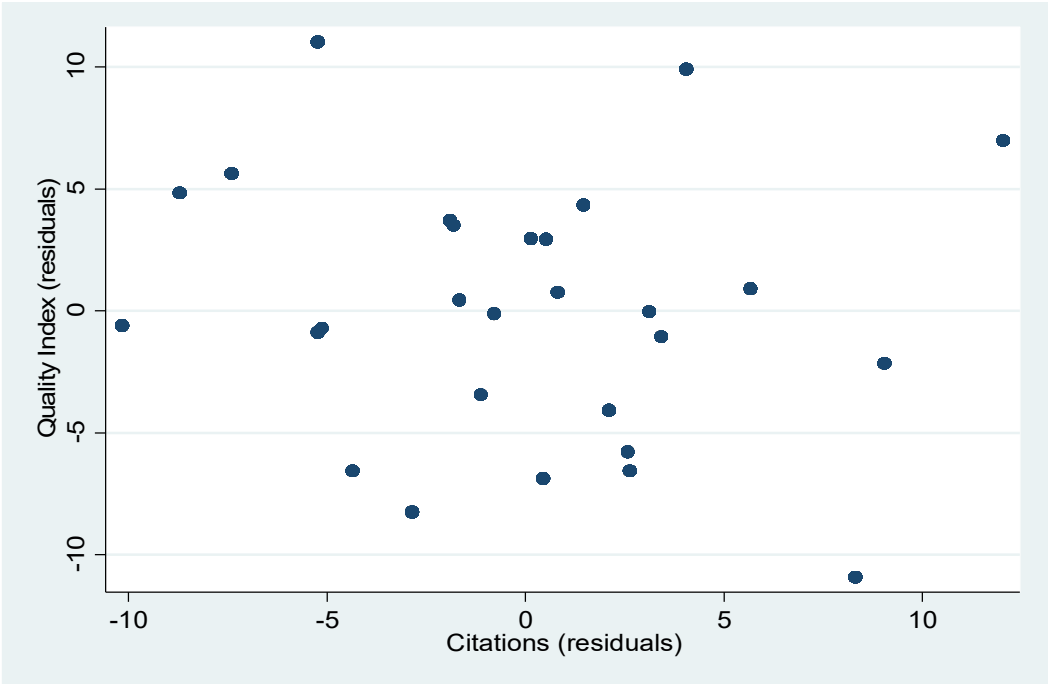
Note. All panels are derived by ordering outputs by their Article Influence Score (AIS), from the highest to the lowest. A classification of outputs based solely on AIS is maintained throughout. For each institution, outputs with the highest AIS are assigned four stars proportionately to the REF classification in column (2) of Table 1. The remaining outputs are classified using columns (3), (4) and (5) of Table 1 to assign three, two and one stars, respectively. Panel A in this figure 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, all publications in journals at least equal to the American Economic Review (AER) would be awarded four stars. In 43% of institutions, the Economic Journal (EJ) would represent the pass-mark between three and four stars. The Journal of Health Economics (JHE) would determine publications below four stars in all institutions. Panel B in this figure 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, the JHE would be the critical threshold. See Section 4 for details.

