Measuring research quality in a more inclusive way: evidence from the UK Research Excellence Framework*

Erin Hengel
Brunel University

Almudena Sevilla
LSE

Sarah Smith
University of Bristol

November 2023

Abstract. Evidence suggests that common metrics of research quality—*e.g.*, journal publications and citations—are systematically biased against certain groups. But does relying solely on them to evaluate quality lead to lower diversity in academia? In this paper, we start to answer this question by analysing data from the UK's nationwide research assessment exercise, the Research Excellence Framework. We find that narrowly focussed output-based measures of departmental research quality do indeed negatively correlate with the diversity of departmental staff, while measures of research impact and of the quality of the research environment correlate positively. An aggregate measure that incorporates all three components is therefore likely to better promote staff diversity compared to more narrowly defined output-focused measures. More generally, our results suggest that comprehensive definitions of research quality may be more effective at promoting diversity in academia compared to narrower measures. We further argue that funding decisions informed by broader measures result in more efficient resource allocations across the higher education sector.

JEL: I23, J16

Keywords: Performance measurement; Research quality; Diversity.

*We are grateful to the editor and three referees for valuable comments. Neither the Higher Education Statistic Agency (HESA) nor HESA Services Limited are responsible for any inferences or conclusions derived from the data or other information supplied by HES Limited or HESA Services Limited.

1 Introduction

Common metrics of research quality and productivity—such as citations and publication counts—play a crucial role in academic job-market decisions (*e.g.*, tenure and promotion). They also increasingly inform which projects and individuals are awarded competitive, non-recurrent grant funding. Soon, metrics may even drive the allocation of recurrent research funding across institutions, too (see, *e.g.*, MacIntosh 2021).

While most metrics are easy to compute and readily available, they only proxy for the true quality of a project and the true performance of researchers. Consequently, they are measured with error—and according to many studies, this error correlates with researcher characteristics. For example, Card, *et al.* (2020), Hengel (2022) and Hengel and Moon (2023) show that female economists are held to higher acceptance standards at top economics journals compared to male economists. Men are also better connected to their academic networks (Ductor, Goyal and Prummer 2023) which probably facilitates their outcomes in peer review (for evidence, see, *e.g.*, Colussi 2018). Meanwhile, Ferber (1986; 1988), Dion, Sumner and Mitchell (2018) and Koffi (2021) show that journal articles written by men are less likely to cite women than they are to cite other men, and Larivière, *et al.* (2013) find that articles with a first or last female author are cited less than observably equivalent articles with male authors in the same positions.

If certain groups publish less, and are less well-cited compared to other groups, then relying solely on metrics based on publications or citations may advantage the latter at the expense of the former. In contrast, broader measures of research quality may impose less of a disadvantage on under-represented groups. To date, however, there is little evidence on the practical use of such measures or on whether they are better at increasing the diversity of academic staff compared to narrower measures. This is the contribution of our paper.

In particular, we provide new evidence on the relationship between the multiple, expansive measures of research quality in the UK's Research Excellence Framework (REF) and the diversity of academic staff. Every 6–7 years, the quality of the research produced by academic departments at UK universities is evaluated in a nation-wide exercise known as the REF. The REF's scope of assessment

is defined broadly to include departments' "outputs" (academic publications), "impact" (case studies documenting how research has changed policy and practice) and "environment" (narrative accounts of how departments "support the production of excellent research"). Government research funding is then allocated to universities according to a weighted average of their departments' performance in each of these three elements.

Combining departmental-level evaluation data from the 2014 REF with data on departments' academic staff diversity from the UK's Higher Education Statistical Agency, we ask whether the broad scope of the REF—and in particular the inclusion of impact and environment in its definition of research quality—is more likely to promote diversity among academic staff compared to an alternative, narrower definition that considers only outputs.[1]

We find that the output score negatively correlates with our measure of diversity on both counts: departments that scored higher for their outputs were not only less diverse at the time of REF submission but were also less likely to increase their diversity in subsequent years. By contrast, the impact score positively correlates with our measure of diversity at the time of submission, suggesting that more diverse departments produce better impact. Although the environment score negatively correlates with staff member diversity at the time of REF submission, it positively correlates with departments' subsequent progress on diversity, in line with its more forward-looking nature. This evidence highlights that measures of research quality correlate with diversity but the direction of that correlation depends on how quality is defined. Our findings also suggests that comprehensive measures of quality could mitigate distortions caused by individual, narrowly-defined metrics.

Our study is related to a long-standing literature in economics and management on the challenge of rewarding performance in the face of multiple and competing objectives (Holmstrom and Milgrom 1991; Kerr 1975). In particular, our evidence highlights the tension of "rewarding A, while hoping for B" (Kerr 1975) in the context of measuring and incentivising research quality in higher education. Our

[1] Our measure of diversity is the share of staff members in a department who are not white men. Although this measure does not capture all dimensions of diversity that matter, it does measure the presence of historically under-represented groups. It is also easily measurable on a consistent basis over time and across institutions using existing administrative data on the population of academics.

evidence on the relationship between the scope of research quality measurement and diversity can contribute to the discussion in several countries—particularly those that have national research assessment processes—about the use of different measures for evaluating research quality (see, *e.g.*, Bishop 2021). We also contribute to an ongoing debate about promoting diversity in higher education (see, *e.g.*, Gamage and Sevilla 2019, Gamage, Sevilla and Smith 2020, Lundberg and Stearns 2019, Bateman, *et al.* 2021) by providing new evidence on how the choice of performance metrics impacts the diversity of academic staff.

