Development of a tool to accurately predict UK REF funding allocation

Shahd Al-Janabi[1], Lee Wei Lim[2], and Luca Aquili[1,2*]

[1]College of Health & Human Sciences, Charles Darwin University, Darwin, Northern Territory, Australia.

[2]Neuromodulation Laboratory, School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, China.

*Corresponding author: Dr Luca Aquili. Email: luca.aquili@cdu.edu.au

Keywords: REF; impact factor; metrics; funding

Abstract

Understanding the determinants of research funding allocation by funding bodies, such as the Research Excellence Framework (REF) in the UK, is vital to help institutions prepare for their research quality assessments. In these assessments, only publications ranked as 4* or 3* (but not 2* or less) would receive funding. Correlational studies have shown that the impact factor (IF) of a publication is associated with REF rankings. Yet, the precise IF boundaries leading to each rank are unknown; for example, would a publication with an IF of 5 be ranked 4* or less? Here, we provide a tool that predicts the rank of each submitted publication to (1) help researchers choose a publication outlet that would more likely lead to the submission of their research output(s) by faculty heads in the next REF assessment, thereby potentially improving their academic profile; and (2) help faculty heads decide which outputs to submit for assessment, thereby maximising their future REF scores and ultimately their research funding. Initially, we applied our tool to the REF 2014 results for Neuroscience, Psychiatry, and Psychology, which predicted publications ranked 4* with 95% accuracy (IF $\geq$ 6.5), 3* with 98% accuracy (IF= 2.9-6.49), and 2* with 95% accuracy (IF= 1.3-2.89). We then generalised these findings to another REF unit of assessment: Biological Sciences to further demonstrate the predictive capacity of our tool.

Introduction

The latest Research Excellence Framework (REF) exercise in 2014 distributed £1.6 billion per year of research funding to universities and research institutions in the United Kingdom (UK). In determining how funds are allocated, research outputs (65% weightage) submitted to the REF are assessed by a panel of experts who assign a research quality grade or ranking of 4* (world leading), 3* (internationally excellent), 2* (internationally recognised), 1* (nationally recognised), or U/C (unclassified; below national standard) to each institution. Importantly, only outputs judged to have a score of 4* or 3* are funded, with 4* (ranging from £7504 to £14,639 depending on discipline) receiving four times as much money as 3* (ranging from £1876 to £3659 depending on discipline) (Koya and Chowdhury 2017). The method describe in this paper focused on the funding allocations based on submitted research outputs. It should be noted that funding is also distributed on the basis of research impact (20% weightage) defined by the REF as "an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia", which is a measure of Ph.D. completions, and also on the basis of research infrastructure (i.e., laboratory facilities; cumulative 15% weightage). These weightages and definition apply to both REF 2014 and the incoming REF 2021.

To gauge how well a university department might fare in the REF assessments, institutions run mock REF exercises, in which research outputs are graded by other colleagues. These mock assessments are time consuming (Farla and Simmonds 2015) and expensive (Jump 2015). Moreover, for academics who do not perform well in these internal mock assessments, it can alarmingly lead to active research contracts turning into teaching-only or even potential job losses (UCU 2013). As a result, there have been calls for the implementation of (costly)

alternatives to the peer-reviewed mock assessments (Stern and Nurse 2014), which are primarily based on citation counts (Harnad 2009; Norris and Oppenheim 2003).

To date, some studies have shown that the H-index of a department correlates with the REF rankings or REF GPA (grade point average) (Oppenheim 1995, 1997); however, other investigations have not replicated these findings (Mryglod et al. 2015). Using a machine learning approach, Balbuena (Balbuena 2018) demonstrated that the best predictors of REF GPA were the number of Web Science documents, entry tariff, and the proportion of students coming from independent schools. Additionally, Chowdhury et al. (Chowdhury et al. 2016; Koya and Chowdhury 2017), reported a linear relationship between the number of upper quartile (Q1 or top 25%) impact factor (IF) publications and the REF GPA. Although IF is widely criticized by academics on numerous grounds (Callaway 2016; Hicks et al. 2015; Paulus et al. 2018; Smaldino and McElreath 2016; Berenbaum 2019), unlike other metrics, it has been shown to predict the future success of scientists (Acuna et al. 2012; Győrffy et al. 2020; Van Dijk et al. 2014), and has been used by some universities to motivate staff to publish in the most prestigious journals (e.g., through the provision of cash incentives for publications with a high IF) (Abritis and McCook 2017; Quan et al. 2017). Furthermore, IF is positively correlated with retraction rates (Fang and Casadevall 2011), possibly due to the perceived benefits to one's academic standing driving publications in a high IF journal. However, some of the above reviewed studies had limitations in that they examined variables (e.g., H-index, entry tariff) that cannot be controlled by individual authors/faculties; hence, knowledge of such relationships with REF GPA does not afford researchers any advantage. Moreover, there is no correlational evidence between the controllable metrics (such as IF of the journal) and REF GPA that allow the determination of the likely cut-offs leading to a 4*, 3*, or 2* rating.

In this research, we set out to describe a tool that predicts the rank of each submitted publication. For authors, the purpose of this tool is to identify journals or, more precisely, the impact factor associated with a journal that most likely leads to an award of a 4* or 3* rating (for research outputs contributing 65% to the REF assessment). The importance of IF could help individual authors to (1) plan research studies in such a way that they can publish in outlets likely to result in a 4*/3* rating; and (2) be selected by their Faculty Head for the REF assessment, thus enhancing their academic profile. For faculty heads, this tool would be most beneficial in allowing them to strategically pick the research outputs of individual staff members that have the best chance of being awarded a 4* or 3*, thereby potentially increasing the funding allocation. Researchers and Faculty Heads already have a good sense of what leads to a "good quality" paper; however, we outline here a novel, practical, and accurate quantitative approach to define this "sense". A discussion of the benefits and limitations (methodological and philosophical) of this tool is also provided.

Results

Accuracy of GPA prediction

We assessed the accuracy of our tool by comparing our predicted GPAs with the REF 2014 GPAs for each institution. We compared the two sets of results for each institution by a paired t-test (Fig. 1a). The analysis showed the actual REF 2014 GPAs (M = 2.71, SD = 0.43) and our predicted GPAs (M = 2.66, SD = 0.62) were not statistically different (t (80) = 1.43, p = 0.16). These results indicate our tool for assigning rank to each publication based on IF cut-off values can accurately predict the ranking of publications, as our predicted GPAs for each institution were comparable to the REF 2014 GPAs.

Accuracy of the percentage of outputs per rank prediction

To further check the accuracy of our tool, our predicted percentages of publications with ranks ranging from 4* to U/C were compared with the percentage of publications awarded each rank by REF 2014. First, we calculated the difference scores between each of the two variables for each rank in each institution. Second, we conducted a one-way repeated measures ANOVA with Bonferroni corrections on the mean difference scores of each rank across institutions. Mauchly's Test of Sphericity was violated; hence, Lower-bound corrections were used. We found no significant differences between rank categories ($F(1, 80) = 2.93$, $p = 0.09$), which indicates the accuracy of our predicted percentage of outputs was similar across ranks compared to the actual REF 2014 percentage of outputs (Fig. 1c). The accuracy of our tool in predicting the percentage of outputs can be seen in Figure 1b. This accuracy was calculated by summing the percentage of publications in each rank as identified by the REF 2014 versus our tool. The error rate was calculated as the difference between the two sums and was subtracted from the true positive (sum of the percentage of publications in each rank identified by REF 2014) as an indication of our hit rate. The percentage accuracy was then calculated using our hit rate divided by the true positive.

Accuracy of rankings prediction

We ordered the predicted GPAs to identify the predicted (output-based) ranking of each institution. With this information, we conducted a Wilcoxon signed-rank test to compare the REF 2014 output-based institution rankings with our predicted rankings to ascertain the accuracy of our tool. The test showed our predicted institution rankings did not statistically differ from that of the REF 2014 institution rankings ($Z = -0.082$, $p = 0.94$). The level of agreement between the institution rankings derived from our IF-based tool versus the REF 2014 results was further confirmed by the Bland-Altman plot (Fig. 2). Specifically, the

Bland-Altman plot revealed no systematic differences or consistent bias between the institution rankings derived using our tool versus the REF 2014 results. Indeed, few values (<4%) lay outside of the limits of agreement (mean of differences ± 1.96 SD), indicating agreement between the two measurements.