Figure 6: REF Quality Index and citation counts

Panel A. Elsevier's Scopus

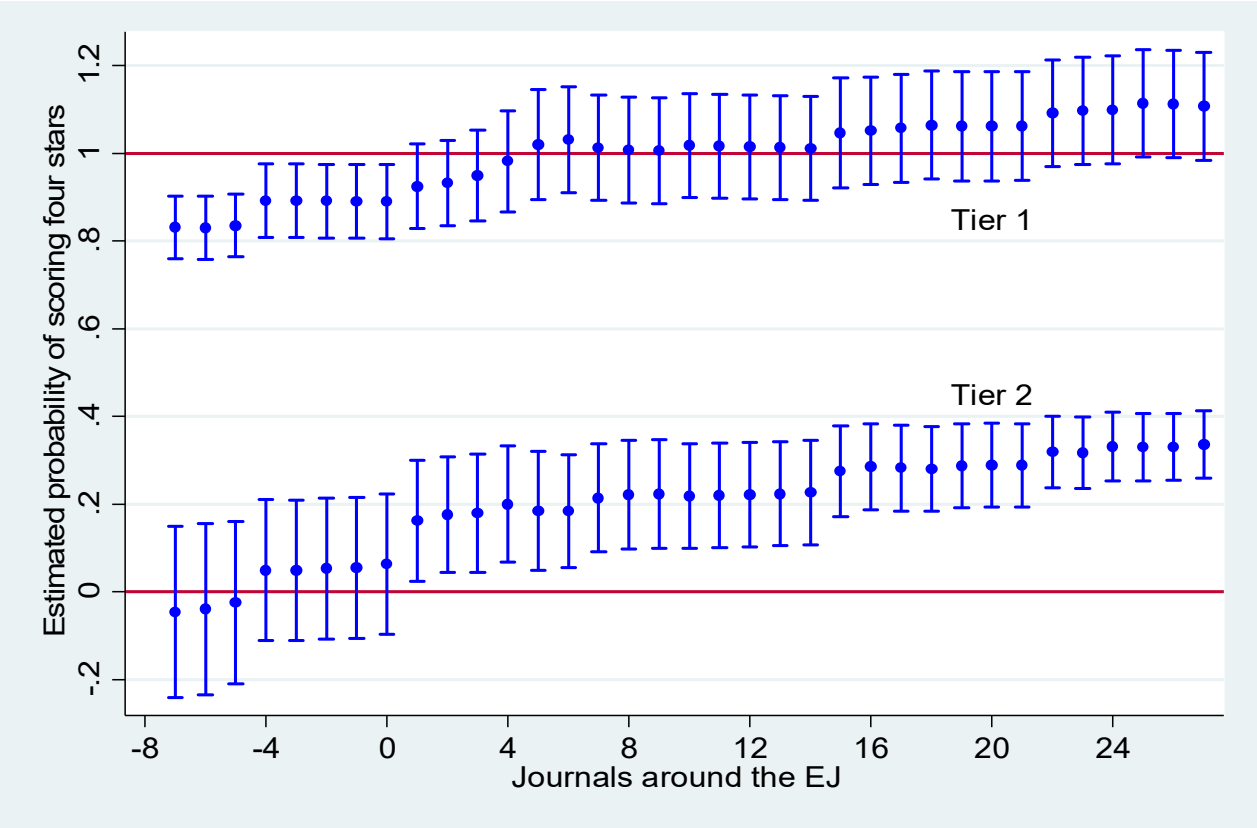


Panel B. Google Scholar



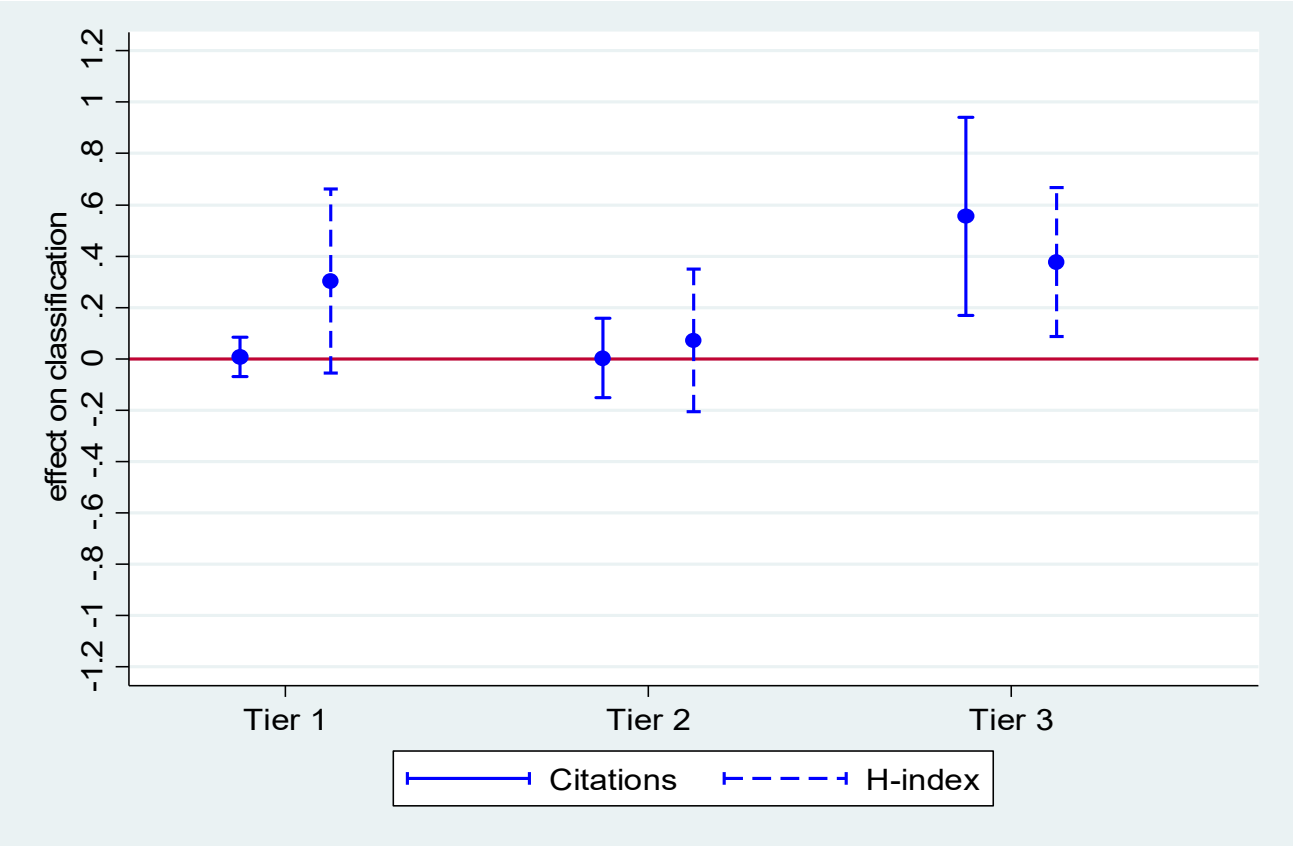
Note. This figure shows partial correlation graphs between institution Quality Index, on the vertical axis, and average number of citations, on the horizontal axis. Reported are scatterplots of residuals of these two variables from regressions on the average Article Influence Score and average H-index of outputs submitted. Panel A and Panel B use citations from Elsevier's Scopus and Google Scholar, respectively. See Section 4 for details.