The paper is organised as follows. The next section provides more detail on the UK REF process. Section 3 explains our methodology, while Section 4 presents the results. In Section 5, we discuss the implications of our findings.

2 The REF 2014

Since 1986, research in UK higher education institutions (HEIs) has been subject to a thorough, national assessment process known originally as the Research Assessment Exercise (RAE) and, since 2014, as the Research Excellence Framework (REF).[2] The results of the process—which takes place (roughly) every six–seven years—are primarily used to allocate ~£2 billion per year of central government research funding across universities, but they are also included in various league tables (for example, the university rankings produced by both the *Complete University Guide* and the *Guardian University Guide* incorporate REF scores) and promoted by individual HEIs in order to attract staff and students. Thus, REF outcomes directly and indirectly determine how resources are allocated between institutions and have had a profound impact on universities' research investment strategies and hiring and promotion decisions (De Fraja, Facchini and Gathergood 2019).

REF submissions are made at the level of Units of Assessment (UoAs) which correspond broadly to academic departments.[3] In REF 2014, assessment of research quality was carried out by 36 subject sub-panels, consisting of academic and external assessors. The sub-panels were organised into four

[2] Similar assessments have been introduced in the Netherlands (Observatory of Science and Technology), Italy (Triennial Research Evaluation), Australia (Excellence in Research for Australia) and New Zealand (Performance-based Research Fund).

[3] Universities can—and do—decide which sub-panel to submit particular staff to. For example, economics staff can be submitted either to the economics and econometrics sub-panel or the business and management sub-panel.

main panels covering medicine, health and life sciences (panel A), physical sciences, engineering and maths (Panel B), social sciences (Panel C), and arts and humanities (Panel D).

In 2014, each UoA submitted the following three elements to the REF:

- A curated collection of its staff members research *outputs* (*e.g.*, books and academic articles).
- A limited number of *impact* case studies documenting the wider social impact of staff members' research (*e.g.*, the change in policy and practice that their research achieved).
- A narrative account of the UoA's research *environment*, covering the following four dimensions: (i) the coherence of the UoA's research agenda; (ii) resources, facilities and infrastructure; (iii) external engagement; and (iv) "people", which included the promotion of equality and diversity among the UoA's staff members.[4]

While outputs and impact evaluated the quality of departments' research, the REF 2014's environment score instead measured their strategies, processes and culture for supporting that research. It was also the only component that was explicitly forward-looking, in that it intended to identify departments that could sustain a positive research environment going forward.[5]

REF 2014's sub-panel members read and assessed the quality of every submitted UoA's outputs, impact and environment without making formal use of metrics such as citations and journal rankings.[6] Instead, quality was assessed subjectively against the following broad criteria: outputs were judged on their "originality, significance and rigour"; impact case studies were judged for their "reach and significance"; and research environments were judged for their "vitality and significance". Against these criteria, research quality was graded from 4* (highest) to 1* (lowest) according to the broad standards summarised in Table 1. Each sub-panel and main panel additionally conducted benchmarking exercises to agree on more specific standards for each grade. Many sub-panels also double-scored

[4] Alongside the environment statement, UoAs were required to provide information on grant income and numbers of PhD students. However, these data were contextualised within the environment statement itself.

[5] See Appendix C for further discussion and insights on what, precisely, the environment score in the REF 2014 was measuring.

[6] Several sub-panels (*e.g.*, clinical medicine, physics and economics and econometrics) had access to citation data but these data were used to supplement rather than replace peer assessment.

submissions to improve the consistency of assessment, and impact case studies were also evaluated by external assessors working outside academia.

At the end of the exercise, the shares of each UoA's outputs, impact and environment that were graded 4*, 3*, *etc.* were published on the REF 2014 website. Each UoA also received an overall grade profile that was a weighted sum of the grades given to each of the three elements.[7] This final grade profile was used to determine the allocation of government funding, with zero weight given to 1* and 2* research and the highest weight given to 4* research (De Fraja, Facchini and Gathergood 2019).

3 Methodology

3.1 Conceptual framework

In this section, we present a very simple framework to clarify our thinking on distortions introduced when imperfectly measuring research quality and to motivate the empirical approach that follows. Assume that the quality of research in department d is determined by $f(T_d)$, where T_d is the talent of department d. Assume also that the government would like to distribute funds to different departments according to $f(T_d)$, but this is unobserved. Instead, only a proxy of it, $g(T_d)$, is observed. $g(T_d)$ is assumed to positively correlate with $f(T_d)$—*i.e.*, departments with higher $f(T_d)$ usually also have higher $g(T_d)$—but is also systematically biased in favour of people from certain groups. For example, suppose g mapped the number of citations accruing to T_d. Given evidence of bias in the decision to cite (Ferber 1986; Ferber 1988; Dion, Sumner and Mitchell 2018; Larivière, *et al.* 2013; Koffi 2021), g would likely underestimate the quality of female talent in department d and over-estimate the quality of its male talent.

Suppose department d hired staff to maximise the quality of the research it produced. It would therefore optimally choose T_d^* to maximise $f(T_d)$, *i.e.*,

$$T_d^* = \text{argmax}_{T_d} f(T_d).$$

[7] The weights given to outputs, impact and environment were, respectively, 65%, 20% and 15%.

But if department d only cared about research quality to the extent that it leads to more funding, then it would instead optimally choose T_d' to maximise $g(T_d)$, *i.e.*,

$$T_d' = \text{argmax}_{T_d}\, g(T_d).$$

T_d' maximises $g(T_d)$; it does not maximise $f(T_d)$. Thus,

$$f(T_d') \leq f(T_d^*), \tag{1}$$

or in other words, research quality is (weakly) lower when the proxy of quality is used to allocate funding than it would be if funding were allocated according to actual quality.