Strength of GPA predictions

We conducted two regression analyses to ascertain the extent to which our IF-based tool can predict REF 2014 quality profile metrics; specifically, GPA. The first simple linear regression analysis assessed whether our predicted GPA can be used to predict actual REF 2014 GPAs. We found that our tool could accurately predict the GPA compared to the actual GPA (unstandardized $\beta = 0.62$, $p < 0.0001$; overall model fit of $R^2 = 0.80$) (Fig. 3). This effect remained even after conducting a hierarchical regression that accounted for each institution's total output and 2013 University ranking. The model predicting GPA was the only statistically significant model ($p < 0.0001$ vs. $p > 0.09$ for all other models), indicating our tool based on IF cut-off values can predict GPA in the REF 2014.

Generalisability of tool

We examined the generalisability of our tool by applying it to a different area evaluated in the REF 2014 Unit of Assessment: Biological Sciences. Such an investigation should confirm the predictive value of our tool. First, we assessed the accuracy of our tool by comparing our predicted GPAs with the REF 2014 GPAs for each institution who submitted publications to the Biological Sciences Unit of Assessment. As above, we compared the two sets of results for each institution by a paired t-test (Fig. 5a). The analysis showed the actual REF 2014 GPAs ($M = 2.89$, $SD = 0.40$) and our predicted GPAs ($M = 2.93$, $SD = 0.43$) were not statistically different ($t(43) = 1.50$, $p = 0.14$). The accuracy of our tool in predicting the

percentage of outputs in Biological Sciences can be seen in Figure 5b. Second, we conducted a simple linear regression analysis to probe whether or not our tool could accurately predict the GPAs compared to the actual GPAs of institutions who submitted to the Biological Sciences Unit of Assessment. This analysis showed our method was accurate (unstandardized $\beta = 0.86$, $p < 0.0001$; overall model fit of $R^2 = 0.85$) (Fig. 6). This effect remained even after conducting a hierarchical regression that accounted for each institution's total output and 2013 ranking. As before, the model predicting GPA was the only statistically significant model ($p < 0.0001$ vs. $p > 0.14$ for all other models). These results collectively indicate that our tool can accurately predict GPA in the REF 2014 for another Units of Assessment besides Neuroscience, Psychiatry, and Psychology.

Discussion

The primary goal of this study was to identify the IF threshold values for predicting the different REF rankings and GPAs. These findings have the potential to be used by individual authors and faculty heads to make informed decisions on (1) where to submit and publish articles; and (2) which research outputs should be submitted to the next REF assessment to maximise funding allocation. We first briefly summarise the key findings and then provide an in-depth interpretation of these results and their implications.

Our analysis of the REF 2014 Unit of Assessment: Neuroscience, Psychiatry, and Psychology demonstrated accuracy scores in excess of 95% for predicting the percent of publications ranked 4*, 3*, or 2*. Specifically, we revealed that to receive the highest possible score of 4*, a publication submitted to REF must have an IF equal to or greater than 6.5. We estimated with 98% accuracy that a publication with an IF of 2.9 or less receiving a rank of 2* or below was unlikely to receive funding. We replicated these findings in the REF unit of

assessment: Biological Sciences. Although the IF boundaries changed, the accuracy of the model predictions for publications ranked 4*, 3*, or 2* remained largely similar. Our findings significantly expand on previous correlational studies by showing that the IF of publications can be used to predict REF ranking and GPA (Chowdhury et al. 2016; Koya and Chowdhury 2017). We have developed a precise numerical tool to predict the research output quality score based on REF 2014 results.

Our estimates for the IF cut-offs are based on the total GPA achieved by each institution and the corresponding percentages of 4*, 3*, 2*, 1*, and U/C as assigned in the REF 2014 published by the *Times Higher Education*. Figure 1a shows the similarity of our mean estimated REF GPAs for all universities (2.66) compared to the actual REF results (2.71); Figure 1b shows the similarities of our estimated percentages of research outputs with 4*, 3*, 2*, 1*, and U/C scores versus the assigned scores, and Figure 4 shows the similarities between the university REF GPA rankings and our predictions, overall confirming the accuracy of our estimated IF cut-offs used in the REF 2014 Unit of Assessment. It is important to note, however, that our IF cut-offs are not predictions at the individual output level, as this information is not publicly available. The report commissioned by the Higher Education Funding Council for England (HEFCE) is the only study that has attempted to match individual research outputs to REF scores (Wouters et al. 2015). The authors concluded that “The statistics presented do not overwhelmingly support the use of metrics as a replacement for a peer-review-driven model of research quality assessment”. However, it is impossible to further comment on their findings, as the methodological details in their report were scarce and IF was not used as a metric to derive prediction cut-offs.

Interestingly, our hierarchical regression analyses showed that both the total number of research outputs and university rankings (aka: league tables, as determined by The Guardian University guide and ranking for 2013) were not significant predictors of REF GPA. These results suggest that the evaluators in the assessments, such as REF 2014, primarily (if not solely) rely on the journal IF to make their decisions on the rank (4* to U/C) of a submitted publication.

It is important to acknowledge the current study has limitations and the results should be considered with a note of methodological and philosophical caution. The specific IF boundaries established in this study are limited to predicting the REF rankings from two Units of Assessment: (1) Neuroscience, Psychiatry, and Psychology, and (2) Biological Sciences; thus, future studies are needed to confirm the accuracy of our tool when applied to other fields. Nevertheless, given that other disciplines, such as Clinical Medicine, Agriculture, Veterinary and Food Science, Chemistry, Economics, and Econometrics, have been reported to have similar correlations between REF quality and SCImago (an alternative measure of The Journal Citation Reports IF, which is highly correlated with IF; Rocha-e-Silva 2010) to that in Neuroscience, Psychiatry, and Psychology and Biological Sciences ($R \geq 0.437$; (Wouters et al. 2015), we expect our tool will have predictive capacity when also applied to these fields. It is important to stress that our tool may be less precise for some Units of Assessment (e.g., Law, Business and Management Studies, Arts and Design, Social Work and Social Policy, etc.) where a predominant proportion of the assessed research outputs were not from journal articles (e.g., from chapters in a book, conference contributions, patents, exhibitions, etc.), which do not have an IF.

It should also be noted that our IF cut-offs were determined post-hoc (i.e., after the REF 2014 data was made available). The next REF assessment is due to take place this year (2021), and

our IF cut-offs will need to be adjusted to take into account any increases in journal IF values since 2013 (Althouse et al. 2009). The median IF for Neuroscience-related journals increased by 22% from 2007 to 2013, whereas the increase from 2013 to 2020 has so far been about 5%. Additionally, individual researchers/faculties using our tool need to take the 2020 IF values corresponding to the bottom publications in each rank (4* - U/C) and recreate the confidence intervals (CI) around the new mean IF of each rank across institutions. It is likely that these changes will not grossly affect the accuracy of our model[1]. The expectation that our tool will continue to be generally accurate relies on two assumptions. First, there should be a high correlation in the proportions of 4* and 3* assigned to each institution between REF 2014 and REF 2021, as we showed in our analyses of the Research Assessment Exercise (RAE) 2008 and REF 2014[2]. The pattern of results indicates that the number of outputs receiving each rank may be stable over time. Second, the distribution of REF quality ratings should remain largely similar to that in 2021. This was not the case between RAE 2008 and REF 2014, where the number of 4* and 3* ratings doubled in 2014; however, this difference has been widely acknowledged as an inflation of grades (Marginson 2014). Our IF cut-offs for the next REF assessment (2021) also need to take into account any new REF 2021 submission rules different from REF 2014 (e.g., number of outputs per individual (REF

1 To test the accuracy of our model with different IF cut-offs, under the assumption that the IFs of publications submitted to exercises, such as the REF 2014, may change over time, we examined the accuracy of our model with cut-offs that are 10% greater and 10% lower than the cut-offs (lower-bound confidence interval) used in this study. The variations in our IF cut-offs indicate that our model remained accurate, particularly for the first two ranks. For 10% greater cut-offs: 84% accuracy for 4*, 91% for 3*, 89% for 2*, 86% for 1*, and 21% for U/C; and for 10% lower cut-offs: 75% accuracy for 4*, 95% for 3*, 78% for 2*, 50% for 1*, and 24% for U/C. This analysis indicated our tool can predict possible REF rankings and GPAs with changing IF cut-offs.