Figure 7: Sensitivity analysis for the definition of Tier 2



Note. The figure shows estimated probabilities of scoring four stars by journal tier as a function of Tier 2 cut-offs, starting from journal tiers defined in Section 4. The 95% confidence interval is plotted for each estimate. It presents the sensitivity of the results to shifting the Tier 2 upper bound around the Economic Journal (EJ), when the lower bound is the Journal of Health Economics (JHE). The lower bound of Tier 1 is the journal ranked x positions above the EJ in the AIS distribution. See Section 6 for details.

Figure 8: Citations and h-index effects



Note. Reported are the estimated effects on the probability of scoring four stars (in Tier 1 and Tier 2) or three stars (in Tier 3) from a change in citations (solid lines) or h-index (dashed lines). The effects correspond to a change from the tenth percentile in tier to the ninetieth percentile in tier of citations count or h-index. The 95% confidence intervals are obtained from the specifications in columns (7)-(9) of Table 6.

Appendix

Derivation of graphs in Section 4

The following algorithm was considered in deriving Panels C and D of Figure 2. To fix ideas, consider the P pairs that can be formed by considering journals j and l among the J that were submitted. First, we construct the average AIS in this pair after weighting by the number of submissions by institution i to the two journals. This procedure defines $28 \times P$ different values of AIS. Second, we flag pairs for which the number of submissions (i.e., the sum of submissions to journals j and l) exceeds the number of four-star outputs (Panel C) or the number of one- or two-star outputs (Panel D) for at least one institution. The lines in Figure 2 are then computed as described in the main text.

Model specification and estimation

Outputs not in scientific journals (e.g., books) cannot be attributed a value of the AIS (W_j), although other attributes (like citations, Z_{jk}) are observed. Define the dummies B_j, C_j and P_j for books, book chapters and working papers respectively.[21] Equation ((5)) is modified as follows:

$$\alpha_j^d = \alpha_{0\tau}^d + \alpha_{1\tau}^d W_j(1 - B_j - C_j - P_j) + \alpha_{2\tau}^d B_j + \alpha_{3\tau}^d C_j + \alpha_{4\tau}^d P_j,$$

and W_j is set to zero for outputs without the AIS. Define:

$$X_{i\tau} \equiv \left(\sum_{j\in\tau} X_{ij}\right), \qquad WX_{i\tau} \equiv \left(\sum_{j\in\tau} W_j S_j X_{ij}\right), \qquad BX_{i\tau} \equiv \left(\sum_{j\in\tau} B_j X_{ij}\right),$$

$$CX_{i\tau} \equiv \left(\sum_{j\in\tau} C_j X_{ij}\right), \qquad PX_{i\tau} \equiv \left(\sum_{j\in\tau} P_j X_{ij}\right), \qquad Z_{i\tau} \equiv \left(\sum_{j\in\tau} \sum_{k=1}^{X_{ij}} Z_{jk}\right),$$

where $S_j \equiv (1 - B_j - C_j - P_j)$ is an indicator for outputs in scientific journals. Let K and H be the number of regressors in $Z_{i\tau}$ and $WX_{i\tau}$, respectively, with a slight abuse of notation to avoid the use of matrices.

The restrictions on the classification probabilities in Section 5 adds the following con-

[21] Articles in scientific journals without AIS, as well as other type of outputs, are included in the latter category.

straints:

$$\alpha_{i\tau}^{2}=0, \qquad \gamma_{\tau}=0, \qquad \tau=1,2, \qquad i=0,\ldots,3,$$
$$\alpha_{i\tau}^{1}=0, \qquad \gamma_{\tau}=0, \qquad \tau=1,2 \qquad i=0,\ldots,3,$$
$$\alpha_{i3}^{4}=0, \qquad \gamma_{3}=0, \qquad i=0,1,3,4.$$

By imposing these constraints one gets the following system of equations in $7\times(K+H+1)$ unknowns, plus additional 12 unknown parameters on output type dummies:

$$E\left[Y_i^4|\mathbf{X}_i,\mathbf{Z}_i\right]=\sum_{\tau=1}^{2}\alpha_{0\tau}^{4}X_{i\tau}+\sum_{\tau=1}^{2}\alpha_{1\tau}^{4}WX_{i\tau}+\sum_{\tau=1}^{2}\alpha_{2\tau}^{4}BX_{i\tau}+\alpha_{32}^{4}CX_{i\tau}+\sum_{\tau=1}^{2}\gamma_{\tau}^{4}Z_{i\tau},$$

$$E\left[Y_i^3|\mathbf{X}_i,\mathbf{Z}_i\right]=\sum_{\tau=1}^{3}\alpha_{0\tau}^{3}X_{i\tau}+\sum_{\tau=1}^{3}\alpha_{1\tau}^{3}WX_{i\tau}+\sum_{\tau=1}^{2}\alpha_{2\tau}^{3}BX_{i\tau}+$$
$$+\sum_{\tau=2}^{3}\alpha_{3\tau}^{3}CX_{i\tau}+\alpha_{43}^{3}PX_{i\tau}+\sum_{\tau=1}^{3}\gamma_{\tau}^{3}Z_{i\tau},$$

