

BAM 2023 CONFERENCE

US
UNIVERSITY
OF SUSSEX

BUSINESS
SCHOOL

The role of algorithmic management in assessing output quality: a study of REF 2021

Adina Dudau

University of Glagsow

Serafeim Tsoukas

University of Glagsow

Anna Morgan-Thomas

University of Glagsow

Pawel Gaska

University of Glagsow

Beyond declarations: Journal rankings, REF outcomes and responsible assessment

July 13, 2023

Abstract

Responsible assessment is an evaluation practice that relies on peer review and expert judgement rather than quantitative research metrics. By departing from objective indicators such as journal impact factors or journal rankings, the practice strives to produce more valid, accurate and fairer assessments. Combining data for journal outputs from REF2021 for unit of assessment 17 (business and management studies) with other indicators (journal rankings, publication citations and DORA signatories), this paper contrasts the outcomes of responsible assessment with metrics. The analysis finds evidence that CABS journal ranking positively and significantly affects the research quality of the institutions, as measured by the REF outcomes. Further, when we introduce institution-level heterogeneity, we find that the effect is stronger for high-profile outputs, while DORA affiliation is quantitatively unimportant. The implications of these findings are profound; they enhance our understanding of the responsible use of metrics in research evaluations, and they pave the way to a more nuanced deployment of responsible assessment.

Keywords: Journal quality; Ranking; Research evaluation; Research Excellence Framework

1 Introduction

For individuals and institutions alike, publications in peer-reviewed academic journals remain a principal marker of academic worth. Indeed, evidence from diverse disciplines shows that publication record influences access to academic jobs, promotion prospects, career goals, funding, teaching loads and pay (DeFraja et al. 2019, Anderson et al. 2021). At the institutional level, collective evaluations of academic output drive reputation, staff retention and recruitment efforts, student enrolment and institutional funding (Hicks 2012, Ryazanova et al. 2017). Consequently, publications serve as a proxy for productivity, performance and quality for individuals, departments, schools and universities. For better or worse, publishing in academic journals persists as the dominant logic of academic practice (Aguinis et al. 2020, Osterloh and Frey 2020, Linton and Xu 2022).

Against this backdrop, long-standing debate concerns methods and tools for the assessment of publication quality (Thomas et al. 2020, Williams 2020). Competing regimes of evaluation have emerged. There has been a significant rise in bibliometric methods such as citation counts or citation-based indexes including journal impact factors (Ramani et al. 2022) and journal rankings (Vogel et al. 2017, Osterloh and Frey 2020). As computational powers have increased, additional indicators have emerged, including web citations and altmetrics derived from social media (Wilsdon et al. 2015). However, the burgeoning landscape of quantitative indicators and increasing reliance on those in quality evaluations (Anderson et al. 2021, Ramani et al. 2022) have received significant backlash from the academic community (Mingers and Willmott 2013, Chapman et al. 2019, Guthrie et al. 2019).

The calls for more nuanced quality judgement involving peer review have recently coalesced under the umbrella of responsible assessment (Curry et al. 2022, The European Commission 2022). In essence, responsible assessment counters the reliance on aggregate metrics such as journal impact factors in evaluating research. Instead, it argues for the use of qualitative evaluations and peer assessment concurrent with broadening the scope and range of evidence that underpins academic productivity (Curry et al. 2022). In relation to publications, it encourages the use of peer reviews

of individual outputs (Bertocchi et al. 2015, Checchi et al. 2021).

Several major developments underpin the emphasis on responsible assessment. For example, the San Francisco Declaration on Research Assessment (DORA), with its opposition to journal impact factors, has provided much impetus for change. Similar policy initiatives include Leiden Manifesto, the Responsible Metrics movement in the UK (Wilsdon et al. 2015) and Coalition for Advancing Research Assessment in the EU (The European Commission 2022, Curry et al. 2022). National evaluation systems seem to embrace responsible assessment in the UK, Italy, China, Latin America and Switzerland (Abramo et al. 2019, Delegado-Lopez-Cozar et al. 2021). Further evidence of change includes a growing number of institutional DORA signatories (Chapman et al. 2019) and case studies of universities implementing the principles (Saenen et al. 2021). Given the above, we seem to be actively moving away from research metrics and towards responsible assessment.

Yet, beyond declarations, the evidence concerning the responsible assessment practice remains patchy at best. In North America, a recent study of institutional policies guiding academic promotion and tenure evaluations shows continuing dependence on journal impact factor (McKiernan et al. 2019). Some new national assessment systems explicitly adopt metrics (Fry et al. 2023, Civera et al. 2020). At the academic-field, data from Economics and Econometrics shows persistent alignment between peer-reviewed REF scores and the metrics of journal rankings (Hole 2017, Battistin and Ovidi 2022, Linton and Xu 2022). This corroborates well with recent evidence of journal rankings continuing to influence the assessment of individual performance, attitudes, behaviours and goals (Walker et al. 2019, Currie and Pandher 2020, Osterloh and Frey 2020, Ramani et al. 2022). Preference for non mainstream (Chavarro et al. 2017) or lower ranked outlets (Salandra et al. 2022) seems limited, the marginalisation of individuals and declining intellectual diversity seem to persist in multiple fields (Guthrie et al. 2019, Anderson et al. 2021).

In our paper, we interrogate the divide between declarations and evidence of practice. Specifically, using empirical evidence from Research Excellence Framework 2021 (REF2021), we contrast the outcomes of responsible assessment with metrics to explore whether the two diverge. Thus we ask: have responsible research assess-

ment declarations made an impact on the outcomes of institutional research evaluation? We combine REF2021 with other data (journal rankings, publication citations, DORA signatories) to infer correspondence between metrics and peer review.

The analysis centers on the UK and REF for two reasons. First, REF is the largest and the longest running research assessment exercise in the world (Banal-Estañol et al. 2023, Checchi et al. 2021). Second, unlike other national equivalents, REF overtly embraces responsible assessment by (1) relying on peer-review, (2) explicitly discouraging the use of the objective indicators and (3) encouraging submission of varied research outputs, not just journal articles (Blackburn et al. 2023). Although REF assessment goes beyond outputs to include research environment and impact, following other studies (Bertocchi et al. 2015, Hole 2017, Traag and Watman 2019) the paper focuses on outputs.

We further narrow the scope to one Unit of Assessment (UoA) within REF: UoA17, Business and Management (B&M). The subject represents the single largest UoA by the number of submitting institutions and the second largest in terms of the number of staff. Given the relative research underfunding of BM (CABS 2020), the quality of outputs plays a disproportionately important role in assessing research performance. Considering intense global competition for staff and students, business schools are subject to multiple ranking and accreditation regimes, and the use of journal rankings is common and broadly accepted (Walker et al. 2019). Well established rankings such as the Academic Journal Guide, produced by Chartered Association of Business Schools (CABS), offer a seemingly appropriate proxy for objective measurement (Vogel et al. 2017).

The question guiding our research is whether the data on the ground supports the declared preference for responsible assessment. In answering this question, we make two contributions to responsible assessment. First, we bring clarity to the relationship between the outcomes of metric assessment and of responsible assessment, as we investigate the relationship between discourse and practice in national research evaluation projects such as REF. Second, we reflect on how we might create conditions for further engagement with, and implementation of, responsible assessment that does not rule out the use of metrics in national research evaluation exercises, at

institutional and national levels. In doing so, we can assist in (un)covering any gaps and therefore pave the way for responsible research assessment in REF2028, as well as similar national research evaluations exercises around the globe.

The remainder of the paper is structured as follows. Section 2 discusses the literature on assessment of research outputs and stipulates the research questions. Section 3 describes the data, presents summary statistics and outlines our methodology. Section 4 presents the main empirical results. The paper concludes with discussion of findings and their implications for policy and future research.