2 We extracted the percentage of publications awarded a 4* and 3* during RAE 2008 for the Unit of Assessment: Psychology (note that Psychiatry and Neuroscience were measured separately in very few entries). With this information, we conducted a correlational analysis between the percentage outputs for RAE 2008 and REF 2014 for each rank. This analysis revealed a medium-to-large correlation between the percentage outputs receiving a 4* ($r = 0.78$, $p < 0.0001$) and 3* ($r = 0.50$, $p < 0.0001$) in the 2008 versus 2014 exercises, suggesting the number of outputs receiving each rank may be stable over time.

2019)). As our tool was based on REF 2014, any predictions of results not yet published are bound to be less precise.

Whilst our tool was highly accurate (i.e., above 90%) in identifying the ranks of 2*, 3*, and 4*, it was less accurate for U/C and 1* ranks. We can provide some plausible explanations as to why this pattern of results may have occurred. First, we automatically assigned a rank of U/C to any research article which we could not derive an IF (see Methods for the derivation of an IF for each publication). This strategy likely resulted in over assigning U/C ranks when a panel of reviewers would have judged the publications differently. Take for example the case of a newly formed journal from an established publishing group (e.g., *Nature*, *Frontiers*). Because these publishing groups are well-known by academics, it is foreseeable that a panel of reviewers would have assigned a higher score than U/C to a publication e.g., in *Nature Human Behaviour* and/or *Frontiers in Physiology*, despite them not yet having an IF or a stable IF pattern. Second, the predicted IF range for an output classified as U/C and 1* was much smaller than for 2*, 3* and 4* ranks (see Fig. 1 and Fig. 5). This pattern of results may lead to minor IF deviations that result in underclassifying and/or overclassifying a research output to one rank versus another. Third and related to the second point, we can speculate that when the IF of a journal is very low, reviewers who may not be familiar with these journals may use other heuristics that are as salient as IF for determining the rank of a paper, such as the institutional ranking of the corresponding author. Lastly, the sample size for U/C and 1* ranks was between one tenth and one twentieth of 2*, 3*, and 4* ranks, which likely increased the error rate. Indeed a previous study on the Finnish ranking of research publications found that misclassification/model errors occurred even when citation-based metrics predicted most of the expert-based rankings for papers (Saarela et al. 2016). Despite these methodological limitations, it is important to note that the practical utility of our tool

297 remains largely unaffected, given that we identified that only the outputs assigned a value of
298 3* and 4* will receive funding under the REF scheme.
299 Our analyses indicate that funding bodies, such as REF 2014 in the UK, may base their
300 allocation decisions on the IF of the outputs submitted for assessment by each institution.
301 Given the speed in which these decisions must be made, this strategy makes sense insofar as
302 journals with higher IFs are more visible, hence, one can use this metric as a 'quick and dirty'
303 index for (a) the importance of each output to a wide readership; and (b) the (world-class)
304 success of the researchers/institution of each output. If our assumption are indeed true then an
305 obvious outcome, as previously stated, is that it may be advantageous for researchers to
306 publish in higher IF journals and/or for Faculty Heads to submit only high IF outputs for
307 assessment in order to increase funding allocation to the institution. Although this strategy is
308 *implied* by our findings, our view is that researchers, Faculty Heads, and funding bodies
309 should consider the ethics of using IF as a prime measure of deciding where to publish or for
310 allocating funding, primarily because IF was conceived as a metric of journal usage not
311 author scholarship. The formula for IF is the number of cited articles published within a given
312 period divided by the total number of citable outputs published within that period (Jones
313 2013; Marson 2020). This formula (a) does not index the way in which an article was read or
314 used following publication, thus obscuring the assessment of the publication's impact or
315 importance; (b) does not index the actual merit of a publication in that journal because the
316 increased citation could be due to weaknesses and flaws in the publication as opposed to its
317 strengths, thus obscuring the assessment of the researcher's success; (c) does not index
318 publications that have been cited in text books and not in or in addition to an indexable
319 journal, thus obscuring both the above assessments; and (d) does not have high reliability
320 (Greenwood 2007), though it may have high validity in some fields (Saha et al. 2003, also see
321 Jarwal et al. 2009; Law and Leung 2020)). In listing the limitations of applying IF as a

metric, we seek to highlight the ethical issue(s) that may arise from its use in strategy, planning, and decision-making, as per the implications of our findings. In the future, the increasing number of easily accessible tools that chart bibliometrics may encourage funding bodies to make more use of an output's citation count as opposed to IF in decision-making (Marson 2020).

In summary, our results provide a simple practical tool that can help individual researchers and/or heads of departments/faculties in the Unit of Assessment: Neuroscience, Psychiatry, and Psychology and Biological Sciences to accurately estimate the UK REF quality score based on the IF of their publications. Although our findings were based on REF 2014, our additional analyses [1] demonstrate that the tool likely has predictive capabilities for REF 2021 (despite the approaching REF 2021 submission deadline). Our findings largely agree with previous literature on the association between various publication metrics and scientific success/research funding, but further expands on these studies by providing a more cost and time effective approach for evaluating research papers in the context of REF compared to mock peer-reviews (Farla and Simmonds 2015; Jump 2015). It is possible that this tool could be applied to other research funding allocation exercises (e.g., ERA Research Assessment in Australia) with similar utility after some modifications.

Methods

Data sources and variables

The primary data for this study was the REF 2014 results (Research Excellence Framework 2014) in the UK. As proof of concept, we focused on the Units of Assessment: Neuroscience, Psychiatry, and Psychology, and Biological Sciences to assess the accuracy of our tool in predicting REF rankings. These units were chosen because 99% of their outputs were

research articles (and hence likely to have an associated IF). There were 81 institutions who submitted publications for the Unit of Assessment: Neuroscience, Psychiatry, and Psychology totalling a combined 9,121 publications. On the other hand, there were 44 institutions who submitted publications for the Unit of Assessment: Biological Sciences totally a combined 8,608 publications. For each Unit of Assessment, we extrapolated the output and quality profiles from the REF 2014 dataset: each institution name, the journals of the submitted publications, and the evaluations for each institution, including the percentage of publications awarded each rank (4* to U/C) at the culmination of the REF 2014 assessment. We added four variables to this extrapolated dataset. The first variable was the total number of outputs (i.e., publications) submitted by each institution to the Unit of Assessment for evaluation. The second variable was the IF of each journal as of 2013 (sourced from Journal Citation Reports 2014). Publications for which a 2013 IF could not be sourced were either assigned the IF for the preceding or following year. We were unable to source an IF in 1.29% of publications for the Unit of Assessment: Neuroscience, Psychiatry, and Psychology, and 1.20% of publications for the Unit of Assessment: Biological Science. The third variable added was the Grade Point Average (GPA) awarded to each institution for each Unit of Assessment (Research Excellence Framework 2014: Institutions Ranked by Subject 2014), which is calculated by multiplying the percentage of publications in each rank by its rating (adding the total across all ranks and dividing by 100), giving a GPA index of an institution's overall quality of research. Finally, the fourth variable added was the ranking of each institution in 2013 (University guide 2013: University league table 2013).