$$E\left[Y_i^2|\mathbf{X}_i,\mathbf{Z}_i\right]=\alpha_{03}^{2}X_{i\tau}+\alpha_{13}^{2}WX_{i\tau}+\alpha_{33}^{2}CX_{i\tau}+\alpha_{43}^{2}PX_{i\tau}+\gamma_{3}^{2}Z_{i\tau},$$

$$E\left[Y_i^1|\mathbf{X}_i,\mathbf{Z}_i\right]=\alpha_{03}^{1}X_{i\tau}+\alpha_{13}^{1}WX_{i\tau}+\alpha_{33}^{1}CX_{i\tau}+\alpha_{43}^{1}PX_{i\tau}+\gamma_{3}^{1}Z_{i\tau}.$$

The system is estimated from seemingly unrelated regressions imposing the following $3\times(K+H+1)$ adding up conditions for the classification probabilities, and 5 additional restrictions that involve the coefficients on the output type dummies:

$$\alpha_{0\tau}^{3}+\alpha_{0\tau}^{4}=1, \qquad \tau=1,2,$$

$$\alpha_{i\tau}^{3}+\alpha_{i\tau}^{4}=0, \qquad \tau=1,2, \qquad i=1,2,3,$$

$$\gamma_{\tau}^{3}+\gamma_{\tau}^{4}=0, \qquad \tau=1,2,$$

$$\alpha_{03}^{1}+\alpha_{03}^{2}+\alpha_{03}^{3}=1, \qquad i=1,3,4,$$

$$\alpha_{i3}^{1}+\alpha_{i3}^{2}+\alpha_{i3}^{3}=0,$$

$$\gamma_{3}^{1}+\gamma_{3}^{2}+\gamma_{3}^{3}=0.$$

After imposing these constraints, tier-specific intercepts and regressors in $Z_{i\tau}$ and $WX_{i\tau}$ yield $4\times(K+H+1)$ unknowns; 7 additional unknowns arise from output type dummies.

Our baseline specification estimates only tier-level intercepts, imposing $\gamma_\tau^d = 0$ and $\alpha_{i\tau}^d = 0$ for all τ's and d's and for $i = 0, \ldots, 4$. The regression adjustment includes tier-specific linear and quadratic terms in AIS. In the latter case, two additional sets of constraints are imposed to the estimation. First, continuity of classification probabilities is forced across tiers boundaries. Formally, we impose:

$$\alpha_{01}^4 + \alpha_{11}^4 W_1^{min} = \alpha_{02}^4 + \alpha_{12}^4 W_2^{max},$$

$$^4_{02} + \alpha_{12}^4 W_2^{min} = 0,$$

$$\alpha_{03}^3 + \alpha_{13}^3 W_3^{max} = 1,$$

where W_τ^{min} and W_τ^{max} are the lowest and highest value of the AIS, respectively, in tier τ. The number of parameters estimated in the various specifications is shown in Table A.6. Second, we impose that the two top journals in Economics (according to the AIS) among ones frequently submitted (see Table (3)) are deterministically awarded four stars.

Academic network

We scraped from the web curricula of the Economics and Econometrics panel members, ignoring assessors and secretariat staff who joined the panel near the submission deadline. The panel comprised 18 members, including a chair and a deputy chair. We determined the academic network by considering institutions where panel members and their co-authors were appointed throughout their professional career. We assigned to each member the institutions with which she had a professional position. For all co-authors, we then retrieved their affiliations and number of articles written with the panellists. We used this information to define various indicators of academic ties.

First, we compute the number of panellists ever employed by institution i. Define a dummy E_{ip} that takes value one if institution i has employed or is now employing panellist p. The index C_i is then computed as the sum of these dummies over the 18 panellists:

$$C_i = \sum_{p=1}^{18} E_{ip}.$$

Second, we compute the number of panellists' co-authors ever employed by institution i.

Define N_{ip} as the number of co-authors of panellist p affiliated with institution i.[22] The index C_i^{coaut} is then computed as the sum over the 18 panellists:

$$C_i^{coaut} = \sum_{p=1}^{18} N_{ip}.$$

[22] Alternatively, one could consider the frequency of each co-author in panellists' scientific production or take into account the different sizes of networks across panellists. The conclusions in Section 7 are robust to these alternative definitions of ties.

Table A.1: Comparison with the classification in Hudson (2013)

	(1) **Tier 1**	(2) **Tier 2**	(3) **Tier 3**
4*	73.33%	13.33%	13.33%
probable 4*	66.67%	33.33%	0.00%
possible 4*	50.00%	33.33%	16.67%
3*	21.95%	26.83%	51.22%
probable 3*	0.00%	14.29%	85.71%
possible 3*	0.00%	11.11%	88.89%
2*	0.00%	2.56%	94.87%
probable 2*	0.00%	0.00%	100.00%
possible 2*	0.00%	0.00%	100.00%
1*	0.00%	0.00%	100.00%

Note. The table shows the comparison between journal tiers defined in Section 5 and the classification in Hudson (2013). Cells report the share of journals in Tier 1, Tier 2 and Tier 3 for each of Hudson's (2013) categories. For example, 73.33% of journals classified as unambiguously four stars in Hudson (2013) belong to our Tier 1—see column (1); all journals classified as unambiguously one star belong to our Tier 3—see column (3). See Section 5 for details.