2 Assessment of research outputs

Research quality is a growing concern for multiple constituencies in the academic landscape. The purpose of quality assessment is to judge the overall caliber of research ([Wilsdon et al. 2015](#)) and the process is underpinned by varying definitions of quality, excellence and impact. The assessment is supported by a broad range of tools and methods that produce measures of excellence ([Thomas et al. 2020](#), [Williams 2020](#)), which in turn classify individuals, subjects and institutions in order of diminishing quality ([Hicks 2012](#), [Civera et al. 2020](#)). Policy makers use the assessment to demonstrate public value and justify ongoing support for academic research ([Hicks 2012](#), [Banal-Estañol et al. 2023](#), [Fry et al. 2023](#)). For research funders, excellence as a marker of past success, is an important predictor of scholarly potential, thus driving decisions about allocation of financial support. Institutions benefit from research esteem indirectly through reputation, staff and student recruitment ([Ryazanova et al. 2017](#)); they benefit directly through allocations of public funds that disproportionately support high performers ([Civera et al. 2020](#), [Banal-Estañol et al. 2023](#)). For individuals, excellence means access to academic jobs as well as enhanced job conditions, pay, promotion and recognition ([DeFraja et al. 2019](#), [Walker et al. 2019](#)). Although the notion of research excellence is broader than publications ([Williams 2020](#)), the outputs remain a key measure of performance in academic assessment ([Bertocchi et al. 2015](#), [Traag and Watman 2019](#)). Acknowledging that research assessment has other dimensions ([Curry et al. 2022](#), [The European Commission 2022](#)),

this paper focuses on outputs.

2.1 Metrics assessment

In general, the assessment of outputs adheres to two regimes of evaluation: a metrics model or a responsible assessment model (Wilsdon et al. 2015). The metrics model quantifies research excellence by using objective proxies. In the context of outputs, quality typically means scholarly impact, which is the work’s influence on further research (Aguinis et al. 2020, Williams 2020), and citations and citation based indexes are the most prevalent objective proxies of excellence (Chapman et al. 2019, Checchi et al. 2021). Whether in normalized (field weighted) or crude form, citation counts may be the sole metric of quality or transformed to produce indexes (e.g., h-index, Journal Impact Factor) or rankings (e.g., CABS) (Ramani et al. 2022). For example, the Journal Impact factor (JIF) reflects the average number of citations per year that articles published in a particular journal receive during the previous two years. Although JIF attempts to evaluate relative importance and influence of a journal within a field, it is commonly deployed to infer the quality of individual research outputs (e.g., Ramani et al. 2022) and thus the scholarly capability of the authors (Delegado-Lopez-Cozar et al. 2021, Battistin and Ovidi 2022). JIF has been deployed in national assessment exercises (Civera et al. 2020, Fry et al. 2023). A recent study by McKiernan et al. (2019), for example, shows that JIF remains widely used in the United States and Canada in the institutional policies and guidelines concerning review, promotion and tenure evaluation; 87% of institutions support its use and none prohibits it.

In addition to journal-level metrics, several specialist indexes attempt to capture excellence at an individual level. For example, the h-index is a proxy of researcher quality, taking into account the number of publications and their scholarly impact in citations (Chapman et al. 2019). As a summary metric, h-index is an indicator the individual’s productivity and scholarly impact via citations (Mryglod et al. 2015). The h-index combines the volume of outputs and citations over time, thus inherently advantaging established scholars. Various extensions (hc-index, hIa-index; q-index)

aim to address biases related to the impact of time or self-citation (see, for example, [Ramani et al. 2022](#)).

Metric indicators of research excellence also include journal rankings. Unlike citation counts and their derivatives, most journal rankings rely on algorithms to calculate a relative position within a list ([Buttler 2008](#), [Anderson et al. 2021](#)). That is, rankings combine citations with other indicators to arrive at a rank. In business and management, most widely used journal lists include the Academic Journal Guide produced by CABS (hereafter CABS AJG). CABS AJG assigns a numerical value to journals within the subfields of business and management, ranging from 4 to 1. Although the rank of an individual journal is metrics informed, peer review plays a role: the final rank takes into account scientific committee recommendations based on the subject expert judgments ([Guthrie et al. 2019](#)).

Metric indicators of performance have grown in popularity for several reasons. Academia is a highly competitive profession where intense rivalry concurs with the lack of clear-cut indicators of quality, productivity and success ([Chapman et al. 2019](#)). Metrics fill this gap: they provide independent, seemingly objective and bias-free measures ([Wilsdon et al. 2015](#)). Because the same metrics apply across different research fields, cross-comparisons of items, individuals, groups or institutions are relatively straightforward ([Anderson et al. 2021](#), [Ramani et al. 2022](#)). Research assessment embraces metrics in part, to keep the cost of research evaluations down and in part due to isomorphic pressures for audit and evaluation ([Haley et al. 2017](#)). By most accounts, their popularity in research assessment is growing ([Aguinis et al. 2020](#), [Osterloh and Frey 2020](#)).

However, the widespread use of metrics faces strong opposition ([Wilsdon et al. 2015](#), [Guthrie et al. 2019](#), [Saenen et al. 2021](#)). For example, JIF has been criticized not only for the flawed way in which it may infer individual performance ([Haley et al. 2017](#), [Ramani et al. 2022](#)), but also for its disregard of the number of citations per article, the age of the articles and the self-citation rate ([Ramani et al. 2022](#)). Additionally, fields of research have different citation patterns, making it difficult to compare JIFs across fields ([Snyder and Bonzi 1998](#)). As such, there are calls to use JIF in conjunction with other indicators ([Bornmann and Marx 2016](#)) or to

abandon it altogether (Ramani et al. 2022, Saenen et al. 2021). The continued use may have caused toxic effects on careers, from jeopardizing PhD awards, job applications, tenure decisions, rewards and recognition and even grant programs (see, for example, Delegado-Lopez-Cozar et al. 2021, Ramani et al. 2022). More broadly, metrics are linked to an erosion of research integrity and the societal impact of science (Anderson et al. 2021, Williams 2020). Importantly, metrics also lead to goal displacement (Civera et al. 2020) where achievement of excellence in research is replaced with the goal of obtaining a high metric. Indeed, gaming behaviors, short-termism, erosion of creativity and general “dehumanization” of research cultures are among the unintended and undesirable consequences of using metrics to assess research (Hicks 2012, Chapman et al. 2019, Guthrie et al. 2019).

2.2 Responsible assessment

Against the tide of metrics (Wilsdon et al. 2015), calls for responsible assessment are gaining prominence (Saenen et al. 2021, Curry et al. 2022). For example, the San Francisco Declaration on Research Assessment (DORA) has now been signed by over 20,000 research institutions worldwide.¹ In Canada, the support for DORA now includes all major funding agencies as signatories (Chapman et al. 2019). In Europe, the 2015 Leiden Manifesto articulated the principles of qualitative judgement (REF) leading to the Coalition for Advancing Research Assessment (CoARA) (Curry et al. 2022). A recent EU research policy review also embraces responsible assessment (The European Commission 2022). At national levels, the UK and Italy continue to promote expert judgments (Franco-Santos and Otley 2018) and similar initiatives have been reported in China, Latin America and Switzerland (Delegado-Lopez-Cozar et al. 2021). Common to these movements is the notion that quality objective metrics do not reflect quality (Blackburn et al. 2023). Nuance is arguably needed to account for idiosyncratic differences among scholarly domains, research production and the consequent authoring and publishing variations (Traag and Watman 2019). The key role of academic peers as arbiters of judgement within fields is essential and should

¹See <https://sfdora.org>.

not be undermined by automating the review process (Curry et al. 2022).

Admittedly, much of the impetus for responsible assessment originates from DORA. The declaration maintains that research assessment should be based on the quality of research output, not the impact factor of the journal that publishes it. Moreover, an evaluation should take into account a broad range of research outputs, such as publications, data, software, and other non-conventional items. Furthermore, research assessment should be based on peer review. This principle goes right to the heart of academic publishing, where peer review is widely accepted as a reliable method of scrutiny over research outputs that ensures experts in the relevant field evaluate the research. Finally, the assessment should take into account the context in which the research was produced. Indeed, research outputs should not be evaluated in a vacuum; factors such as academic discipline, funding environment, institutional culture, and available resources matter. The DORA principles offer a new approach to evaluating research outputs that attempt to alleviate key disadvantages of metrics (Ramani et al. 2022). Publishing guidelines on how the principles may be implemented in practice and introducing the TARA toolkits offer means of implementation.

Research Assessment Exercise in the UK explicitly embraces the principles of responsible assessment. Indeed, the UK pioneered performance management through national research evaluations, with the first exercise taking place in 1986 (Cooper and Otley 1998, Kellard and Śliwa 2016, Banal-Estañol et al. 2023). REF2021 was the eighth iteration, with assessment involving peer evaluation of research outputs, impact case studies, and research environments at the institution and discipline levels (Blackburn et al. 2023). With respect to outputs, the criteria of significance, originality and rigor denote excellence. The process of assessment relies on panels of experts from multiple disciplines who exercise academic judgment to identify multiple markers of excellence (Blackburn et al. 2023). Similar to its predecessors, REF 2021 is a significant event in the UK higher-education landscape, as it determines future distribution of research funding (DeFraja et al. 2019). In addition, universities use REF results to signal reputation, benchmark their research performance and set future research priorities (Pidd and Broadbent 2015, Banal-Estañol et al. 2023).