Equation (1) highlights an important implication of using metrics as a measure of research quality: unless the proxy perfectly captures the underlying construct of interest, it will result in a misallocation of money within the sector—*e.g.*, money will go to institutions that produce the most-highly cited publications which, in a world where citations are biased in favour of a particular group, are unlikely to be the most diverse institutions.

Furthermore, by rewarding $g(T_d)$ instead of $f(T_d)$, departments are incentivised to reduce diversity of T_d. Since T_d' was chosen to maximise $g(T_d)$ and g is systematically biased in favour of people from certain groups, then T_d' will likely be less diverse than T_d^*. This is because departments choose to hire a pool of talent T_d' that is disproportionately composed of group members that are advantaged by g. An implication of this is that departments that perform well on the basis of a biased measure of research quality will tend to be less diverse than those that perform less well.

One way to move T_d' closer to T_d^* would be to augment g with a complementary measure that positively correlates with diversity. In principle, the impact and environment measures in the REF 2014 may have fulfilled this role—indeed, the people element of the environment score explicitly included the promotion of equality and diversity, and panel members were required to consider this as part of their assessment. This insight motivates our empirical analysis, described in the next section, which examines the relationship between the different measures of quality in the REF and (a measure of) diversity.

3.2 Empirical approach

To determine the extent to which the different measures of departmental research quality in REF 2014 correlate with a measure of departments' academic staff diversity, we estimate the following equation using OLS:

$$D_{dsi} = \beta_0 + \beta_1 \text{Outputs}_{dsi} + \beta_2 \text{Impact}_{dsi} + \beta_3 \text{Environment}_{dsi} + \phi_i + \varphi_s + \varepsilon_{dsi}. \quad (2)$$

The dependent variable, D_{dsi}, is a measure of the diversity of academic staff in department d in higher education institution i allocated to sub-panel s. We regress this on a weighted sum of the shares of outputs, impact and environment that were rated 3* and 4*, *i.e.*, $4 \times \text{percentage } 4^* + 3 \times \text{percentage } 3^*$.[8] In order to remove systematic variation across HEIs and subjects, Equation (2) additionally controls for fixed effects for institutions (ϕ_i) and sub-panels (φ_s).

In Equation (3) we investigate how REF scores correlate with (future) improvements in diversity by regressing department d's change in diversity over the five years following its REF 2014 submission, ΔD_i, on its output, impact and environment scores:

$$\Delta D_{dsi} = \beta_0 + \beta_1 \text{Outputs}_{dsi} + \beta_2 \text{Impact}_{dsi} + \beta_3 \text{Environment}_{dsi} + \varphi_s + \phi_i + \varepsilon_{dsi}. \quad (3)$$

β_1, β_2 and β_3 are our coefficients of interest in both Equations (2) and (3). Negative coefficients indicate that higher measured research quality is associated with lower departmental diversity at the time of UoA submission (Equation (2)) and a decline in diversity in the years following submission (Equation (3)). Assuming that research quality does not systematically vary (for whatever reason) by researcher characteristics, these estimates provide suggestive evidence that REF 2014's proxies of research quality under-estimate the quality of research by under-represented groups, lead to less diversity in the sector and may even lower the true quality of the research it produces.

To estimate Equations (2) and (3), we measure D_{dsi} as the percentage of academic staff in a department who were non-white and/or female in 2013, the year UoAs made their REF submissions. To capture ΔD_{dsi}, we subtract D_{dsi} in 2013 from D_{dsi} in 2018. Although these proxies of diversity are

[8] This was also the formula used to determine funding allocations after the REF 2014 concluded (De Fraja, Facchini and Gathergood 2019).

by no means comprehensive, they do capture important dimensions of under-representation (gender, for example) that have been shown to matter in terms of publications and citations.

Our measure of D_{dsi} comes from the academic staff census data collected by the Higher Education Statistical Agency (HESA). HESA staff data are reported by universities and cover all individuals on a contract of employment with a publicly funded higher education provider in the UK during a given academic year (1 August to 31 July). To identify academic staff, we restrict our data to non-administrative staff members on academic contracts who are engaged in teaching and/or research. We additionally exclude senior management (including heads of school and function heads) and staff members employed by professional service departments (*e.g.*, central administration, staff and student facilities, and catering).

We merge our HESA data on departments' demographic profiles with publicly available information on departments' REF 2014 performance using the mapping described in Appendix A.[9] After merging, our final dataset covers 1,736 academic departments across 36 different disciplines at 151 UK higher education institutions.[10] Basic summary statistics are provided in Appendix A.

4 Results

4.1 Main results

Our main regression results are presented in Table 2. Panel A displays results from estimating Equation (2) using the 2013 percentage of non-white-male staff members as the dependent variable. Panel B shows results from estimating Equation (3) using the post-REF improvement in diversity (2013–2018) as the dependent variable. Column (1) includes no controls; columns (2) and (3) add, respectively, sub-panel and HEI fixed effects.

[9] In 13 departments, REF environment, impact and output results were not published because the number of submitted staff was three or fewer.

[10] For several departments, HESA data on staff demographics in either 2013 or 2018 were unavailable. We also exclude 65 observations corresponding to multiple submissions from the same department. (For example, University of Chester made two environment submissions to the "Geography, environmental studies and archaeology" UoA, one for "Geography and development studies" and another for "Archaeology". Both observations are excluded from the analysis.) As a result, the final main estimation samples shown in Table 2 and Table 3 include only 1,635 observations when the dependent variable is D_i and 1,598 when the dependent variable is ΔD_i.