Table A.2: Estimation results controlling for publication characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Tier			Tier			Tier			Tier		
	1	2	3	1	2	3	1	2	3	1	2	3
4* ("world leading")	0.720***	0.090***		0.731***	0.081***		0.725***	0.086***		0.663***	0.131***	
	(0.027)	(0.026)		(0.024)	(0.024)		(0.027)	(0.026)		(0.028)	(0.026)	
3* ("internationally excellent")	0.280***	0.910***	0.120***	0.269***	0.919***	0.113***	0.275***	0.914***	0.115***	0.337***	0.869***	0.104***
	(0.027)	(0.026)	(0.036)	(0.024)	(0.024)	(0.034)	(0.027)	(0.026)	(0.034)	(0.028)	(0.026)	(0.031)
2* ("internationally recognised")			0.741***			0.738****			0.742***			0.754***
			(0.035)			(0.035)			(0.034)			(0.033)
1* ("nationally recognised")			0.139***			0.149***			0.143***			0.142***
			(0.022)			(0.021)			(0.020)			(0.016)
Google Scholar citations	Y	Y	Y	N	N	N	Y	Y	Y	Y	Y	Y
H-index of authors	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Economic journal indicator	N	N	N	N	N	N	N	N	N	Y	Y	Y
Number of journals in tier	53	88	143	53	88	143	53	88	143	53	88	143
Number of publications in tier	888	1,067	645	888	1,067	645	888	1,067	645	888	1,067	645
Mean of standardized AIS in tier	2.37	0.33	-0.31	2.37	0.33	-0.31	2.37	0.33	-0.31	2.37	0.33	-0.31

Note. The table shows estimation results when adding different publication characteristics to the baseline specification in columns (4)-(6) in Panel B of Table 5. Columns (1) to (3) include standardized Google Scholar citation count. Columns (4)-(6) includes the highest h-index among authors of a research output. Columns (7)-(9) combine all variables included in the previous columns. Columns (10)-(12) add an indicator for publication in an economic journal. The estimating equations are discussed in Section 5 and in the Appendix. *** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Tier classification for books and book chapters

	baseline specification			alternative specification		
	(1)	(2)	(3)	(4)	(5)	(6)
	Tier 1	Tier 2	Tier 3	Tier 1	Tier 2	Tier 3
			Panel A: books			
Cambridge University Press	X				X	
Harvard University Press	X				X	
John Wiley & Sons Ltd		X				X
Lambert Academic Publishing		X				X
Oxford University Press	X				X	
Princeton University Press	X				X	
Routledge		X				X
			Panel B: book chapters			
Cambridge University Press		X				X
Canadian Tax Foundation			X			X
Elsevier			X			X
Emerald Publishing			X			X
Harvard University Press		X				X
North-Holland			X			X
Oxford University Press		X				X
Palgrave Macmillan			X			X
Princeton University Press		X				X
Springer			X			X
University of Chicago Press		X				X

Note. The table lists editors of books and book chapters included among the REF submissions. Panels A and B present the tier classification of books and book chapters, respectively. Columns (1)-(3) show the baseline specification. Columns (4)-(6) show an alternative allocation used as a sensitivity check. See Section 5 for details.

Table A.4: Estimation results adjusting for unobserved quality

	(1)	(2)	(3)	(4)	(5)	(6)
		Tier			Tier	
	1	2	3	1	2	3
4* ("world leading")	0.688***	0.117***		0.672***	0.127***	
	(0.025)	(0.023)		(0.024)	(0.023)	
3* ("internationally excellent")	0.312***	0.883***	0.097**	0.228***	0.873***	0.098***
	(0.025)	(0.023)	(0.039)	(0.024)	(0.023)	(0.036)
2* ("internationally recognised")			0.782***			0.782****
			(0.038)			(0.037)
1* ("nationally recognised")			0.121***			0.120***
			(0.024)			(0.021)
Google Scholar citations (linear and quadratic terms)	Y	Y	Y	Y	Y	Y
H-index of authors	N	N	N	Y	Y	Y
Number of journals in tier	53	88	143	53	88	143
Number of publications in tier	888	1,067	645	888	1,067	645
Mean of standardized AIS in tier	2.37	0.33	-0.31	2.37	0.33	-0.31

Note. The table shows the estimation results after adding a quadratic term in citations to the specifications presented in Table A.2. Columns (1)-(3) include a linear and a quadratic term in Google Scholar citations. Columns (4)-(6) control for h-index of authors. See Section 7 for details. *** p<0.01, ** p<0.05, * p<0.1.