Peer review, the cornerstone of responsible assessment practice, is not free from limitations (Curry et al. 2022). For instance, inconsistency is a key problem (Traag and Watman 2019); expert judgments of the same outputs do not necessarily tally, and some evidence suggests that inter-rater variability is quite significant (Bertocchi et al. 2015). Review process is necessarily slow, inefficient, expensive, subjective and non transparent (Wilsdon et al. 2015). Furthermore, peer-review can lead to biases in favor of dominant institutions or research paradigms and against emerging or interdisciplinary fields (Chapman et al. 2019). Subjectivity and non transparency surrounding outcomes present a particular challenge for institutions and individuals. For example, in the UK, detailed peer review results of the REF at the output level are not publicly available (Blackburn et al. 2023). Yet, institutions and research administrators require inoperable and objective indicators of quality because cross-comparisons are needed for policy development and research management (Anderson et al. 2021). For individuals, transparent markers of quality determine the rules of the game and signal the academic worth of different options (Williams 2020). Thus, the lack of transparency surrounding responsible assessment brings uncertainty and confusion (Chapman et al. 2019).

2.3 Responsible assessment versus research metrics

To date, empirical evidence that robustly compares and contrasts the outcomes of metrics and responsible assessment is only partial and largely inconclusive (Thomas et al. 2020). In general, critiques of objective measures tend to be one-sided. A typical paper argues against metrics by discussing the shortcomings of a particular index; for example, JIT (Ramani et al. 2022) or h-index (Chapman et al. 2019). Sometimes, a critique may be accompanied by a suggestion of an alternative way forward (Osterloh and Frey 2020). Although valuable, the critiques typically do not offer empirical evidence that shows divergence between metrics and alternative approaches. This is understandable; an alternative approach to evaluation may yet to be deployed in assessment practice (e.g Osterloh and Frey 2020) or its use may be too limited to permit comparisons. Nonetheless, the critiques are one-sided because

they tend to exclude evidence that alternatives result in fundamentally different outcomes of the assessment exercise ([Thomas et al. 2020](#)).

Few studies take a more evidence-based approach to comparison. For instance, past analyses contrast bibliographic metrics (citation counts and citation based indexes, such as JIT or h-score) with peer review to produce mixed results. [Wilsdon et al. \(2015\)](#), for instance, analyse the correlation between field-normalized citations based on Scopus data and peer review. Their findings show low correlation values ranging from 0.2 to 0.4. [Battistin and Ovidi \(2022\)](#) support these findings using REF2014 data for economics. With regard to h-index, [Mryglod et al. \(2015\)](#) find that the departmental h-index does not predict peer review score. Nonetheless, evidence to the contrary also exists with several authors reporting high correlations between citations and peer review scores ([Pride and Knoth 2018](#)). [Bertocchi et al. \(2015\)](#), for example, uses data from the national research assessment in Italy to show significant correspondence between h-index and peer review. Similarly, [Checchi et al. \(2021\)](#) compare REF with Italian VQR outcomes to find a high level of correspondence between peer assessment and bibliometrics.

When it comes to journal rankings, the evidence showing correspondence between metrics and peer assessment is also largely inconclusive. For example, [Taylor \(1994\)](#) analyzes B&M data and finds a clear correlation between peer-review score and journal rank. [Linton and Xu \(2022\)](#) use REF2021 data on unit of assessment 16 (Economics and Econometrics) to show that the CABS journal ranking is very highly correlated with the REF GPA score. In this sub-field, evidence points to consistent and high correspondence between journal ranks and responsible assessment ([Bertocchi et al. 2015](#), [Hole 2017](#), [Linton and Xu 2022](#)). However, the evidence from [Blackburn et al. \(2023\)](#) for UoA17 in REF2021 contradicts these findings; it reports low correlations between peer-review and CABS journal rank.

In a recent paper, [Traag and Watman \(2019\)](#) shed some light on the possible sources of disparity by highlighting two points. First, they argue that comparisons in the outcomes of metrics and responsible assessment should consider the level of aggregation (individual output versus aggregate outputs at an institutional level). Alignment should be sought at a higher level of aggregation (institution) and not

individual output. Using evidence from REF2014, their findings show high levels of agreement between metrics and peer review at the institutional level. Second, their findings highlight that correlation analysis may not be the most appropriate technique for capturing the agreement between metrics and peer review because the coefficient is dependent on the sample size (e.g., the number of outputs or individuals). That is, size-dependent Pearson’s correlation may be high or low depending on the number of items analyzed and not because of the underlying strong or weak relationship between metrics and peer review. Based on the evidence from multiple disciplines, their study calls for further investigations at an institutional level that focus on subject fields and move away from correlation as the key measure of relationship.

Admittedly, comparative analyses of responsible assessment have been hampered by the lack of transparency inherent in peer review (Thomas et al. 2020). To illustrate, our literature search fails to identify studies that examine how DORA principles on assessment affect outcomes either with regard to research outputs or other aspects of research performance. By contrast, broader evidence supporting the use of metrics does exist. For example, journal rankings continue to be a good predictor of individual performance, as reflected by career progression or pay (Hamermesh 2018, Walker et al. 2019).

This leaves academia in a conundrum. On the face of it, the move toward responsible assessment is gaining momentum. The number of signatories for DORA is steadily increasing; policy within the broader research environment and institution is changing (The European Commission 2022); research funders are embracing responsible assessment (Wilsdon et al. 2015) as do national research assessment exercises (Curry et al. 2022). Yet, evidence showing persistent reliance on metrics is also plentiful (Bertocchi et al. 2015, McKiernan et al. 2019, Walker et al. 2019, Aguinis et al. 2020, Ramani et al. 2022, Fry et al. 2023). The current paper addresses this problem by asking:

- 1) What is the correspondence between the outcomes of assessment when using metrics (journal rankings, citations) versus peer-assessment scores?
- 2) Do institutional declarations in support of responsible assessment produce dif-

ferences in submissions and their outcomes?

3 Analytical approach

3.1 Data

We collect public information on the outputs submitted in the REF2021 Unit of Assessment (UoA) 17-Business and Management.² In total, 108 institutions employing 6,995 staff submitted 16,038 outputs including 14,354 journal articles amounting to 90% of submissions. The analysis below focuses only on journal articles because objective metrics tend not to cover other outputs. We thus leave out books and working papers, and we drop unclassified outputs (Linton and Xu 2022). In addition, following selection criteria used in previous studies (e.g, Hole 2017), we eliminate from the data set outputs which were published in atypical B&M journals, the titles of which appear fewer than five times in the dataset. Table 1 shows the number of outputs by institution, with Cardiff the largest (415 journal outputs) and the University of Bolton the smallest (1 output).

Insert table 1 here.

The analysis deploys two proxies of responsible assessment: REF outputs Gross Point Average (*GPA*) and DORA signatories' list. REF GPA for an institution is derived from peer review of individual outputs. Although REF GPA is publicly available, data concerning peer scores for individual outputs is not publicly available. In its absence, we use output GPA and apply an algorithm (e.g, Hole 2017, Linton and Xu 2022) to infer the correspondence with objective indicators.

Further, we use institutional DORA commitments as an additional proxy for responsible assessment. DORA signatories signal their approval for and adoption of

²The results of the 2021 assessment were published in May 2022; see <https://results2021.ref.ac.uk> (accessed 28 January 2023)

assessment that rests on peer review and includes unconventional outputs as markers of academic productivity. Although other declarations are available, DORA is most widely accepted. Publicly available information on DORA signatories marks the adoption of responsible assessment principles by individual higher education institutions.³ Our data set of 108 institutions includes 68 DORA signatories and 40 non-signatories.

Considering objective indicators, we use two types of measures. First, we deploy CABS journal rank that assigns a score of 4,3,2,1, or “unclassified” to a journal.⁴ Second, we collect paper-specific citation information from Scopus and Google Scholar and the variable captures an average citation for each of the outputs.

In addition, our confirmatory analysis controls for the time since publication and the nature of the institution. Specifically, we distinguish two institutional affiliations: the *Russel* group and the “1994” group. We also include an additional marker for a relatively newly established university: *New Universities* (former polytechnics). Table 2 presents the full list of variables and their definitions.