Looking first at panel A, higher output scores are associated with lower shares of non-white-male staff members at the time of REF submission (*i.e.*, $\beta_1 < 0$), indicating that departments with higher scoring outputs were generally less diverse in 2013. Adding sub-panel fixed effects, the coefficient remains negative but becomes smaller and statistically insignificant, indicating that there are subjects that have systematically high output scores/low levels of diversity. By contrast, a department's impact score positively correlates with our measure of diversity (*i.e.*, $\beta_2 > 0$), and this is robust to including sub-panel and institution fixed effects. This indicates that more diverse departments produce better impact. Perhaps surprisingly, the coefficient on environment score is also negative (*i.e.*, $\beta_3 < 0$), including sub-panel and institution fixed effects. Although panel members are asked to consider measures to promote diversity and equality as part of the environment evaluation, the relationship between the environment score and our measure of diversity at the time of submission is negative.

Estimating Equation (3)—*i.e.*, using the change in non-white-male staff (2013–2018) as the dependent variable (ΔD_{asi})—we see that departments that scored highly on outputs experienced a decline in diversity (an increase in the share of white men) in the years following the REF (*i.e.*, $\beta_1 < 0$). Adding sub-panel and institution fixed effects does not change the magnitude of the coefficient but increases the standard error such that the estimated effect is no longer statistically significant (column (3)). The correlation between the impact score and post-REF change in diversity is very close to zero.

By contrast, there is a positive relationship between the environment score and future diversity improvements (*i.e.*, $\beta_3 > 0$). This suggests that the environment score—the only component of the REF that is forward looking—may capture aspects of departmental strategies, processes and culture that are important for promoting diversity. The size of the implied effect is quite large. Across the 36 sub-panels, the (average) inter-quartile range of (our transformed) environment score is around 200, which would equate to a 1.80 percentage point increase in the share of non-white-male staff (based on the coefficient of 0.009) compared to a mean increase over the period of 3.93 percentage points.

4.2 Further analysis

Is it possible to say whether there are particular aspects of departmental processes and culture that are associated with improvements in our measure of diversity? This insight could help to inform

strategies to increase representation from under-represented groups. It could also form the basis for designing alternative—and more targeted—approaches that reward specific drivers of increased representation, rather than the broad and subjective environment measure. We consider two possible candidates—Athena SWAN accreditation (a UK-wide initiative aimed at improving gender equality in higher education) and the quality of management practices at the department level.

4.2.1 Athena Swan

The Athena SWAN Charter was launched in 2005 to advance the careers of women initially in STEM fields but later across all academic fields. Athena SWAN awards are given—at bronze, silver or gold level—to universities and, separately, to individual departments that can demonstrate a commitment to gender equality. The submission process, which typically takes a couple of years, requires a comprehensive audit of gender equality, and a set of concrete proposals for change (see Gamage, Sevilla and Smith 2020). Many environment statements refer to Athena SWAN—either because the departments already have an award or because they are in the process of applying for one.

We re-run Equations (2) and (3), additionally including a binary indicator ("Athena") which takes the value 1 if the departmental statement includes a mention of Athena SWAN. The results are reported in columns (1) and (3) of Table 3. We find that departments that refer to Athena SWAN tend to have lower diversity at the time of REF submission. This seemingly counter-intuitive result may suggest that the decision to apply for Athena SWAN is a response to low diversity; however, the results in column (3) show that departments with a mention of Athena SWAN also make more subsequent improvement in diversity in the years following REF submission. Nevertheless, neither of these correlations is statistically significant.

The coefficients on the environment score in columns (1) and (3) of Table 3 are similar to those shown in Table 2. Thus, controlling for whether a department's environment statement mentions Athena SWAN has little impact on the magnitude or significance of the correlation between a department's environment score and the diversity of its staff. This suggests that the environment score's relationship with promoting equality and diversity captures more than just whether a department has (or is applying for) an Athena SWAN award. (For further analysis and discussion of what the environment measures, see Appendix C.)

4.2.2 Management practices

The way a department is run may be an important factor in determining its strategies, processes and culture, *i.e.*, the environment score may reflect the quality of management of a department. There is a body of literature in economics on measuring management quality in different organisations, showing that the quality of an organisation's (measured) management practices in relation to operations and people correlates positively with its overall performance (see Bloom, *et al.* 2014). This relationship holds for many different sectors, including UK higher education (McCormack, Propper and Smith 2014). There is also evidence that better managed organisations have practices that facilitate a better work-life balance, including part-time work flexibility, time off for family duties, childcare support and the ability to work from home (Bloom and Van Reenen 2006). This suggests that better-managed organisations might have environments that are more conducive to a higher share of women, but this has not been tested explicitly.

Scores reflecting the quality of management at the departmental level (specifically, operations management quality and people management quality) were collected for around 160 departments (covering English, Psychology, Business and Computer Science) in 2012 by McCormack, Propper and Smith (2014). We add these (standardised) management scores as further controls in Equations (2) and (3) to see whether there is any evidence that management practices can explain the observed environment effect. We find that better managed departments—particularly in the dimension of people management—are indeed more diverse, both at the time of REF submission (close to when the management scores were collected) and afterwards (columns (2) and (4), respectively). However, after including the management practice scores, the positive relationship between environment score and post-REF improvement remains, and increases in magnitude, although it also becomes insignificant, likely due to smaller samples.

Given small samples, these results are only suggestive. Moreover, we do not know if they only hold for specific departments (business, computer science, English and psychology) or instead extend more broadly. Nevertheless, they provide preliminary evidence that people management processes may be an important component of a positive environment that can increase representation from historically under-represented groups and may be a direction for future research.