Calculation of the prediction cut-off values

We first defined the IF boundaries or cut-off values for each of the REF ranks in each Unit of Assessment using the 2014 dataset. We calculated the number of publications assigned each rank of 4*, 3*, 2*, 1*, and U/C for each institution in each Unit of Assessment. Next, we

ordered the publication ranks by their IF for each institution in each Unit of Assessment and identified the IF value for the publication that was likely to be at the bottom of each rank in each institution. The IF cut-offs were then identified by calculating the number of publications assigned a rank from 4* to U/C in each institution. For example, for a given institution with six publications assigned a 4* by REF 2014, we ranked these submitted publications by their IF and identified the publication with the lowest IF for that institution in that rank. We then calculated the mean IF for each rank across institutions in each Unit of Assessment with their corresponding confidence intervals. The lower bound confidence interval served as our prediction cut-off for each rank in each Unit of Assessment (see Fig. 1b and Fig. 5b).

Predicting the ranking of publications and GPA

Using IF cut-off values for each Unit of Assessment, we aimed to predict the rank of each publication submitted to REF 2014. For example, all publications with an impact factor equal to or above the lower bound confidence interval for a 4* (e.g., 6.54 for Neuroscience, Psychiatry, and Psychology) were assigned a rank of 4*, whereas all publications with an impact factor less than the lower bound confidence interval for a 1* (e.g., 0.71 for Neuroscience, Psychiatry, and Psychology) were assigned a rank of U/C. Next, we calculated the percentage of publications in an institution that we predicted to be awarded each rank. With that information, we also calculated the predicted GPA for each institution. This calculation used the same formula described above (see Fig. 4 for comparisons).

Figure 1. Panel A shows the mean GPA awarded by REF 2014 versus the predicted mean GPA determined by our tool for the Unit of Assessment: Neuroscience, Psychiatry, and Psychology. **Panel B** shows the percentage of publications awarded each rank across institutions by REF 2014 versus our predicted percentage of publications for the Unit of Assessment: Neuroscience, Psychiatry, and Psychology using our tool based on impact factor cut-off values. We used those cut-off values to predict the number of publications receiving each rank from 4* to U/C according to the REF 2014 Unit of Assessment: Neuroscience, Psychiatry, and Psychology. **Panel C** shows the difference score between REF 2014 and our predicted percentage of publications in each rank for the Unit of Assessment: Neuroscience, Psychiatry, and Psychology to validate the accuracy of the tool. All error bars reflect standard error of the mean.

Figure 2. Bland-Altman plot for institution rankings derived from our tool versus the REF 2014 results for the Unit of Assessment: Neuroscience, Psychiatry, and Psychology. The centre line indicates the mean difference, whereas the lines on either side are the limits of agreement (mean ± 1.96 SD).

Figure 3. The relationship between the predicted GPA using our tool and REF 2014 GPA for each institution that submitted publications to the Unit of Assessment: Neuroscience, Psychiatry, and Psychology. The blue line is the best fit (r = 0.90) and green dotted lines are the upper and lower confidence interval boundaries.

Figure 4. The GPA awarded by REF 2014 (in green) and the predicted GPA by our tool (in blue) for each institution that submitted publications to the Unit of Assessment: Neuroscience, Psychiatry, and Psychology. The GPA is used to assess the quality of research published in the fields of Neuroscience, Psychiatry, and Psychology that takes into account the percentage of outputs in each institution assigned a rank from 4* to U/C. Panel A represents institutions that are in the upper GPA quartile of REF 2014, panels B and C represent institutions in the middle quartiles (2nd and 3rd, respectively), and panel D represents institutions in the lower GPA quartile.

Figure 5. shows the mean GPA awarded by REF 2014 versus the predicted mean GPA determined by our tool for the Unit of Assessment: Biological Sciences. **Panel B** shows the percentage of publications awarded each rank across institutions by REF 2014 versus our predicted percentage of publications in the Unit of Assessment:

Biological Sciences using our tool based on impact factor cut-off values. We used those cut-off values to predict the number of publications receiving each rank from 4* to U/C according to the REF 2014 Unit of Assessment: Biological Sciences.

Figure 6. The relationship between the predicted GPA using our tool and REF 2014 GPA for each institution that submitted publications to the Unit of Assessment: Biological Sciences. The blue line is the best fit (r = 0.92) and green dotted lines are the upper and lower confidence interval boundaries.

Declarations

Funding

The study received no funding.

Competing interests

The authors declare no competing interests.

Availability of data and material

Data available from public resources (REF 2014, The Guardian University Guide) and the Journal of Citation Reports.

Code availability

N/A

Author contributions

LA conceived the study; SA and LA developed the methodology and analysed the data; SA, LWL and LA wrote the manuscript.

REFERENCES

Abritis, A., & McCook, A. (2017). Cash incentives for papers go global. *Science, 357*(6351), 541-541, doi:10.1126/science.357.6351.541.

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Predicting scientific success. *Nature, 489*(7415), 201-202, doi:10.1038/489201a.

Althouse, B., West, J., Bergstrom, T., & Bergstrom, C. (2009). Differences in Impact Factor Across Fields and Over Time. *Journal of the American Society for Information Science and Technology, 60*, doi:10.1002/asi.20936.

Balbuena, L. D. (2018). The UK Research Excellence Framework and the Matthew effect: Insights from machine learning. *PLOS ONE, 13*(11), e0207919, doi:10.1371/journal.pone.0207919.

Berenbaum, M. R. (2019). Impact factor impacts on early-career scientist careers. *Proceedings of the National Academy of Sciences, 116*(34), 16659-16662, doi:10.1073/pnas.1911911116.

Callaway, E. (2016). Beat it, impact factor! Publishing elite turns against controversial metric. *Nature News, 535*(7611), 210.

Chowdhury, G., Koya, K., & Philipson, P. (2016). Measuring the impact of research: lessons from the UK's Research Excellence Framework 2014. *PLOS ONE, 11*(6), e0156978.

Fang, F. C., & Casadevall, A. (2011). Retracted Science and the Retraction Index. *Infection and Immunity, 79*(10), 3855-3859, doi:10.1128/iai.05661-11.

Farla, K., & Simmonds, P. (2015). REF 2014 accountability review: Costs, benefits and burden. *Brighton, UK: Technopolis| group|. Retrieved October, 1*, 2019.

Greenwood, D. C. (2007). Reliability of journal impact factor rankings. *BMC medical research methodology, 7*(1), 1-6.

Győrffy, B., Herman, P., & Szabó, I. (2020). Research funding: past performance is a stronger predictor of future scientific output than reviewer scores. *Journal of Informetrics, 14*(3), 101050, doi:https://doi.org/10.1016/j.joi.2020.101050.

Harnad, S. (2009). Open access scientometrics and the UK Research Assessment Exercise. *Scientometrics, 79*(1), 147-156.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.

Jarwal, S. D., Brion, A. M., & King, M. L. (2009). Measuring research quality using the journal impact factor, citations and 'Ranked Journals': Blunt instruments or inspired metrics? *Journal of Higher Education Policy and Management, 31*(4), 289-300.

Jones, J. (2013). The impact of impact factors and the ethics of publication. Springer.

Journal Citation Reports (R) (2014). In T. Reuters (Ed.).

Jump, P. (2015). Winners and losers in HEFCE funding allocations. *Times Higher Education, 26*, 6-9.

Koya, K., & Chowdhury, G. (2017). Metric-based vs peer-reviewed evaluation of a research output: Lesson learnt from UK's national research assessment exercise. *PLOS ONE, 12*(7), e0179722.