Insert table 2 here.

3.2 Sample description

Table 3 reports summary statistics for the 108 academic departments that make up our dataset. The average institution has a *GPA* of 2.86, ranging from 1.45 to 3.72. The calculated GPA, based on the CABS AJG ranking (*CabsGPA*), is slightly higher at around 3.2. The number of citations and the number of months of the institution’s papers are 69 and 67, respectively. Moreover, on average, the institutions submit around 133 outputs, ranging from 1 to 415.

Insert table 3 here.

³The data on the organisations that have signed DORA are available here: <https://https://sfdora.org/signers/> (accessed 01 March 2023).

⁴The methodology of the CABS Academic Journal Guide 2021 and the actual ranking is available here: <https://charteredabs.org/academic-journal-guide-2021/>.

Next, we show summary statistics for our key variables. We split the sample between *DORA* and non-*DORA* to gauge any differences across the two groups. Means and standard deviations are presented in table 4. In the last column of the table we report p-values for the tests of equality of means across the above-mentioned groups. The breakdown reveals that the average *GPA* is 2.92 for *DORA* members, and 2.76 for their counterparts. The mean difference is statistically significant. The *CabsGPA* is also significantly different across the two groups albeit at a lower level of significance. The remaining variables do not display any significant differences between the two groups.

Insert table 4 here.

The pairwise correlation coefficients in table 5 show the bivariate relationships of the main variables of this study. We observe a positive association between GPA and the implied GPA score. The coefficient is large in magnitude (0.93) and statistically significant at the 1% level. This statistic is consistent with [Linton and Xu \(2022\)](#), who find that the implied GPA score for the Economics unit of submission is positively associated with REF scores, with a pairwise coefficient of 91%.

Insert table 5 here.

In summary, our preliminary statistical analysis highlights three main points. First, the descriptive data on outputs offer a first glimpse of the likely gap between rhetoric and practice, because journal publications prevail. Given that most institutions in our sample are *DORA* signatories, then one would expect a higher incidence of non-journal outputs. However, this is not borne by the data. Second, the tests point to significant differences between *DORA* and non-*DORA* organisations, which indicate that there is a correlation between *DORA* membership and an improved GPA score. In other words, there is considerable cross-sectional variation in REF scores. Finally, there is a strong positive association between implied GPA scores and REF scores. Regression analysis determines whether these bivariate relationships carry over to a multivariate framework, and that is where we now turn our attention.

3.3 The algorithm

In the absence of peer-review scores due to REF confidentiality, we use an algorithm to infer peer-review scores for all outputs submitted by each institution. The algorithm allocates a rank to each submitted output taking into account the journal it was published in and compares the predicted share of outputs at the institution level to the actual shares. Therefore, the algorithm allows us to infer the average quality of the papers published in the journal, as judged by the REF sub-panel, based on the outputs submitted to REF 2021 for UoA17, Business and Management. In line with [Hole \(2017\)](#) and [Linton and Xu \(2022\)](#), we assign an initial rank to all submitted articles. We use these initial values to make predictions about the proportions of 4-star (“world leading”), 3-star (“internationally excellent”), 2-star (“internationally recognised”) and 1-star (“nationally recognised”) submissions for each institution. Next, we calculate the squared difference between the predicted and actual proportions for each category and sum the squared differences, where $i = 1, 2, \dots, N$ refers to a cross-section of units (institutions in this study) and $j = 1, 2, \dots, 4$ denotes journals’ categories. The sum of squared differences (SSD) is weighted by the number of submissions from each institution (N_i) as follows:

$$SSD = \sum_{i=1}^{108} \sum_{j=1}^4 N_i (p_{ij} - \hat{p}_{ij})^2 \quad (1)$$

where p_{ij} is the REF-panel-determined proportion of j -star submissions in each institution, and $\hat{p}_{ij}$ is the predicted proportion:

$$\hat{p}_{ij} = \frac{1}{N_i} \sum_{n=1}^{N_i} I(r_{ni} = j) \quad (2)$$

Further, r_{ni} is the rank assigned to output n submitted by institution i . $I(\cdot)$ is the indicator function equal to one if the expression in the parenthesis is true, and zero otherwise. We sort the journals in random order and run the algorithm in order to minimise SSD. Specifically, we begin with the CABS AJG journal ranking to assign initial journal rankings and then proceed sequentially through the journals, changing

the assigned category if it improves the SSD. We repeat this until the improvement in SSD is small; that is, the algorithm converges when an iteration decreases the SSD by < 0.0001 .⁵ We report the average over a thousand alternative sequences along with the range of outcomes. The outcome is a percentage count of each journal r in each quality category q .

3.4 Baseline model

We construct an indicator to rank departments in national league tables (DeFraja et al. 2019, Linton and Xu 2022). Specifically, grade point average (GPA) is a weighted average of the scores, with the proportion in each category as the weight:

$$GPA_i = \sum_{q=1}^4 p_{ij} \quad (3)$$

We wish to assess the the role of the standard journal rating system in explaining the observed REF GPA using an approach that fully reflects the impact of various factors observed at the institution level embedded in the empirical analyses by Battistin and Ovidi (2022) and Linton and Xu (2022). We consider the following baseline model:

$$GPA_i = \alpha + \beta_1 CabsGPA_i + \beta_2 Citations_i + \beta_3 Age_i + \beta_4 Size_i + \epsilon_i \quad (4)$$

where α is an intercept and GPA_i denotes the dependent variable. The explanatory variables include the AJG CABS journal classification ($CabsGPA$), the average citation of the institution’s papers ($Citations$), the average age (in months) of the institution’s papers (Age) and the total number of submissions ($Size$). ϵ is the error term. The coefficient of interest in equation 4 is β_1 , which measures the impact of the GPA of the institution’s papers implied by the CABS AJG journal classification on the institution’s GPA. Obtaining a positive and significant coefficient for β_1 , would support the notion that implied GPA scores matter for determining research quality

⁵For details on the algorithm see Hole (2017) and Linton and Xu (2022).

of the institutions, as measured by the GPA.

In order to enrich our understanding of our baseline findings, we also explore whether different institution types affect the results. We follow [DeFraja et al. \(2019\)](#) to generate four categories: Russel group, 1994 group, new universities (former polytechnics), and other institutions.

3.5 The role of institutional heterogeneity

Next, we investigate whether variation in institutional output quality affects the relationship between CABS AJG implied GPA and institution GPA. In order to look at this issue, we estimate a variant of equation 4 as follows:

$$GPA_i = \alpha + \beta_1 CabsGPA_i * HighProfile_i + \beta_2 CabsGPA_i + \beta_3 Citations_i + \beta_4 Age_i + \beta_5 Size_i + \epsilon_i \quad (5)$$

where *HighProfile* is a dummy variable equal to 1 for institutions whose publications are above the median of the distribution of publications in journals with a CABS AJG 2021 grade of 4-star, and 0 otherwise. We construct this variable in the spirit of [Banal-Estañol et al. \(2023\)](#), who measure publications in top journals as those recorded in journals with CABS AJG 2018 grades of 3-star or 4-star. The interaction term in β_1 gauges the change in the response to *CabsGPA* for institutions that publish in high-profile journals compared to their counterparts. If the point estimate for the interaction term is positive and significant, we can conclude that the effect of the implied CABS AJG GPA score on institution GPA is more potent for the high-profile journal categories relative to publications in journals with 2-star and 1-star ratings.

Finally, we take into account how *DORA* adoption affects the relationship between the CABS AJG implied GPA and the REF implied GPA. To do so, we estimate the following model:

$$GPA_i = \alpha + \beta_1 CabsGPA_i * DORA_i + \beta_2 CabsGPA_i + \beta_3 DORA_i + \beta_4 Citations_i + \beta_5 Age_i + \beta_6 Size_i + \epsilon_i \quad (6)$$

where *DORA* is a dummy variable equal to 1 for organisations that have signed the DORA declaration, and 0 otherwise. If responsible assessment matters, then we expect β_1 to be positive and significant. On the other hand, if DORA adoption does not play a material role in determining REF implied scores, we expect that the effect of CABS AJG implied ranking is not significantly different from the non-adopters. Therefore, the coefficients associated with β_1 should not be statistically significant.