5 Discussion

Our paper provides new evidence from REF 2014 on the relationship between alternative measures of the quality of research in an academic department and the diversity of its academic staff, measured by the share of historically under-represented groups. The main findings are that a measure of output research quality is negatively correlated with this measure of diversity, while measures of the impact of research and the quality of the research environment positively correlate with it.

One implication is that the choice with respect to the scope of research quality matters for diversity in higher education. In several countries which have national research assessment processes, there have been debates on the best way to assess research quality. There is a push for metric-based systems for assessing outputs, which have the attraction of being cheaper to implement. The Australian Research Council, for example, has used an evaluation system strongly supported by bibliometric indicators in its Excellence in Research for Australia assessments (Arnold, *et al.* 2018). To the extent that narrow and metrics-based approaches are biased against certain groups, however, then our evidence indicates that this approach will result in a misallocation of resources within the sector.

A second implication is that broadening the scope of research quality measures can mitigate some of the negative effects on diversity. By incorporating measures of research impact and environment quality alongside a measure of output quality, the REF allocated more funding to departments that increased diversity than it otherwise would have done, albeit the differences are small. To quantify the effect of incorporating the environment score, we can compare the average post-REF change in diversity, weighted by the amount of funding that departments receive (according to the funding formula $4 \times \text{percentage } 4^* + 3 \times \text{percentage } 3^*$) first, based only on outputs and second, incorporating environment scores. The output-weighted increase in diversity (the reduction in the share of white men) is 3.73 percentage points. Adding environment score increases this to 3.76. This is a positive effect, but small, partly because the environment score carries a small weight in the overall REF 2014 assessment (0.15 compared to 0.65 for outputs) and partly because the environment score is closely correlated with the output score. If outputs and environment were weighted equally, the weighted increase in diversity

would be 3.81. The plans for REF 2029 are to reduce the weight given to outputs and to increase the weight given to environment.

The point of this paper is to provide evidence that the scope of research quality measurement matters and can have implications for under-represented groups. We have considered the UK REF because it offers a range of different measures of quality, but this is not an endorsement of current REF measures. As has been discussed, there may be a high level of subjectivity in the assessments. There is also ambiguity in the four different elements that are included in the environment measure. Unpacking this—and understanding exactly which elements of the research environment are beneficial for diversity—remains a topic for further discussion and research.

References

Arnold, Erik, Paul Simmonds, Kristine Farla, Peter Kolarz, Bea Mahieu, and Kalle Nielsen. 2018. *Review of the Research Excellence Framework.* Evidence Report, technopolis.

Bateman, Victoria, Danula K Gamage, Erin Hengel, and Xianyue Liu. 2021. *The gender imbalance in UK economics.* Silver Anniversary Women's Committee Report, London: Royal Economic Society.

Bishop, Dorothy. 2021. *Is the benefit of the REF really worth the cost?* 28 April. Accessed April 14, 2022. https://www.timeshighereducation.com/opinion/benefit-ref-really-worth-cost.

Bloom, Nick, and John Van Reenen. 2006. "Management practices, work-life balance, and productivity: a review of some recent evidence." *Oxford Review of Economic Policy* 22 (4): 457–482.

Bloom, Nicolas, Renata Lemos, Raffaella Sadun, Daniela Scur, and John Van Reenen. 2014. "The new empirical economics of management." *Journal of the European Economic Association* 12 (4): 835–876.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. 2020. "Are referees and editors in economics gender neutral?" *Quarterly Journal of Economics* 135 (1): 269–327.

Colussi, Tommaso. 2018. "Social ties in academia: a friend is a treasure." *Review of Economics and Statistics* 100 (1): 45–50.

De Fraja, Gianni, Giovanni Facchini, and John Gathergood. 2019. "Academic salaries and public evaluation of university research: evidence from the UK Research Excellence Framework." *Economic Policy* 34 (99): 523–583.

Dion, Michelle L, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. 2018. "Gendered citation patterns across political science and social science methodology fields." *Political Analysis* 26 (3): 312–327.

Ductor, Lorenzo, Sanjeev Goyal, and Anja Prummer. 2023. "Gender and collaboration." *Review of Economics and Statistics* 105 (6): 1366–1378.

Ferber, Marianne. 1988. "Citations and networking." *Gender and Society* 2 (1): 82–89.

Ferber, Marianne. 1986. "Citations: are they an objective measure of scholarly merit?" *Signs* 11 (2): 381–389.

Gamage, Danula K., Almudena Sevilla, and Sarah Smith. 2020. "Women in economics: a UK perspective." *Oxford Review of Economic Policy* 36 (4): 962–982.

Gamage, Danula K., and Almudena Sevilla. 2019. "Gender equality and positive action: evidence from UK universities." *AEA Papers and Proceedings* 109: 105–109.

Hengel, Erin. 2022. "Publishing while female. Are women held to higher standards? Evidence from peer review." *The Economic Journal* 132 (648): 2951–2991.

Hengel, Erin, and Eunyoung Moon. 2023. "Gender and equality at top economics journals." Mimeo.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask principal-agent analyses: incentive contracts, asset ownership, and job design." *Journal of Law, Economics, and Organization* 7 (1991): 24–52.

Kerr, Steven. 1975. "On the folly of rewarding A, while hoping for B." *Academy of Management Journal* 18 (4): 769–783.

Koffi, Marlène. 2021. “Innovation ideas and gender inequality.” Mimeo.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. “Bibliometrics: global gender disparities in science.” *Nature* 504 (7479): 211–213.