Table A.5: Differences between predicted quality and REF outcomes

VARIABLES	(1)	(2) Dummy for representation in the panel	(3) Panellists' co-authors affiliated
Dummy for one-star publications	-0.006	0.011	-0.029
	(0.025)	(0.028)	(0.027)
Dummy for two-stars publications	0.029	0.057	0.029
	(0.018)	(0.036)	(0.018)
Dummy for three-stars publications	-0.007	0.025	0.020
	(0.026)	(0.043)	(0.034)
Institution quality profile in RAE 2008	-0.000	-0.000	-0.001
	(0.001)	(0.001)	(0.001)
Institution impact profile in REF 2014	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Institution environment profile in REF 2014	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Relationship of institutions with panellists x dummy for one-star publications		-0.008	-0.000
		(0.015)	(0.005)
Relationship of institutions with panellists x dummy for two-stars publications		-0.014	-0.019**
		(0.028)	(0.009)
Relationship of institutions with panellists x dummy for three-stars publications		-0.006	0.002
		(0.023)	(0.009)
Relationship of institutions with panellists x dummy for four-star publications		0.029	0.018
		(0.020)	(0.013)
Constant	-0.001	-0.012	0.023
	(0.023)	(0.023)	(0.027)
Observations	112	112	112
R-squared	0.089	0.109	0.118
Method	OLS	OLS	OLS

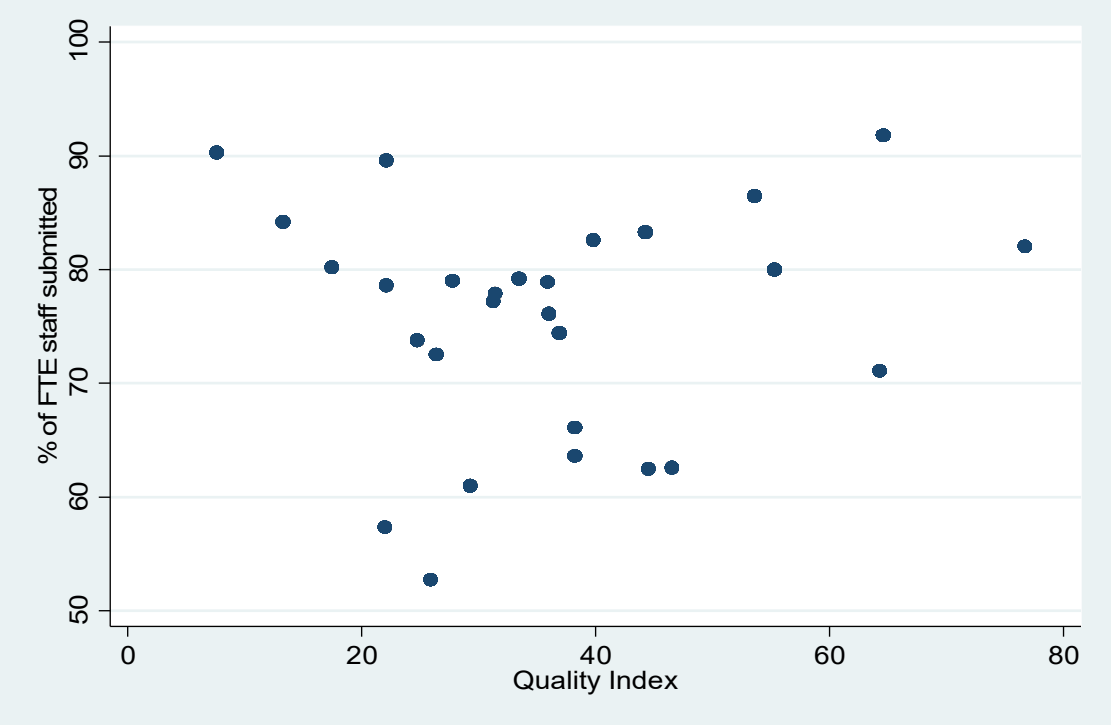
Note. The table reports the results of OLS regressions using, on the left hand side, the difference between the number of outputs awarded d stars and the number predicted by our model. The latter quantity is obtained from a model that adjusts for AIS, publication characteristics and unobservable research quality – see Table A.4. Column (1) includes number-of-stars dummies, and indicators of the research performance of institutions in the REF and in the 2008 RAE. Columns (2) and (3) add measures of connection to experts on the panel interacted with dummies for number of stars. Column (2) considers a dummy equal to one if at least one panellist was ever employed at the institution. Column (3) considers panellists' co-authors ever employed at the institution. Standard errors are clustered at the institution level. See section 7 for details. *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Number of unknown parameters

	(1) N. of W_j	(2) N. of Z_{jk}	(3) Output type dummies	(4) N. of parameters
Baseline specification	0	0	N	4
Number of parameters after adding:				
Article Influence Score	2	0	Y	19
Citations	2	1	Y	23
H-index	2	2	Y	27
Dummy for Economics outputs	3	2	Y	31

Note. The table presents the number of parameters estimated in each specification considered. Column (1) shows the number of regressors in the journal characteristics vector, denoted H in the Appendix. Column (2) reports the number of regressors in the publication characteristics vector, denoted K in the Appendix. The latter number is 2 when adjusting for AIS since a quadratic term is included. Column (3) indicates whether dummies for output type (e.g., book, book chapter) are included. Column (4) reports the total number of unknowns, equal to 4(K+H+1), plus 7 when output type dummies are included. This follows from summing the conditions for classification probabilities and our classification restrictions, as discussed in the Appendix.