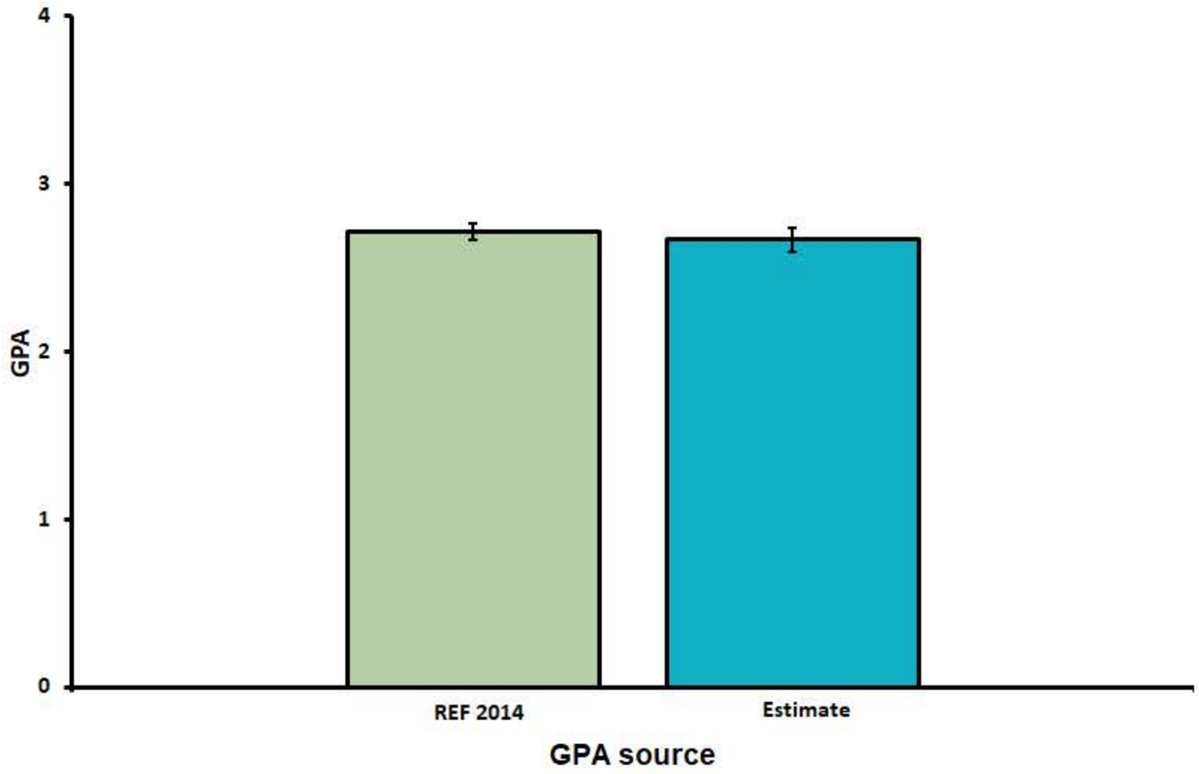
Law, R., & Leung, D. (2020). Journal impact factor: A valid symbol of journal quality? *Tourism Economics, 26*(5), 734-742.

Marginson, S. (2014). UK research is getting better all the time–or is it? *The Guardian*.

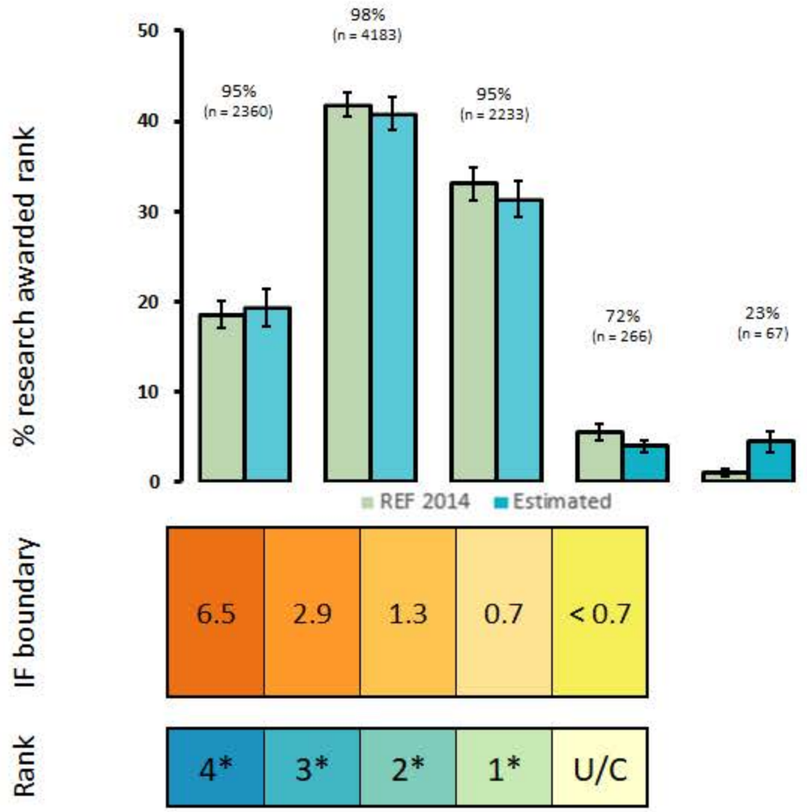
Marson, S. (2020). Editorial: Is the Impact Factor (IF) Ethical To Use for

Promotion and Tenure Decisions? *Journal of Social Work Values and Ethics, 17*, 2-5.
Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2015). Predicting results of the Research Excellence Framework using departmental h-index. *Scientometrics, 102*(3), 2165-2180.
Norris, M., & Oppenheim, C. (2003). Citation counts and the research assessment exercise V. *Journal of Documentation*.
Oppenheim, C. (1995). The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British library and information science university departments. *Journal of Documentation*.
Oppenheim, C. (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation*.
Paulus, F. M., Cruz, N., & Krach, S. (2018). The impact factor fallacy. *Frontiers in psychology, 9*, 1487.
Quan, W., Chen, B., & Shu, F. (2017). Publish or impoverish. *Aslib Journal of Information Management*.
REF (2019). Guidance on Submissions. Research England Bristol.
Research Excellence Framework (2014). http://www.ref.ac.uk/2014/.
Research Excellence Framework 2014: Institutions Ranked by Subject (2014). https://www.timeshighereducation.com/sites/default/files/Attachments/2014/12/17/g/o/l/sub-14-01.pdf.
Rocha-e-Silva, M. (2010). Impact factor, scimago indexes and the brazilian journal rating sytem: where do we go from here? *Clinics, 65*(4), 351-355.
Saarela, M., Kärkkäinen, T., Lahtonen, T., & Rossi, T. (2016). Expert-based versus citation-based ranking of scholarly and scientific publication channels. *Journal of Informetrics, 10*(3), 693-718.
Saha, S., Saint, S., & Christakis, D. A. (2003). Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association, 91*(1), 42.
Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society open science, 3*(9), 160384.
Stern, N., & Nurse, P. (2014). It's our duty to assess the costs of the REF. *Times Higher Education. https://www. timeshighereducation. com/comment/letters/its-our-duty-to-assess-the-costs-of-the-ref/2017479. article*.
UCU (2013). The Research Excellence Framework UCU Survey Report. UCU London.
University guide 2013: University league table (2013). https://www.theguardian.com/education/table/2012/may/21/university-league-table-2013.
Van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology, 24*(11), R516-R517.
Wouters, P., Thelwall, M., Kousha, K., Waltman, L., De Rijcke, S., Rushforth, A., et al. (2015). The metric tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management). *London: Higher Education Funding Council for England (HEFCE)*.