Lundberg, Shelly, and Jeanna Stearns. 2019. “Women in economics: stalled progress.” *Journal of Economic Perspectives* 33 (1): 3–22.

MacIntosh, Robert. 2021. “Radical rethink of UK's excelence frameworks is needed.” *Times Higher Education*, 16 April.

McCormack, John, Carol Propper, and Sarah Smith. 2014. “Herding cats? Management and university performance.” *The Economic Journal* 124 (578): F534–F564.

REF 2014. 2012. *Panel criteria and working methods.* REF 2014.

REF 2021. 2020. *Panel criteria and working methods.* REF 2021.

Thorpe, Andy, Russell Craig, Dennis Tourish, Glenn Hadikin, and Sasa Batistic. 2018. “‘Environment’ submissions in the UK’s Research Excellence Framework 2014.” *British Journal of Management* 29 (3): 571–587.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, *et al.* 2015. “The Metric Tide.” Report of the role of metrics in research assessment and management.

Table 1 Broad standards for REF 2014 grading

	Outputs	Impact	Environment
4*	World-leading	Outstanding	World-leading
3*	Internationally excellent… but falls short of the highest standards	Very considerable	Internationally excellent… but falls short of the highest standards
2*	Recognised internationally	Considerable	Recognised internationally
1*	Recognised nationally	Recognised but modest	Recognised nationally

Table 2 Correlations between REF 2014 scores and departmental diversity

	(1)	(2)	(3)
Dependent variable: D_{dsi}			
Output	-0.047***	-0.005	-0.011
	(0.011)	(0.010)	(0.011)
Impact	0.023***	0.009*	0.010**
	(0.006)	(0.005)	(0.005)
Environment	-0.015**	-0.020***	-0.015***
	(0.006)	(0.005)	(0.005)
Constant	52.846***		
	(1.815)		
Env.-out.	0.047	0.269	0.771
Env.-imp.	0.000	0.000	0.001
Out.-imp.	0.000	0.216	0.094
Sub-panel f.e.		✓	✓
Institution f.e.			✓
No. obs.	1,635	1,635	1,635
R-squared	0.041	0.448	0.538
Dependent variable: ΔD_{dsi}			
Output	-0.013**	-0.014*	-0.014
	(0.006)	(0.007)	(0.009)
Impact	0.000	0.000	-0.002
	(0.004)	(0.004)	(0.004)
Environment	0.007*	0.008*	0.009**
	(0.004)	(0.004)	(0.004)
Constant	5.038***		
	(1.207)		
Env.-out.	0.027	0.036	0.041
Env.-imp.	0.288	0.234	0.095
Out.-imp.	0.091	0.105	0.213
Sub-panel f.e.		✓	✓
Institution f.e.			✓
No. obs.	1,598	1,598	1,598
R-squared	0.004	0.046	0.166

Note. Results from estimating Equation (2) (panel A) and Equation (3) (panel B). D_{dsi} is the share of non-white-male staff in a department in 2013 (in percentages); ΔD_{dsi} is the change in D_{dsi} between 2013–2018. Scores are the weighted sum of 4* and 3* research ($4 \times$ percentage $4^{*} + 3 \times$ percentage 3^{*}). Sample excludes multiple submissions from the same departments (65 observations). Standard errors clustered at the institution level in parentheses. ***, ** and * significant at the 1%, 5% and 10% level, respectively.

Table 3 Correlations between REF 2014 scores and departmental diversity, additional controls

	(1)	(2)	(3)	(4)
Dependent variable:	D_{dsi}	D_{dsi}	ΔD_{dsi}	ΔD_{dsi}
Output	-0.011	-0.011	-0.014	-0.029
	(0.011)	(0.033)	(0.009)	(0.027)
Impact	0.010**	0.024*	-0.002	-0.038**
	(0.005)	(0.013)	(0.004)	(0.015)
Environment	-0.014***	-0.016	0.009**	0.017
	(0.005)	(0.015)	(0.004)	(0.013)
Athena	-0.666	-2.158	0.119	2.538
	(0.916)	(2.567)	(0.858)	(2.517)
Management		0.319		0.349
		(0.670)		(0.536)
People management		0.462**		0.538**
		(0.213)		(0.217)
Sub-panel f.e.	✓	✓	✓	✓
Institution f.e.	✓	✓	✓	✓
No. obs.	1,635	166	1,598	164
R-squared	0.538	0.877	0.166	0.718

Note. Results from estimating Equation (2) (columns (1) and (2)) and Equation (3) (columns (3) and (4)). D_{dsi} is the share of non-white-male staff in a department in 2013 (in percentages); ΔD_{dsi} is the change in D_{dsi} between 2013–2018. Scores are the weighted sum of 4* and 3* research ($4 \times$ percentage $4^{*} + 3 \times$ percentage 3^{*}). Athena is an indicator variable equal to one if the department mentioned the word "Athena" at least once in its environment statement; Management is a measure of the (average) quality of management practices relating to operations (on a scale of 1–5) collected by McCormack, Propper and Smith (2014); People management is a is a measure of the (average) quality of management practices, relating to personnel. Sample excludes multiple submissions from the same departments (65 observations). Standard errors clustered at the institution level in parentheses. ***, ** and * significant at the 1%, 5% and 10% level, respectively.