4 Results

4.1 Metrics and responsible assessment

We show the predictive accuracy of the algorithm using actual AJG CABS journal rankings. Table 6 illustrates the contingency table of the predicted against actual journal ratings. To ease interpretation, we assign journals to 4-star, 3-star, 2-star and 1-star categories based on cut-off points discussed in [Hole \(2017\)](#).⁶ We find that the algorithm correctly classifies about 70% in the 4-star category and the predictive ability declines as we move to 2-star and 1-star categories. In summary, we conclude that categorization of journals is remarkably consistent in 4-star and 3-star outputs when considering standard journal rankings such as the CABS AJG.⁷ Moreover, we produce a comprehensive journal ranking for business and management taking into account data for REF2021. Table A1 in the appendix reports our journal ranking results, which are the percentages of 1,000 runs for which each journal is classified as 4-star, 3-star, 2-star or 1-star. The table reports the journals and their corresponding

⁶The choice of cut-off points is somewhat arbitrary but widely accepted in the literature ([Hole 2017](#), [Linton and Xu 2022](#)).

⁷In the appendix we present a summary of the algorithm’s predictions by AJG CABS rating (see figure 1).

likelihood belonging to a certain category.

Insert table 6 here.

Next, we estimate a basic model of REF GPA determination as shown in equation 4. In column 1 of table 7 we present the results for all universities. In column 2, we control for the main university groups. We report coefficient estimates and t-statistics, with standard errors clustered by university. The general finding is that *CabsGPA* positively and significantly affects the REF implied GPA, both statistically and economically.

Insert table 7 here.

Our key variable of interest is *CabsGPA*, which shows how implied journal rankings affect the institution *GPA*. We find that this variable is the key predictor of *GPA* score even after controlling for other characteristics such as the number of citations or the size of the submission. The finding is also economically important, because the coefficient of 0.643 implies an elasticity of the REF implied score with respect to *CabsGPA*, evaluated at sample means, of 0.72. A 10% increase in the implied CABS AJG GPA score leads therefore to a 7.2% increase in REF implied score. Our results for UoA 17 are consistent with Linton and Xu (2022), who show that in UoA 16 “institution GPA is largely and almost exclusively determined by the SCOB implied GPA.” The results show qualitatively and quantitatively significant effects. In the following columns of table 7, we rerun the same regressions and find that the main results persist even after controlling for other shocks that could affect universities differently. Turning to our university controls, we find that they exhibit the expected signs. Specifically, Russel group universities are positively associated with higher GPA scores and confirm the finding in DeFraja et al. (2019) that they tend to display the strongest performance among other universities in the research evaluations.

4.2 The role of institutional heterogeneity

Having identified a direct relationship between the CABS AJG implied GPA score and institution GPA, we now dig deeper into how *CabsGPA* affects REF implied scores. In all models we interact the dummy variable for high-profile publications with the variable capturing the CABS AJG implied GPA score to gauge the change in output quality. We find the net response under output quality by summing the coefficients.

Insert table 8 here.

According to the results of estimating equation 5, which are in table 8, the effect of the *CabsGPA* on *GPA* is driven primarily by 4-star outputs (i.e quality that is world-leading in terms of originality, significance and rigour), as opposed to other categories. We document that research quality of the institutions has a significantly different response to *CabsGPA* with respect to output characteristics. This result supports Blackburn et al. (2023), who find that the highest correlation between REF scores and journals rated by the CABS AJG is found for the 4-star outputs (78.58%).

Insert table 9 here.

Finally, we consider how being a *DORA* signatory affects the relationship between REF implied GPA and CABS implied GPA, as shown in equation 6. Our key variable of interest is the interaction between *CabsGPA* and the *DORA* dummy variable (*CabsGPA***DORA*). This shows the impact of the DORA on REF GPA. Controlling for a number of university-level characteristics, DORA's incremental effect is insignificant and quantitatively unimportant, as demonstrated by the coefficient on the interaction term in column 1 of table 9. Therefore, although *CabsGPA* remains the main influence on institution GPA, DORA affiliation has no material effect. The effect remains unchanged in columns 2 of the table. The CABS AJG CABS implied rating is persistent across all specifications.

5 Discussion and concluding remarks

Focusing on research outputs, we interrogate the correspondence in evaluation outcomes resulting from two different modes of assessment: metrics and responsible assessment. In a univariate analysis, we find a significant and high correlation between *CabsGPA* and institution GPA (0.91). In a multivariate setting, we estimate the dependence of peer-review score on an implied GPA score derived from the CABS AJG controlling for the size of institutional submission, the type of institution and DORA signatory status. The results show that CABS AJG rank, a metric indicator of performance, remains an important determinant of peer-derived score. Also significant is membership in the Russel group. By contrast, both citation counts and DORA status have an insignificant effect on the outcome of responsible assessment. These findings offer several important insights into the debate on responsible assessment ([Anderson et al. 2021](#), [Curry et al. 2022](#), [Ramani et al. 2022](#)).

A key finding concerns the close correspondence between peer review and CABS AJG rank. The results show that this metric indicator of performance remains an important determinant of a peer-derived score at the institutional level. This relationship is more potent for higher-ranked journals. The evidence aligns with past studies that observe a similar strong relationship between CABS AJG and REF ([Hole 2017](#), [Traag and Watman 2019](#), [Battistin and Ovidi 2022](#), [Linton and Xu 2022](#)) suggesting that in business and management, journal rankings may be good proxies of quality. Our correlations are somewhat higher than in [Blackburn et al. \(2023\)](#) probably due to the fact that the submission size is the confounding variable ([Traag and Watman 2019](#)). That is, a correlation coefficient typically produces a lower score unless one controls for the effects of size. In past studies, size-dependent measures of correlation systematically produced stronger levels of correlation ([Pride and Knoth 2018](#), [Checchi et al. 2021](#)).

Beyond the effect of size, a further explanation concerns the nature of the peer-review process. As [Bertocchi et al. \(2015\)](#) observes, peer review of published work is necessarily “informed.” Unlike the anonymous and double-blind review of manuscripts prior to publication, an informed review exposes authors’ identities and item prop-

erties, such as, which journal published the paper. A close correspondence between journal rank and peer-review outcome may be indicative of an implicit use of metrics in peer review when reviewers unwittingly depend on rankings in peer judgments (Curry et al. 2022). The findings call for further study of informed peer review.

By contrast, both citation counts and DORA status are insignificant and quantitatively unimportant in determining the outcome of responsible assessment. Our evidence corroborates past studies that also find limited correspondence (Mryglod et al. 2015, Battistin and Ovidi 2022). Considering that citations are misaligned with peer review even when controlling for time since publication, we advise caution when using citation or citation-based indexes (e.g., JIF, h-index) as proxies of research output quality. Our results strongly support calls for abandoning these practices (Ramani et al. 2022, Blackburn et al. 2023), implying that in the field of business in management these indicators should not be used.

Several implications present themselves. From a macro perspective of research policy, correspondence between peer review and journal rank supports a key principle of responsible assessment: the context of assessment matters. The responsible assessment movement explicitly maintains that academic fields vary in their prevailing cultures of knowledge production, in the values that underpin quality judgments, and in the methods that may be most appropriate for capturing excellence (e.g., Guthrie et al. 2019, Anderson et al. 2021, Saenen et al. 2021, Curry et al. 2022). Acknowledging heterogeneity of disciplines, the proponents of responsible assessment have unintentionally suppressed these differences by promoting one type of assessment mode (peer review), leading to homogeneity in the tools of assessment. Our findings point to specificity in business & management, perhaps alongside economics (Hole 2017, Linton and Xu 2022), where mixed-modes of assessment, including metrics, may offer close proxies of research output quality. The point here is not to displace peer review with metrics, but rather to consider how metrics can triangulate and sense-check peer-review outcomes (Wilsdon et al. 2015). In that sense, our study reinforces calls for further research on the responsible use of metrics (Curry et al. 2022).

From a policy perspective, the findings point to an important gap in evidence.

Significant progress has been made in understanding national evaluation systems and their effects on research productivity, as well as the nature of scholarship they induce (Hicks 2012, Fry et al. 2023, Banal-Estañol et al. 2023). Much attention has been paid to negative outcomes of the use of metrics (Civera et al. 2020, Thomas et al. 2020). By contrast, a general dearth of evidence concerns responsible assessment and its effects on the outcomes of assessment. The lack of data concerns both the quality judgements (review scores) and their systematic effects on scholarship: how the scores inform the perceptions of academic worth and translate into research decisions and career pathways (Thomas et al. 2020). More research is needed to fill this void and move responsible assessment towards evidence-based policy and practice.