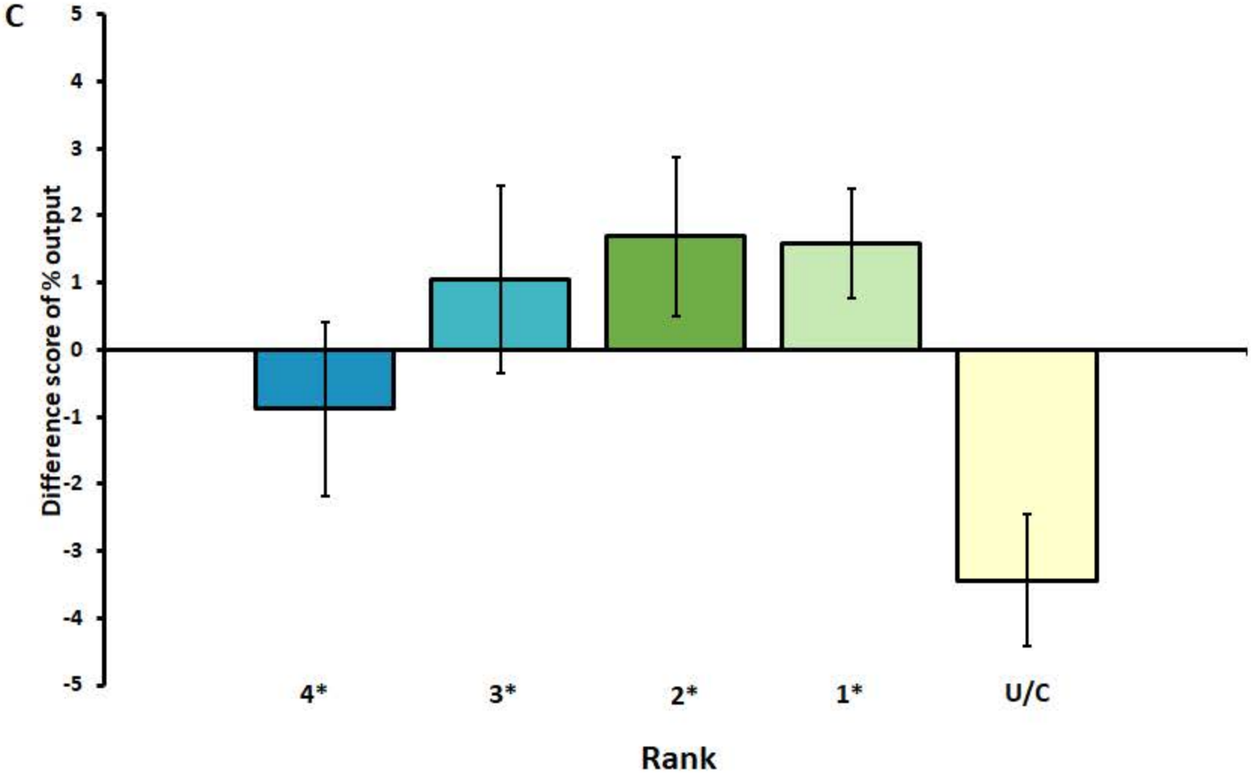
A

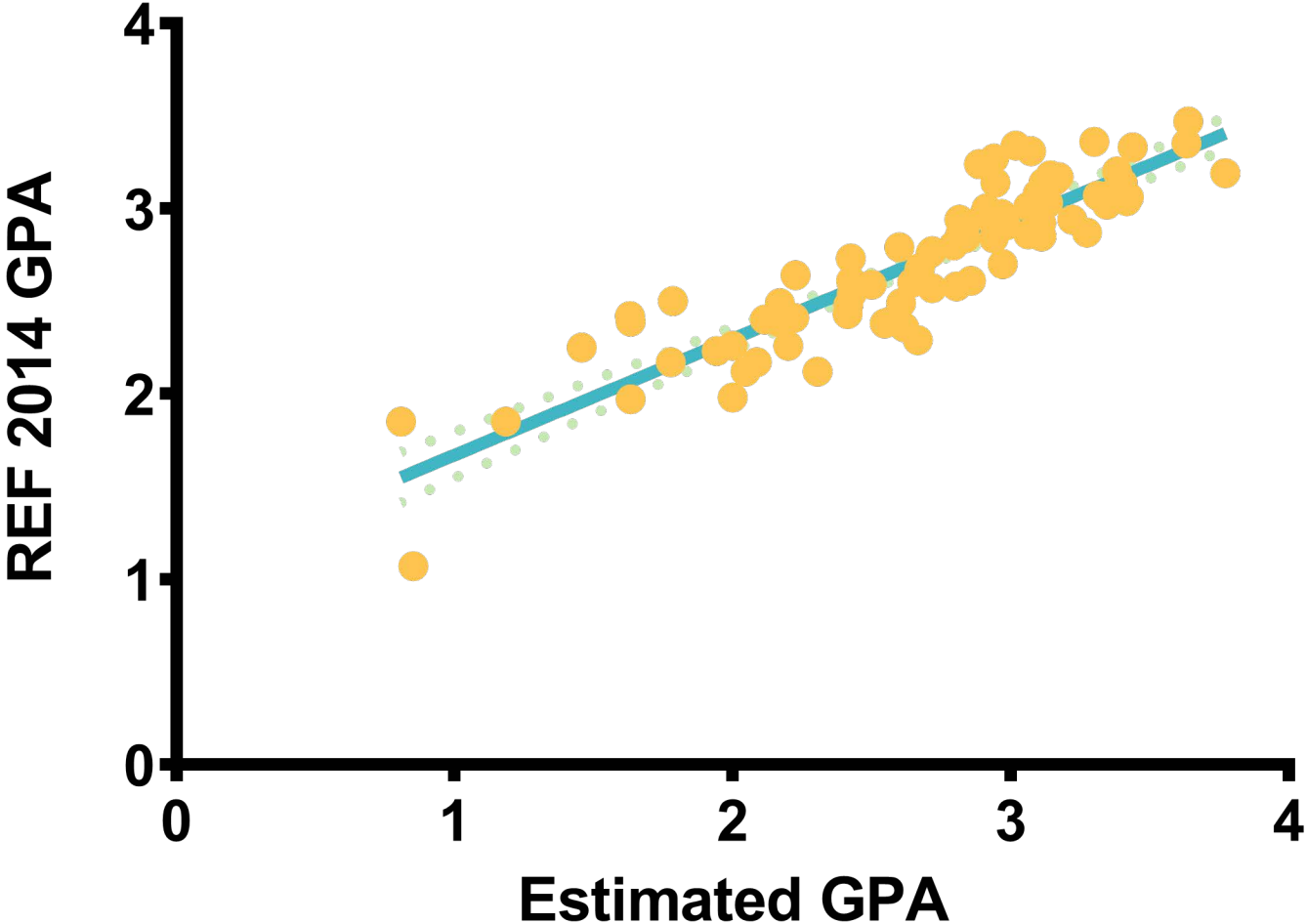

GPA
4
3
2
1
0
REF 2014
Estimate
GPA source

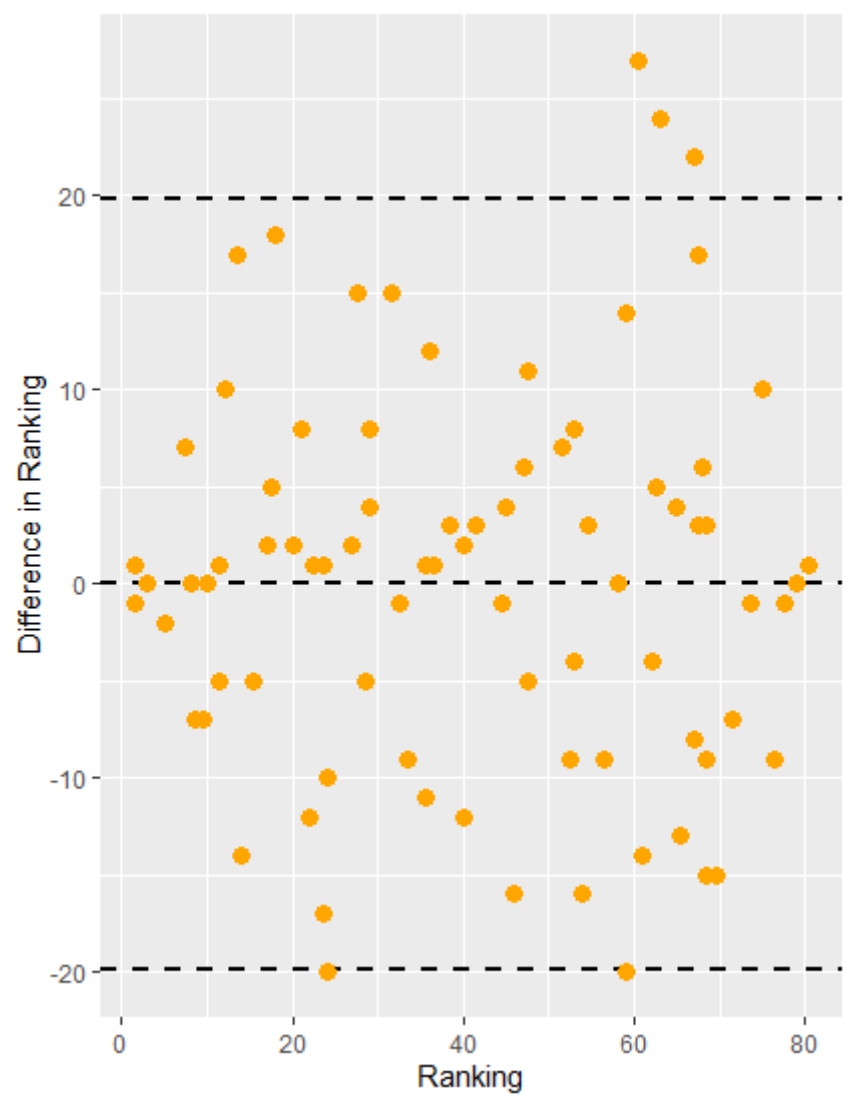
B


% research awarded rank
50
40
30
20
10
0
95%
(n = 2360)
98%
(n = 4183)
95%
(n = 2233)
72%
(n = 266)
23%
(n = 67)
REF 2014
Estimated
IF boundary
6.5
2.9
1.3
0.7
< 0.7
Rank
4*
3*
2*
1*
U/C

C


Difference score of % output
5
4
3
2
1
0
-1
-2
-3
-4
-5
4*
3*
2*
1*
U/C
Rank


REF 2014 GPA
Estimated GPA
0
1
2
3
4
0
1
2
3
4


Difference in Ranking
20
10
0
-10
-20
0
20
40
60
80
Ranking

A

Institute Name	REF 2014	Estimate
University of Oxford	3.47	3.64
University of Cambridge	3.35	3.63
Imperial College London	3.19	3.77
Cardiff University	3.33	3.44
University of York	3.36	3.3
University of Sussex	3.2	3.38
University of Glasgow	3.14	3.4
University of Edinburgh	3.06	3.42
King's College London	3.04	3.42
Newcastle University	3.07	3.35
Royal Holloway, University of London	3.31	3.07
University of Leicester	3.07	3.31
University of Aberdeen	3.34	3.02
Birkbeck College	3.17	3.17
University of Birmingham	3.18	3.14
Bangor University	3.15	3.16
University of Exeter	3.14	3.12
University of Warwick	3.27	2.94
University of St Andrews	3.08	3.1
University of Essex	3.24	2.89
University of Stirling	3.14	2.95

B

Institute Name	REF 2014	Estimate
University College London	3.02	3.35
University of Dundee	3.03	3.14