Appendices

A. Summary statistics

Table 4 Summary statistics

	Departments	Disciplines	Institutions
No. observations	1,736	36	151
Staff composition			
No. of staff submitted to REF 2014	28.5	30.9	25.0
Non-white male staff (2013)	45.2	43.8	47.1
Non-white male staff (2018)	49.1	47.9	50.2
Funding and degrees (2013–2018)			
Research income (mill. £)	13.1	17.6	10.7
No. PhDs awarded	52.7	60.9	39.6
% REF 2014 environment rated as			
Four star	26.5	28.2	21.4
Three star	42.3	43.8	37.5
Two star	24.2	22.1	29.3
One star	6.5	5.4	10.7
Unclassified	0.5	0.4	1.1
% REF 2014 research output rated as			
Four star	19.0	19.3	16.9
Three star	45.2	46.9	41.8
Two star	29.0	28.0	31.9
One star	5.9	5.2	8.0
Unclassified	0.8	0.7	1.4
% REF 2014 impact rated as			
Four star	32.0	33.8	29.7
Three star	41.6	41.7	38.4
Two star	19.6	18.4	21.7
One star	5.1	4.5	7.5
Unclassified	1.6	1.6	2.7

Note. Table presents departmental (first column), discipline (second column) and institutional (third column) counts (first row) and averages (remaining rows). Data from HESA and REF 2014.

B. Mapping HESA cost centres to REF 2014 UoAs

HESA's standardised departmental classification scheme ("cost centres") does not perfectly overlap with the boundaries of REF 2014 discipline-specific assessment panels ("units of assessment" or UoAs). Table 5 shows how we mapped cost centres to UoAs.

Table 5 Mapping from REF 2014 UoAs to HESA cost centres

REF 2014 UoA	HESA cost centre
1: clinical medicine	101: clinical medicine
2: public health, health services and primary care	102: clinical dentistry 105: health and community studies
3: allied health professions, dentistry, nursing and pharmacy	103: nursing and allied health professions 107: pharmacy and pharmacology
4: psychology, psychiatry and neuroscience	104: psychology and behavioural sciences
5: biological sciences	106: anatomy and physiology 112: biosciences
6: agriculture, veterinary and food science	109: veterinary science 110: agriculture, forestry and food science
7: earth systems and environmental sciences	111: earth, marine and environmental sciences
8: chemistry	113: chemistry
9: physics	114: physics
10: mathematical sciences	122: mathematics
11: computer science and informatics	121: IT, systems sciences and computer software engineering
12: aeronautical, mechanical, chemical and manufacturing engineering	116: chemical engineering 120: mechanical, aero and production engineering
13: electrical and electronic engineering, metallurgy and materials	117: mineral, metallurgy and materials engineering 119: electrical, electronic and computer engineering
14: civil and construction engineering	118: civil engineering
15: general engineering	115: general engineering
16: architecture, built environment and planning	123: architecture, built environment and planning
17: geography, environmental studies and archaeology	124: geography and environmental studies 126: archaeology
18: economics and econometrics	129: economics and econometrics
19: business and management studies	133: business and management studies 134: catering and hospitality management
20: law	130: law
21: politics and international studies	128: politics and international studies
22: social work and social policy	131: social work and social polity
23: sociology	132: sociology
24: anthropology and development studies	127: anthropology and development studies
25: education	135: education 136: continuing education
26: sport and exercise sciences, leisure and tourism	108: sports science and leisure studies
27: area studies	125: area studies
28: modern languages	137: modern languages
29: English language and literature	138: English language and literature
30: history	139: history
31: classics	140: classics
32: philosophy	141: philosophy
33: theology and religious studies	142: theology and religious studies
34: art and design: history, practice and theory	143: art and design
35: music, drama, dance and performing arts	144: music, dance, drama and performing arts
36: communication, cultural and media studies, library and information management	145: media studies

C. Understanding the environment score

C.1. What does the environment score measure?

In this appendix, we provide further insights into the environment assessment in the REF. Including an assessment of UoAs' research environments in the REF has not been uncontroversial. A primary concern is that assessment of a narrative statement involves a degree of subjective judgement, arguably to a greater extent than judging research outputs or impact. Furthermore, the environment assessment takes place towards the end of the assessment period, after panel members already know output (and impact) grades; thus panels may calibrate departments' environment grade profiles to match their research output and impact grades, removing independent variability.

To explore this possibility, we estimate the following OLS regression (variables defined as in Section 3.2):

$$\text{Environment}_{dsi} = \beta_0 + \beta_1 \text{ Outputs}_{dsi} + \beta_3 \text{ Impact}_{dsi} + \rho' X_d + \varphi_s + \phi_i + \varepsilon_d. \quad (4)$$

Equation (4) includes a vector of additional controls (X_d) for the number of completed PhDs and total grant income, information that UoAs submitted to REF 2014 alongside their narrative environment statements.

The regression results (reported in Table 6 below) show that higher environment scores are associated with higher output and impact gradings; the correlation is stronger with outputs. The environment score correlates positively with (the natural logs of) total research income and the number of PhD degrees awarded between 2008–2013. The environment score also correlates positively with the number of staff members submitted to REF 2014, which may indicate that larger departments have better research environments, although it was also the case that larger departments (with more submitted staff) were allowed to submit longer environment statements.[11] Adding HEI fixed effects substantially increases the regression's R^2; combined, the controls used in column (4) absorb 74 per cent of the variation in environment scores.

[11] In REF 2014, UoAs were not required to submit research—whether in the form of outputs or impact—from every research active member of staff. "Submitted staff members" are the number of staff for whom the UoA submitted at least one research output or impact case study to the REF 2014.