Unfortunately, the progress may have been hampered by the responsible assessment movement itself. In the context of the dominant discourse of responsibility, any reference to metrics seems outdated, reactionary and objectionable. Recent policy developments side decisively with responsible assessment (Curry et al. 2022, The European Commission 2022). A cursory glance at our reference list shows that the balance of arguments tips decisively against metrics. However, an unintended consequence of the dominant discourse is the suppression of divergent views, including research on the viability of mixed models of assessment where metrics complement peer assessment. Few exceptions that show alignment between the two (Traag and Watman 2019, Battistin and Ovidi 2022) do not seem not to gain much prominence beyond the narrow field in which they have been produced. Their implications fail to register in broader research policy literature (Curry et al. 2022). Serious concerns about bias, clarity, transparency and efficiency in use inherent in peer-review are yet to be resolved.

This brings two significant risks. First, in the absence of clear rules, uncertainty prevails, potentially harming the very constituencies (younger researchers, interdisciplinary fields, highly original contributions) that responsible assessment purports to protect (Chapman et al. 2019). Although imperfect and imprecise (Aguinis et al. 2020, Anderson et al. 2021, Ramani et al. 2022), measures such as a journal rank list can reassure authors about how others may judge their work, thereby enhancing certainty, transparency and objectivity (Walker et al. 2019, Battistin and

Ovidi 2022). Second, the absence of evidence fosters detachment between claims and practices, where overt support for responsible assessment coincides with covert continuation of practices that rely on metrics. Responsible assessment thus becomes another facet of virtue signalling or institutional peacocking (Yarrow and Johnston 2023). As McKiernan et al. (2019) shows, this risk is not insignificant.

From that perspective, our finding that DORA makes no difference to the outcome of responsible assessment is open to several interpretations. DORA signature signals institutional commitment to responsible assessment. Similarity in assessment outcomes (REF GPA) between signatories and non-signatories may be a positive sign of responsible assessments: its implementation does not lead to performance erosion. A greater incidence of non-traditional outputs in UoA17 submission, would have been a very encouraging sign. However, the high incidence of journal articles juxtaposed with convergence in outcomes may point to a more sinister problem: a potential discord between adoption (signature) and implementation of responsible assessment. The latter involves non-trivial changes in multiple policies, processes and practices that accompany research assessment at an institutional level. DORA signature may signify adoption but not necessarily implementation. Of course, implementation takes time. Within a field with long-standing and broadly accepted institutions such as journal rankings and a tradition of submitting articles, change will necessarily require time, particularly when institutional funding is at stake (Banal-Estañol et al. 2023). However, there is a strong possibility that simple declarations mask the complexity, scale and difficulty of transformation needed to adjust the research ecosystem for responsible assessment. After all, how does one replace objectivity in a manner that is convenient, transparent, equitable, unbiased, actionable and speedy? Although some implementation examples have emerged (Saenen et al. 2021), the answers to these questions are yet to receive much attention. Further evidence on the practices of implementation of responsible assessment and their outcomes for institutions and individuals is urgently needed.

Our final point concerns the purpose of assessment. Various objects of assessment (items, individuals, teams, academic disciplines, institutions) may be evaluated at different points in time to support a range of decisions. Considering individuals,

for example, the purpose of assessment may include hiring, promotion, tenure, or workload allocation (Chapman et al. 2019, McKiernan et al. 2019). In fact, academic careers rest on a multitude of judgements that involve a plethora of decisions informed by quality assessment (Walker et al. 2019); these may include whether to support time off, protect staff from teaching and administration, grant access to research support or offer funding (Anderson et al. 2021). All these tend to be made in continuing fashion and are based, at least in part, on some notion of the individual’s research excellence. To date, much of the discussion about research assessment concerns macro-level, performance-based assessments of institutions, where the purpose of assessment is evaluation at a national level and the audit is very infrequent (once every 6 or 7 years), also reflecting the costs of the exercise (Hicks 2012, Civera et al. 2020, Banal-Estañol et al. 2023). Peer review is a major element of the cost. Recently published guidelines concerning REF2028 suggests that institutions will be asked to report on their internal peer-review processes as an aspect of their research environment (Scottish Funding Council 2023). This move will promote responsible assessment at a micro level of institutional and disciplinary policies of research management.

Although the goals are aspirational and noble, the consequences seem uncertain for at least two reasons. First, there are practical concerns. Can a peer review effectively inform every decision in research management? Beyond the cost, timeliness and practicality, is it plausible to generate detailed and interoperable data of this nature? Or, is it sensible to deploy vast swarms of academics to the task of reviewing? Secondly, there are concerns about bias, uncertainty and ultimately, the lack of fairness. Every assessment is a power regime that rests on distinct forms of control underlying the judgments of worth (Williams 2020). Objective metrics promote numbers, thereby privileging certain disciplines, and categories of staff and being subject to manipulation. In contrast, responsible assessment privileges expert reviewers and their subjective judgments. As such, it does not remove power from assessment but displaces one set of forces (metrics) with another (individuals). In this context, the absence of evidence concerning its outcomes is troubling.

This paper is not free from limitations. The first limitation concerns unavailabil-

ity of REF peer-review scores and the consequent use of algorithm to estimate the scores based on % GPA within 1-star, 2-star, 3-star, 4-star categories and journal submissions. Although the analysis evidences a close match between the predicted and observed percentages, the absence of actual scores affects the findings. As responsible assessment gains momentum, and more data becomes available, it may be possible to increase the accuracy by the actual scores.

The second limitation concerns disciplinary focus. The study focuses on one discipline (business and management), where journal outputs represent a dominant mode of knowledge productions and where journal rankings are widely used to compare performance. This affects generalisation of findings. Disciplinary traditions lead to differences in assessment outcomes ([Traag and Watman 2019](#)) and the findings generated in the context of B&M (e.g., importance of journal rankings) may not necessarily transfer to other fields. Future studies could increase heterogeneity by including a broader range of subjects and [Traag and Watman \(2019\)](#) or [Banal-Estañol et al. \(2023\)](#) offer an excellent illustration of how to do this.

Future studies may also engage in micro-level analysis contrasting and comparing outcomes of peer assessment with metrics. A long-term evidence of consequences of responsible assessment on careers is urgently needed.

References

- Abramo, G., D'Angelo, C. and Reale, E.: 2019, Peer review vs bibliometrics: Which method better predicts the scholarly impact of publications?, *Scientometrics* **121**.
- Aguinis, H., C.Cummings, Ramani, R. and Cummings, T.: 2020, Reviewing and theorizing the unintended consequences of performance management systems, *Academy of Management Perspectives* **34**(1), 135–154.
- Anderson, V., Elliott, C. and Callahan, J.: 2021, Power, powerlessness, and journal ranking lists: The marginalization of fields of practice, *Academy of Management Learning Education* **20**, 89–107.
- Banal-Estañol, A., Jofre-Bonet, M., Iori, G., Maynou, L., Tumminello, M. and Vassallo, P.: 2023, Performance-based research funding: Evidence from the largest natural experiment worldwide, *Research Policy* **52**, 104780.
- Battistin, E. and Ovidi, M.: 2022, Rising stars: Expert reviews and reputational yardsticks in the research excellence framework, *Economica* **89**, 830–848.
- Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. and Peracchi, F.: 2015, Bibliometric evaluation vs. informed peer review: Evidence from Italy, *Research Policy* **44**(2), 451–466.
- Blackburn, R., Dibb, S. and Tonks, I.: 2023, Business and management studies in the UK's 2021 research excellence framework: Implications for research quality assessment, *forthcoming in British Journal of Management*.
- Bornmann, L. and Marx, W.: 2016, The journal impact factor and alternative metrics: A variety of bibliometric measures has been developed to supplant the impact factor to better assess the impact of individual research papers., *Technical report*, EMBO reports.
- Buttler, D.: 2008, Free journal-ranking tool enters citation market, *Nature* **451**, 6.
- CABS: 2020, Research income for business and management, *Technical report*, Chartered Association of Business Schools.
- Chapman, C., Bicca-Marques, J., Calvignac-Spencer, S., Fan, P., Fashing, P., Gogarten, J., Guo, S., Hemingway, C., Leendertz, F., Li, B. and Matsuda, I.: 2019, Games academics play and their consequences: How authorship, h-index and journal impact factors are shaping the future of academia, *Proceedings of the Royal Society B* pp. 2019–2047.
- Chavarro, D., Tang, P. and Ràfols, I.: 2017, Why researchers publish in non-mainstream journals: Training, knowledge bridging, and gap filling, *Research Policy* **46**(9), 1666–1680.