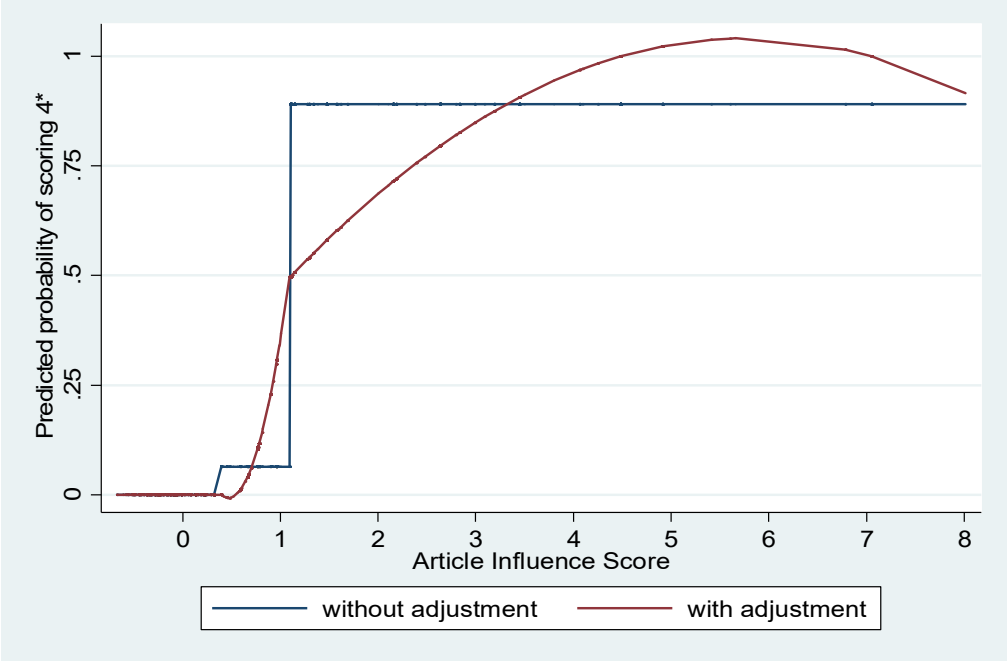
Figure A.1: REF Quality Index and the share of staff submitted



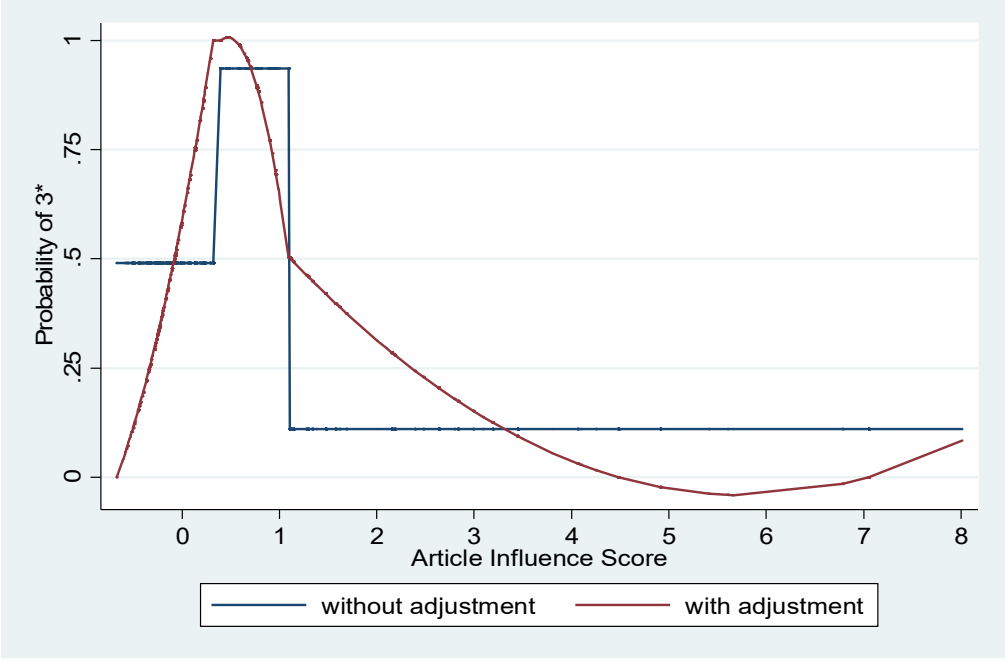
Note. The figure presents a scatterplot of the share of full time equivalent (FTE) members submitted by an institution, on the vertical axis, against the institution's Quality Index, on the horizontal axis. The latter index is computed using the current funding allocation formula, which depends on the incidence of top-quality outputs (80% and 20% classified as four- and three-star research, respectively, and no contribution of remaining outputs). See Section 2 for details and definitions.