Table 6 Correlation between environment scores and other indicators

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable: environment						
Output	0.884***	0.658***	0.661***	0.436***	0.449***	0.442***
	(0.036)	(0.036)	(0.036)	(0.041)	(0.041)	(0.041)
Impact	0.524***	0.405***	0.406***	0.335***	0.320***	0.322***
	(0.021)	(0.021)	(0.021)	(0.022)	(0.022)	(0.022)
Log research income		2.209**	1.846*	1.012	2.054*	2.035*
		(1.082)	(1.092)	(1.154)	(1.156)	(1.158)
Log PhDs		27.043***	24.606***	14.780***	16.157***	16.425***
		(2.257)	(2.489)	(2.776)	(2.756)	(2.785)
No. submitted staff			0.183**	0.124**	0.130	0.133
			(0.079)	(0.084)	(0.083)	(0.085)
% diversity words					41.400***	43.035***
					(7.099)	(7.290)
% non-white-male staff						-0.069
						(0.107)
Constant	-87.616***	-128.464***	-121.366***			
	(6.622)	(11.523)	(11.911)			
No. obs.	1,658	1,646	1,646	1,646	1,644	1,622
R-squared	0.629	0.677	0.678	0.743	0.749	0.748
Institution f.e.				✓	✓	✓

Note. Results from estimating Equation (4). Scores are the weighted sum of 4* and 3* research ($4 \times$ percentage $4^* + 3 \times$ percentage 3^*). Log research income and log PhDs are the (natural logs of) total research income and the number of PhD degrees awarded between 2008–2013. No. submitted staff is the number of staff members (in full-time equivalents) submitted to REF 2014. % diversity words is the share of diversity-related words in a department's environment statement (see Appendix C.2). % non-white-male staff is the percentage of a department's staff members who were women or non-white in 2013 (data from HESA; see Section 3.2 for more details). Standard errors in parentheses. ***, ** and * significant at the 1%, 5% and 10% level, respectively.

What explains the residual variation and to what extent does it capture promotion of equality and diversity? Recall, this was something that sub-panel members were explicitly told to take into account in their assessments of the quality of the research environment. As evidence that panel members paid attention to promotion of equality and diversity, we show that the environment grading captures aspects of the narrative statement that reflect a greater focus on diversity and inclusion. Column (5) in Table 6 presents results showing that environment statements that used more inclusive language—*i.e.*, had a higher share of words related to diversity—scored higher.[12] The results show that inclusive language positively correlates with environment scores.[13] By contrast, actual staff diversity at the time of REF

[12] The variable "% diversity words" measures the prevalence of a set of diversity words in the environment statement. The mean is 0.47% with a range across all departments from 0.05% to 2.21%; see the following section (Appendix C.2) for the list of diversity related words (examples include "URM", "gender", "inclusive" and "culture").

[13] However, the coefficient is not significant if fixed effects for unit of assessment are included ($\beta = -2.48$, standard error 8.06) due to systematic variation in the use of these words across the four main panels. Arts and humanities uses the most inclusive language (mean = 0.60%), followed by social sciences (mean = 0.52%), medicine, health and life sciences (mean = 0.35%) and physical sciences, engineering and maths (mean = 0.29%). Within social sciences, the units of assessment submitting to the economics and econometrics sub-panel used the least inclusive language in their statements (mean = 0.26%).

2014 submission, measured by the percentage of non-white-male staff in 2013, correlates negatively, albeit insignificantly, with the environment score (see column (6)).

One interpretation of the two findings in columns (5) and (6) is that the environment grading captured "words not deeds"; alternatively, given the forward-looking nature of the environment statement, it is also plausible that higher environment scores reflect strategies and procedures to promote equality and diversity that are in place at the end of the REF period that may help to promote diversity and equality going forward.

C.2. List of words related to diversity

The list of words used to define the variable "% diversity words" in Table 6 was put together in the following way. The words from the website https://blog.ongig.com/diversity-and-inclusion/diversity-terms/ formed the starting point. We (the authors) then used our judgement to exclude words that had a limited relationship with diversity in a REF environment statement and to include additional words that we thought would reflect diversity in the UK higher education context (*e.g.*, Athena SWAN).

Table 7 List of diversity-related words

ableism	disability	integration	neurodivergent	separation
accessibility	disabled	intersectionality	neurodiverse	sex
accountability	disablism	justice	neurodiversity	sexism
acculturation	discrimination	lesbian	nonbinary	stereotype
athena	emotional	lgbt*	nurture	supremacy
bame*	equality	lgbta	nurturing	tokenism
belonging	equity	lgbtiqq*	oppression	transexual
bias	ethnic	lgbtq*	pansexual	transgender
bigotry	ethnicity	lgbtq2*	poc*	transitioning
bisexual	exclusion	lgbtqia*	polyamory	transmasculine
black	female	male	polygender	transphobia
bme*	gay	mansplain	prejudice	transvestite
career	gender	marginalisation	privilege	unity
childcare	groupthink	microaggression	qpoc*	urm*
cisgender	heteronormativity	minority	queer	xenophobia
cultural	heterosexuality	misgender	race	
culture	homosexual	multicultural	racism	
decolonisation	identity	multiethnic	racist	
decolonization	inclusion	multiracial	religion	
diaspora	inclusive	neuroatypical	segregation	

*bame: Black and Asian minority ethnicity; bme: Black and minority ethnicity; lgbt: lesbian, gay, bisexual, and transgender; lgbta: lesbian, gay, bisexual, transgender, and asexual/aromantic/agender; lgbtiqq: lesbian, gay, bisexual, transgender, intersex, queer, and questioning; lgbtq: lesbian, gay, bisexual, transgender, and queer (or questioning); lgbtqia: lesbian, gay, bisexual, transgender, queer (or questioning), intersex, and asexual (or allies); lgbtq2: lesbian, gay, bisexual, transgender, queer (or sometimes questioning), and two-spirited; poc: people of colour; qpoc: queer people of colour; urm: under-represented minority.