- Checchi, D., Ciolfi, A., Fraja, G. D., Mazzotta, I. and Verzillo, S.: 2021, Have you read this? An empirical comparison of the British REF peer review and the Italian VQR bibliometric algorithm, *Economica* **88**(352), 1107–1129.
- Civera, A., Lehmann, E., Paleari, S., and Stockinger, S.: 2020, Higher education policy: Why hope for quality when rewarding quantity?, *Research Policy* **49**(8).
- Cooper, C. and Otley, C.: 1998, The 1996 research assessment exercise for business and management, *British Journal of Management* **9**, 73–89.
- Currie, R. and Pandher, G.: 2020, Finance journal rankings: Active scholar assessment revisited, *Journal of Banking and Finance* **111**, 105717.
- Curry, S., Gadd, E. and Wilsdon, J.: 2022, Harnessing the metric tide: Indicators, infrastructures priorities for UK responsible research assessment.
URL: <https://doi.org/10.6084/m9.figshare.21701624.v2>
- DeFraja, G., Facchini, G. and Gathergood, J.: 2019, Academic salaries and public evaluation of university research: Evidence from the UK Research Excellence Framework, *Economic Policy* **34**, 523–583.
- Delegado-Lopez-Cozar, E., Rafols, I. and Abadal, E.: 2021, Letter: A call for radical change in research evaluation in Spain, *Profesional de la informacion* **30**, 300–309.
- Franco-Santos, M. and Otley, D.: 2018, Reviewing and theorizing the unintended consequences of performance management systems, *International Journal of Management Reviews* **20**, 696–730.
- Fry, C., Lynham, J. and S. Tran: 2023, Ranking researchers: Evidence from Indonesia, *Research Policy* **52**(5).
- Guthrie, J., Parker, L., Dumay, J. and Milne, M.: 2019, What counts for quality in interdisciplinary accounting research in the next decade: A critical review and reflection, *Accounting, Auditing and Accountability Journal* **32**, 2–25.
- Haley, U., Page, M., Pitsis, T., Rivas, J. and Yu, K.: 2017, Measuring and achieving scholarly impact: A report from the academy of management's practice theme committee, *Technical report*, Academy of Management.
- Hamermesh, D.: 2018, Citations in economics: Measurement, uses, and impacts, *Journal of Economic Literature* **56**(1), 115–156.
- Hicks, D.: 2012, Performance-based university research funding systems, *Research Policy* **41**(1), 251–261.
- Hole, A.: 2017, Ranking economics journals using data from a national research evaluation exercise, *Oxford Bulletin of Economics and Statistics* **79**(5), 621–636.

- Kellard, N. and Śliwa, M.: 2016, Business and management impact assessment in research excellence framework 2014: Analysis and reflection, *British Journal of Management* **27**, 693–711.
- Linton, O. and Xu, E.: 2022, Auditing the auditors: An evaluation of the REF2021 output results, *Technical report*, Cambridge Working Papers in Economics.
- McKiernan, E., Schimanski, L., Nieves, C. M., Matthias, L., Niles, M. and Alperin, J.: 2019, Use of the journal impact factor in academic review, promotion, and tenure evaluations, *eLife* **8**.
- Mingers, J. and Willmott, H.: 2013, Taylorizing business school research: On the ‘one best way’ performative effects of journal ranking lists, *Human Relations* **66**(8), 1051–1073.
- Mryglod, O., Kenna, R., Holovatch, Y. and Berche, B.: 2015, Predicting results of the research excellence framework using departmental h-index, *Scientometrics* **102**, 2165–2180.
- Osterloh, M. and Frey, B.: 2020, How to avoid borrowed plumes in academia, *Research Policy* **49**(1).
- Pidd, M. and Broadbent, J.: 2015, Business and management studies in the 2014 research excellence framework, *British Journal of Management* **26**, 569–581.
- Pride, D. and Knoth, P.: 2018, Peer review and citation data in predicting university rankings, a large-scale analysis, *Technical report*.
URL: <http://arxiv.org/abs/1805.08529>*arXiv: 1805.08529*
- Ramani, R., Aguinis, H. and Coyle-Shapiro, J.: 2022, Defining, measuring, and rewarding scholarly impact: Mind the level of analysis, *Academy of Management Learning and Education* **21**, 470–486.
- Ryazanova, O., McNamara, P. and Aguinis, H.: 2017, Research performance as a quality signal in international labor markets: Visibility of business schools worldwide through a global research performance system, *Journal of World Business* **52**(6), 831–841.
- Saenen, B., Hatch, A., Curry, S., Proudman, V. and Lakoduk, A.: 2021, Reimagining academic career assessment: Stories of innovation and change, *Technical report*, European University Association.
URL: <https://www.eua.eu/resources/publications/952:reimagining-academic-career-assessment-stories-of-innovation-and-change.html>.
- Salandra, R., Salter, A. and Walker, J.: 2022, Are academics willing to forgo citations to publish in high-status journals? Examining preferences for 4-star and 4-rated journal publication among UK business and management academics, *British Journal of Management* **33**(3), 1254–1270.

- Scottish Funding Council: 2023, *Research Excellence Framework 2028: Initial decisions and issues for further consultation*.
URL: file:///C:/Users/st120c/Downloads/REF_2028_initial_decisions_and_issues_for_further_consultation.pdf
- Snyder, H. and Bonzi, S.: 1998, Patterns of self-citation across disciplines (1980-1989), *Journal of Information Science* **32**, 431–435.
- Taylor, J.: 1994, Measuring research performance in business and management studies in the United Kingdom: The 1992 Research Assessment Exercise, *British Journal of Management* **5**(4), 275–288.
- The European Commission: 2022, *Towards a reform of the research assessment system*, The European Commission, Directorate-General for Research and Innovation.
URL: <https://data.europa.eu/doi/10.2777/707440>
- Thomas, D., Nedeva, M., Tirado, M. and Jacob, M.: 2020, Changing research on research evaluation: A critical literature review to revisit the agenda?, *Research Evaluation* **29**(3), 275–288.
- Traag, A. and Watman, L.: 2019, Systematic analysis of agreement between metrics and peer review in the UK REF, *Palgrave Communications* **5**(29).
- Vogel, R., Hattke, F. and Petersen, J.: 2017, Journal rankings in management and business studies: What rules do we play by?, *Research Policy* **46**(10), 1701–1722.
- Walker, J., Fenton, E., Salter, A. and Salandra, R.: 2019, What influences business academics’ use of the association of business schools ABS list? Evidence from a survey of UK academics, *British Journal of Management* **30**(3), 730–747.
- Williams, K.: 2020, Playing the fields: Theorizing research impact and its assessment, *Research Evaluation* **29**(2), 191–202.
- Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R. and Kerridge, S.: 2015, The metric tide. Report of the independent review of the role of metrics in research assessment and management.
- Yarrow, E. and Johnston, K.: 2023, Athena swan: “Institutional peacocking” in the neoliberal university, *Gender, Work & Organization* **30**, 757–772.

Table 1: Number of journal outputs returned by institutions

The table reports the total number of submissions by institutions in an descending order.

	Institution name	Number of outputs
1	Cardiff University / Prifysgol Caerdydd	415
2	The University of Manchester	375
3	The University of Lancaster	369
4	The University of Warwick	336
5	The University of Nottingham	329
6	The University of Leeds	325
7	Loughborough University	320
8	The University of Birmingham	319
9	The University of Sheffield	318
10	University of Glasgow	308
11	University of Northumbria at Newcastle	302
12	City, University of London	299
13	University of Durham	284
14	The University of Essex	277
15	King's College London	261
16	London Business School	255
17	University of Strathclyde	252
18	The University of Liverpool	250
19	Aston University	247
20	University of Edinburgh	242
21	The London School of Economics and Political Science	237
22	University of Newcastle upon Tyne	236
23	University of Southampton	221
24	University of Portsmouth	215
25	University of Sussex	215
26	The University of Bath	207
27	The University of Reading	201
28	The University of Leicester	196
29	Imperial College of Science, Technology and Medicine	195
30	The University of Huddersfield	195
31	Middlesex University	186
32	De Montfort University	182
33	University of Oxford	179
34	Brunel University London	174