Figure A.2: Predicted classification probabilities as a function of AIS

Panel A. Likelihood of scoring four stars



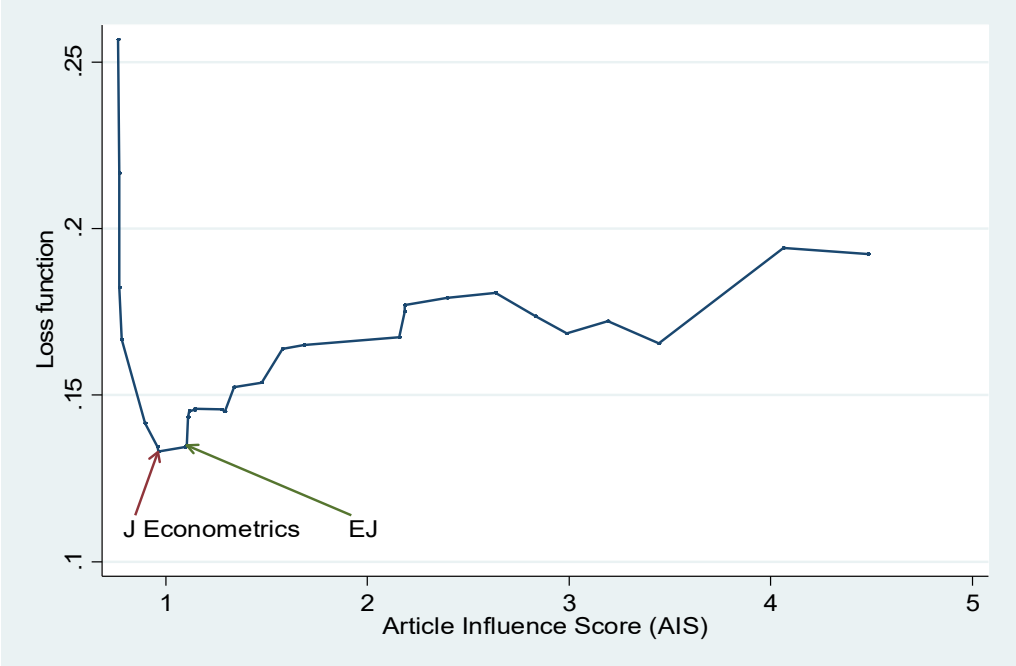
Panel B. Likelihood of scoring three stars



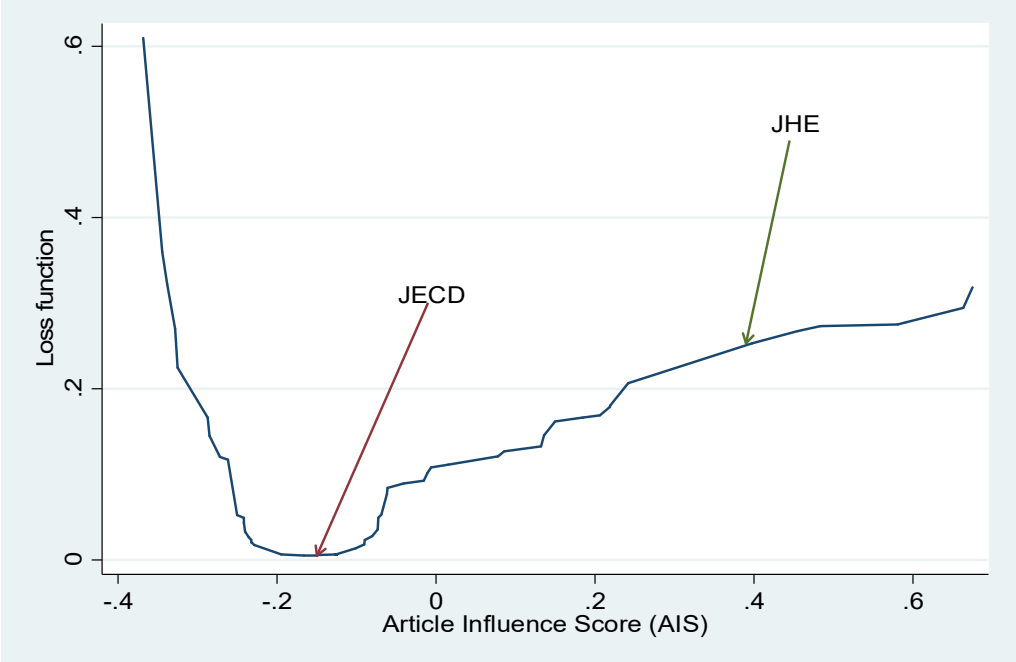
Note. The figure plots predicted probabilities of scoring four stars (Panel A) and three stars (Panel B) as a function of AIS. Estimates are derived from results in Panel A of Table 5. The blue lines show predictions

Figure A.3: Optimal definition of Tier 2

Panel A. Tier 2 upper limit



Panel B. Tier 2 lower limit



Note. The figure is obtained by using a grid search over 60\times60 possible choices and varying the upper and lower limits of Tier 2. We start by setting the lower limit on Tier 2 at JHE. We then select 60 journals with AIS in a window centred on the EJ and use them to define alternative upper limits on Tier 2. This defines a range of 60 possible intervals between JHE and the new upper limit, which we use iteratively to estimate our classification probabilities. We then select the definition of Tier 2 to maximize between-tier distance in classification probabilities while ensuring within-tier homogeneity. Panel A reports the loss function resulting from different choices of the upper limit on Tier 2. Panel B reports the loss function for the lower limit. See Section 6 for details.