University of Manchester	2.94	3.22
University of Bristol	2.87	3.27
University of Nottingham	3.01	3.06
University of Sheffield	2.92	3.11
University of Leeds	2.85	3.11
University of East Anglia	2.97	2.97
University of Durham	2.86	3.06
University of Plymouth	3	2.91
Swansea University	2.91	2.98
University of Southampton	2.96	2.9
Goldsmiths' College	2.84	2.94
Lancaster University	2.94	2.82
University of Liverpool	2.88	2.83
Bournemouth University	2.84	2.84
University of Surrey	2.86	2.82
Anglia Ruskin University	2.8	2.79
University of Reading	2.77	2.72
University of Kent	2.79	2.6

C

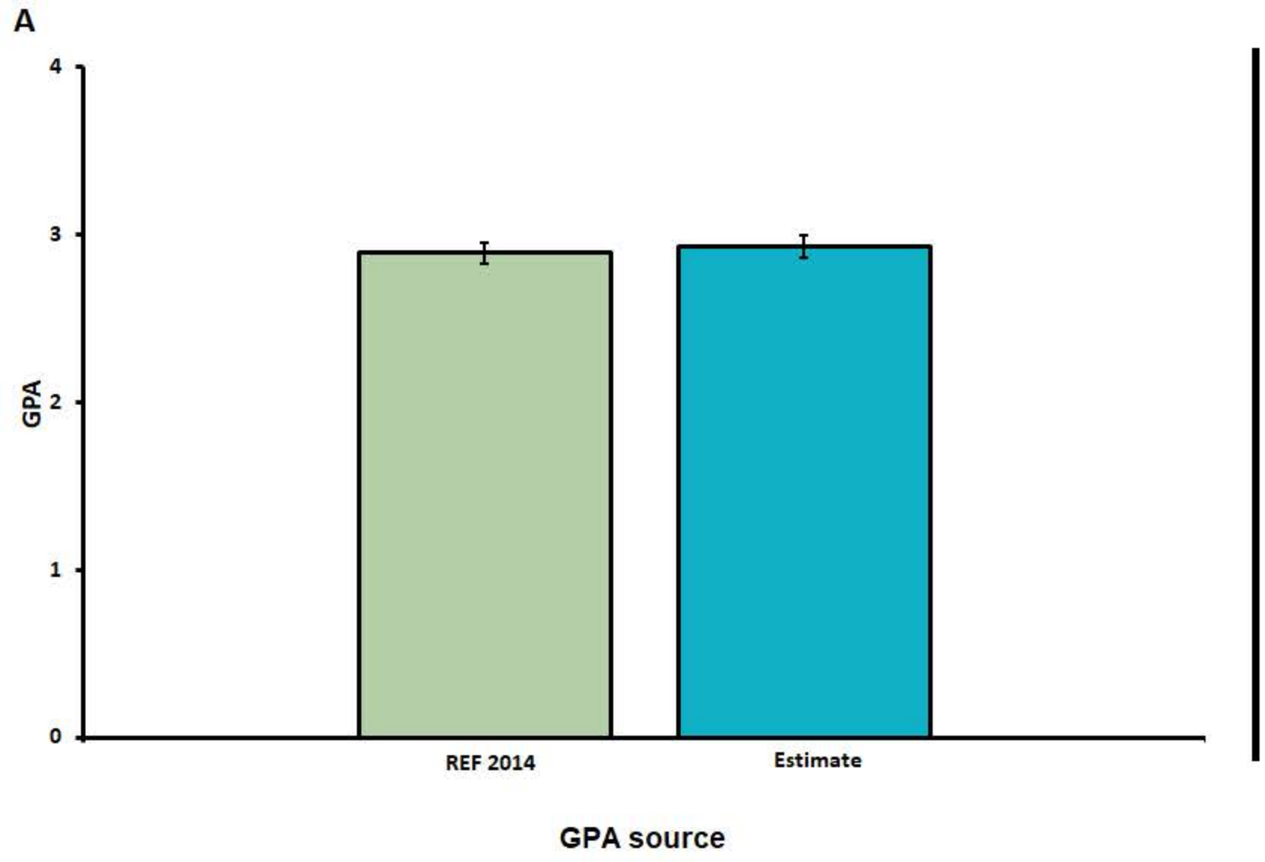
Institute Name	REF 2014	Estimate
City University London	2.7	2.97
University of Westminster	2.61	2.86
Queen's University Belfast	2.76	2.71
Liverpool John Moores University	2.58	2.81
University of Hull	2.68	2.67
University of East London	2.67	2.67
Middlesex University	2.57	2.72
Roehampton University	2.6	2.65
Nottingham Trent University	2.73	2.42
Kingston University	2.49	2.61
University of Ulster	2.59	2.5
University of Portsmouth	2.61	2.43
University of Abertay Dundee	2.54	2.43
University of Strathclyde	2.48	2.42
Edge Hill University	2.64	2.23
Brunel University London	2.43	2.41
Oxford Brookes University	2.49	2.17
University of Central Lancashire	2.41	2.22
York St John University	2.5	1.79
University of Chichester	2.42	1.63