35	University of Greenwich	174
36	Coventry University	170
37	The University of Kent	170
38	University of Exeter	169
39	Manchester Metropolitan University	156
40	University of Bristol	155
41	The Open University	147
42	Royal Holloway and Bedford New College	144
43	Queen's University of Belfast	141
44	The University of East Anglia	141
45	University of York	139
46	Nottingham Trent University	137
47	University of Cambridge	135
48	The University of Stirling	131
49	University of Aberdeen	130
50	Bournemouth University	127
51	Queen Mary University of London	125
52	The University of Surrey	125
53	Birkbeck College	121
54	Swansea University / Prifysgol Abertawe	113
55	The University of Westminster	106
56	Kingston University	105
57	Oxford Brookes University	105
58	Cranfield University	96
59	University of the West of England, Bristol	88
60	The University of Hull	84
61	The University of Bradford	83
62	University College London	81
63	University of St Andrews	80
64	Heriot-Watt University	77
65	University of Ulster	77
66	University of Derby	75
67	University of Plymouth	72
68	University of Lincoln	68
69	Liverpool John Moores University	66
70	Leeds Beckett University	60
71	University of Dundee	56
72	University of Hertfordshire	54
73	University of the West of Scotland	54
74	Birmingham City University	53

75	Sheffield Hallam University	53
76	University of Brighton	53
77	Edinburgh Napier University	50
78	Glasgow Caledonian University	49
79	Goldsmiths' College	48
80	Bangor University / Prifysgol Bangor	47
81	Cardiff Metropolitan University / Prifysgol Metropolitan Caerdydd	44
82	University of South Wales / Prifysgol De Cymru	44
83	Roehampton University	43
84	University of Keele	42
85	School of Oriental and African Studies	41
86	The University of Northampton	41
87	Edge Hill University	40
88	Anglia Ruskin University Higher Education Corporation	39
89	University of Central Lancashire	36
90	University of Wolverhampton	36
91	Aberystwyth University / Prifysgol Aberystwyth	32
92	London South Bank University	29
93	University of Chester	27
94	University of East London	26
95	The University of Salford	24
96	York St John University	24
97	University of Worcester	23
98	London Metropolitan University	21
99	Staffordshire University	21
100	University of Bedfordshire	19
101	University of Gloucestershire	18
102	University of Winchester	17
103	Robert Gordon University	15
104	Bath Spa University	13
105	The University of Cumbria	11
106	Queen Margaret University, Edinburgh	9
107	The University of Chichester	6
108	The University of Bolton	1

Table 2: Variable definitions and sources

<i>A. Dependent variable</i>		
GPA	Calculated by multiplying the percentage of research in each group by its rating, adding them all together and dividing by 100	Authors' calculations
<i>B. Explanatory variables</i>		
CabsGPA	The GPA based on CABS 2021 AJG ranking	CABS Academic Journal Guide
Citations	The average citation of the institution's papers	Scopus and Google Scholar
Age	The average age (in months) of the institution's papers	REF
Size	The total number of submissions	REF
DORA	A dummy variable equal to 1 for organisations that have signed the DORA declaration	DORA
HighProfile	A dummy variable equal to 1 for institutions whose publications are above the median of the distribution of 4-star publications according to CABS AJG 2021.	Authors' calculations
4-star	Quality that is world-leading in terms of originality, significance and rigour	REF
3-star	Quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence	REF
2-star	Quality that is recognised internationally in terms of originality, significance and rigour	REF
1-star	Quality that is recognised nationally in terms of originality, significance and rigour	REF
Unclassified	Quality that falls below the standard of nationally recognised work. Or work which does not meet the published definition of research for the purposes of this assessment	REF

Russel	A dummy variable equal to 1 for the Russell Group's 24 members	https://russellgroup.ac.uk/about/our-universities
Ninetyfour	A dummy variable equal to 1 for 17 universities not invited to join the Russel Group	DeFraja et al. (2019)
New	A dummy variable equal to 1 for former “polytechnic” universities	DeFraja et al. (2019)

Table 3: Descriptive statistics

The table reports summary statistics of the REF outputs for the business and management sub-panel.

Variable	N	Mean	SD	Min	Max
GPA	108	2.86	0.40	1.45	3.72
CabsGPA	108	3.19	0.52	1	3.95
Age	108	67.46	2.82	58.57	73.68
Citations	108	69.75	29.07	5.77	162.47
Size	108	132.90	103.27	1	415

Table 4: Descriptive statistics for different groups

The table reports means and standard deviations in parentheses for DORA and non-DORA signatories. The last column reports the p-value for the test of equality of means between the DORA and non-DORA groups.

Variable	DORA	non-DORA	Difference (p-value)
GPA	2.92 (0.34)	2.76 (0.47)	0.04
CabsGPA	3.25 (0.42)	3.05 (0.63)	0.07
Age	67.35 (2.68)	67.64 (3.08)	0.60
Citations	70.62 (27.80)	68.29 (31.44)	0.69
Size	161.57 (111.94)	126.25 (101.55)	0.10
N	68	40	

Table 5: Correlation matrix

The table reports pairwise correlations of the REF outputs for the business and management sub-panel. * denotes statistical significance at the 1% level.

Variable	GPA	CabsGPA	Age	Citations	Size
GPA	1.00				
CabsGPA	0.93*	1.00			
Age	0.08	0.11	1.00		
Citations	0.67*	0.62*	0.02	1.00	
Size	0.40*	0.40*	0.16	0.24	1.00

Table 6: AJG CABS journal ratings against the predicted ranks

The table reports the predictions of the algorithm. The leftmost column shows actual AJG CABS ratings, while the righthand side columns show the predictions of the algorithm. The green colour defines the predictions that fall within the boundaries of the original AJG CABS rating, amber those that are classified in the nearest neighbour class, and red those where the algorithm misclassified them by 2 rating points.

AJG CABS rating	1-star	2-star	3-star	4-star
Possible 2-star			2	
Probable 2-star		2	2	1
2-star	3	22	22	1
Possible 3-star	1	10	18	4
Probable 3-star		5	27	5
3-star	1	10	71	22
Possible 4-star		3	14	17
Probable 4-star		2	5	24
4-star		2	19	40

Table 7: Baseline results

The table reports the estimates of an OLS model. The dependent variable is GPA score. Robust t-statistics are reported in the parentheses. *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level.

	(1)	(2)
CabsGPA	0.643*** (15.09)	0.570*** (13.45)
Age	-0.003 (-0.53)	-0.004 (-0.87)
Citations	0.002*** (3.10)	0.002*** (2.96)
Size	0.000 (0.93)	0.000 (0.18)
Russel		0.112*** (2.95)
Ninetyfour		0.042 (1.10)
New		-0.016 (-0.22)
Constant	0.837** (2.32)	1.199*** (3.35)
Observations	108	108
R-squared	0.884	0.905

Table 8: The role of high-profile outputs

The table reports the estimates of an OLS model. The dependent variable is GPA score. *HighProfile* is a dummy variable that equals 1 for institutions whose publications are above the median of the distribution of publications in journals with a CABS AJG 2021 grade of 4-star. Robust t-statistics are reported in the parentheses. *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level.

	(1)	(2)
CabsGPA*HighProfile	0.131** (1.96)	0.089* (1.70)
CabsGPA	0.581*** (12.77)	0.535*** (13.78)
HighProfile	-.0319 (-1.24)	-0.199 (-0.76)
Age	-0.004 (-0.75)	-0.005 (-0.97)
Citations	0.002*** (2.85)	0.002*** (2.89)
Size	0.000 (0.83)	0.000 (1.08)
Russel		0.113** (2.69)
Ninetyfour		0.039 (0.75)
New		0.009 (0.12)
Constant	1.075*** (2.96)	1.310*** (3.78)
Observations	108	108
R-squared	0.897	0.914

Table 9: The differential effect of DORA

The table reports the estimates of an OLS model. The dependent variable is GPA score. *DORA* is a dummy variable equal to 1 for organisations that have signed the DORA declaration, and 0 otherwise. Robust t-statistics are reported in the parentheses. *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level.

	(1)	(2)
CabsGPA*DORA	0.013 (0.20)	-0.032 (-0.54)
CabsGPA	0.631*** (13.03)	0.581*** (13.68)
DORA	.009 (-0.04)	0.098 (0.51)
Age	-0.002 (-0.48)	-0.003 (-0.78)
Citations	0.002*** (3.17)	0.002*** (2.99)
Size	0.000 (0.80)	0.000 (0.16)
Russel		0.116*** (3.06)
Ninetyfour		0.045 (1.13)
New		-0.020 (-0.27)
Constant	0.832** (2.32)	1.137*** (3.26)
Observations	108	108
R-squared	0.885	0.905