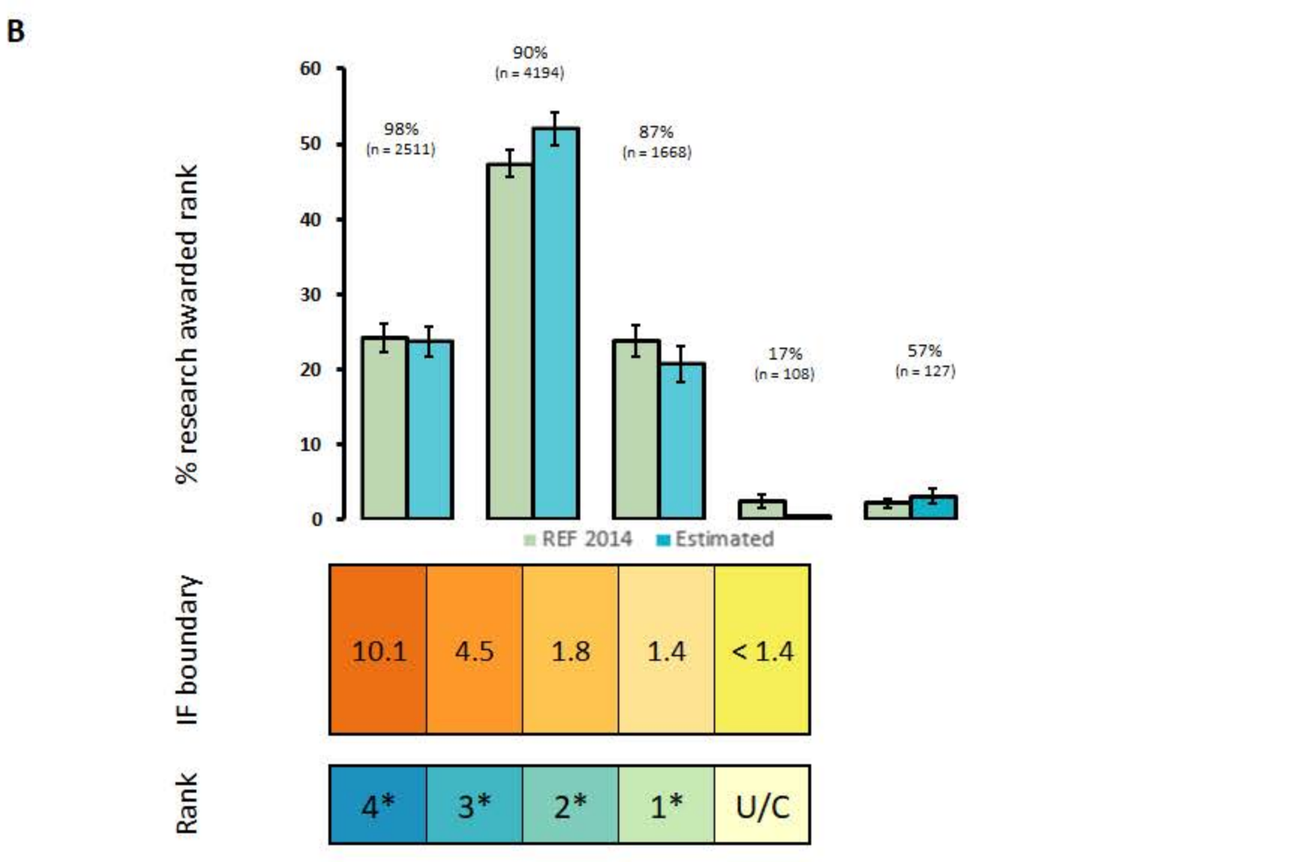
D

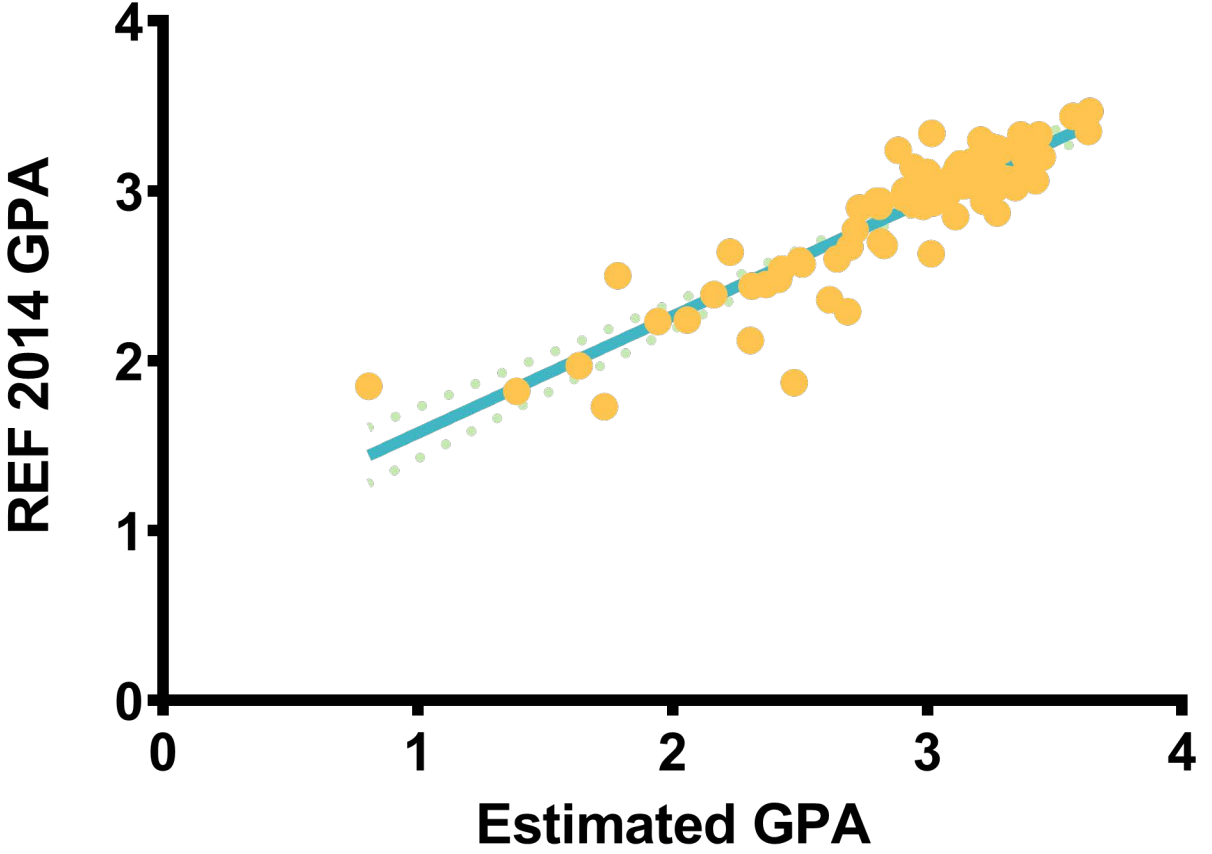
Institute Name	REF 2014	Estimate
Heriot-Watt University	2.36	2.62
University of Hertfordshire	2.39	2.57
University of Lincoln	2.29	2.67
University of Northumbria at Newcastle	2.38	2.55
Keele University	2.39	2.16
University of Greenwich	2.4	2.12
London South Bank University	2.26	2.2
University of Chester	2.12	2.31
Queen Margaret University Edinburgh	2.17	2.09
Liverpool Hope University	2.26	2
Leeds Beckett University	2.12	2.05
University of Derby	2.23	1.94
Coventry University	2.39	1.63
Staffordshire University	1.98	2
University of Winchester	2.17	1.78
University of Bolton	2.25	1.46
University of South Wales	1.97	1.63
University of Worcester	1.85	1.19
Glyndwr University	1.85	0.81
University of Northampton	1.07	0.85

GPA

Type Estimate REF 2014


A
4
3
2
1
0
GPA
REF 2014
Estimate
GPA source


B
60
50
40
30
20
10
0
% research awarded rank
98%
(n = 2511)
90%
(n = 4194)
87%
(n = 1668)
17%
(n = 108)
57%
(n = 127)
REF 2014
Estimated
IF boundary
10.1
4.5
1.8
1.4
< 1.4
Rank
4*
3*
2*
1*
U/C


REF 2014 GPA
Estimated GPA
0
1
2
3
4
0
1
2
3
4