ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Report from the RSS Working Group on Research Excellence Framework (REF) League Tables

Executive Summary

The UK's *Research Excellence Framework* (REF) assessment of research quality in universities serves two main purposes: to determine quality-related (QR) funding through HEFCE and the other national funding councils, and to provide public information on the relative standings of universities and of their departments. The published rankings or 'league tables' that followed REF 2014, as compiled by national newspapers and others, are however based on calculations that do not permit the intended direct comparisons of research quality. The main reason for this is that institutions submitting to the REF were permitted to *select* which researchers to include; and institutions differed substantially in their selectivity.

The 'strategic' exclusion of some eligible research staff from REF 2014 by universities had two main causes:

1. Universities correctly anticipated that the most prominent published rankings would not take account of selectivity.
2. The number of *Impact Case Studies* required in each REF submission was determined by the number of staff included.

The specific **recommendations** made here are based on statistical considerations of comparability and informative presentation. The recommendations are of two types, addressed respectively to the compilers of league tables and to the national funding councils. A common theme is that information to the public about research quality is best served by elimination, as far as possible, of the effects of 'strategic' decisions made by universities about their REF submissions.

> **R1. Published rankings should be more closely aligned with the QR funding outcome per researcher, and this should include the funding outcomes also for any non-submitted staff.** This would eliminate a strategic dilemma otherwise faced by universities, as well as providing a more direct interpretation of the published rankings.
>
> Specifically, for the construction of *more informative* REF-based league tables:
>
> **R1.1 Research quality levels should be weighted *according to the weights used in the QR funding formula* that is in use at the REF submission date.**
>
> **R1.2 In league tables the REF results should be *normalized*, to account for any non-submitted eligible research staff, by using the best available administrative data.**
>
> Taken together, R1.1 and R1.2 imply that league tables should be based on a suitably defined 'quality per researcher' measure, which differs in detail from the similar 'intensity' measure that is already in circulation. R1.2 makes sense if applied together with R1.1, as it is reasonable to assume that staff would only be excluded from REF if their university believed that their exclusion would not substantially reduce QR funding. Note that R1.1 is

based on the principle of aligning decisions about maximising QR funding with those of maximising league table position, and *do not* imply any *statistical* justification for the HEFCE choice of QR formula.

R1.3 Institution-level league tables that are based on such a 'quality per researcher' measure should also publish the number of eligible research staff as a proportion of total academic staff, as contextual information to aid interpretation of the results.

Related recommendations to HEFCE and the other national funding councils are as follows.

R2. Consideration should be given to adjustments to the REF submission and assessment rules, to enhance the public-information value of REF outcomes.

Specifically:

R2.1 The distorting 'threshold' effect of REF 2014's Impact Case Study requirements should be eliminated, as far as possible. It is evident that this aspect of the REF 2014 rules had a substantial effect on universities' decisions about the number of eligible research staff to include in each department's REF submission.

R2.2 It should be considered whether the reported 'overall profiles', derived from the detailed REF assessments of individual departments, can be made more informative. As currently constructed, the overall profiles often yield rankings that are influenced more by the REF assessments of research *Impact* and research *Environment* than by the assessment of research *Outputs*.

Authorship of this Report

The group formed in January 2015 to work on this report, published in May 2015, had the following Fellows of the RSS as its members: Professor Paul Fearnhead (Lancaster University, Department of Mathematics and Statistics), Professor David Firth (University of Warwick, Department of Statistics) and Professor Jonathan Forster (University of Southampton, Mathematical Sciences).

Report from the Working Group on Research Excellence Framework (REF) League Tables

1. Introduction

The 2014 Research Excellence Framework exercise (REF 2014) is a process organised by funding councils to assess the quality of research in UK Universities. The results of REF 2014 were published as a quality profile for each institution, in each of up to 36 units of assessment (UoAs) corresponding to different subject areas. The quality profile summarises the proportion of research activity assessed to be at each of five levels ranging from 4* (highest) to 0* (lowest). Further details of the REF 2014 process are described in [1]. REF 2014 follows on from previous Research Assessment Exercises (RAEs), the most recent being in 2008.

The most significant impact of the results of REF 2014 is through their use by the funding councils in allocation of research funding (called QR funding) to institutions from 2015-16. A stated aim of REF 2014 is also to "provide benchmarking information and establish reputational yardsticks, for use within the higher education (HE) sector and for public information" [2]. The most visible form of this is through the publication of league tables which order institutions according to a summary of research quality derived from REF 2014 results, both within individual UoAs, and overall. In preparing their returns to REF 2014, institutions were therefore required to weigh the relative importance of funding outcome and league table standing. This has been further clouded by the issue that a wide range of different summaries have been used to construct league tables, since the publication of the REF 2014 results.

In this paper, we discuss statistical issues around constructing *research quality* league tables based on REF 2014 results, with the aim of proposing recommendations for future practice. We wish to emphasise that we view such a league table as a summary of the relative success of institutions in the REF. We are not suggesting that any one-dimensional summary can a provide a robust measure of research quality. Care is needed when interpreting any league table which is summarising multi-faceted information, as there is always some arbitrariness in the criteria for ranking. Slight changes in criteria can often lead to substantial differences in rankings. League tables can be, and are, constructed both for institutions as a whole, and for individual UoAs. There are subtle differences in the issues around constructing league tables in these two scenarios. However we consider here a consistent methodology that is appropriate for both cases

The use of REF results to compare institutions, both within a UoA and as a whole, is made more difficult when different institutions, for quite valid reasons, prioritise funding outcome and league-table standing differently. While this remains an issue for any rankings based on REF 2014, closer alignment of the principles by which league tables (for future REFs) are constructed to expected funding outcomes would go some way towards making the resulting institutional comparisons more robust.

We note that REF 2014 results will also be used as a component in more general University league tables, i.e., league tables that take into account also other indicators such as student entry grades, student satisfaction, graduate employability, etc. Much of what follows is also directly relevant there, although we make no specific recommendations.

2. REF league tables

Each published league table summarises the overall research quality of a particular institution in a given UoA by using the REF 2014 'overall profile'$(q_4, q_3, q_2, q_1, q_0)$ to construct a one-dimensional index of research quality, of the form

$$Q = m \times (w_4 q_4 + w_3 q_3 + w_2 q_2 + w_1 q_1 + w_0 q_0) \quad (1)$$

where $w = (w_4, w_3, w_2, w_1, w_0)$ are weights assigned to the different quality grades, and m is a multiplier, derived from staff volume numbers in the institution and UoA concerned. Different choices of w and m can lead to very different orderings. In Table 1, we summarise how various league tables assign w and m.

Measure	**Publisher**	w	m
"GPA"	Various [3,4]	$(4,3,2,1,0)$	1
"Power" (H)	Times Higher [3]	$(4,3,2,1,0)$	Submission volume
"Power" (RF)	Research Fortnight [4]	$(3,1,0,0,0)$	Submission volume
"Intensity"	Times Higher [5]	$(4,3,2,1,0)$	Submission volume ÷ HESA volume

Table 1: A summary of differences between the main REF 2014 League Tables. Submission volume is the FTE staff number submitted for assessment in REF 2014 in the Institution and UoA concerned. HESA volume is the number of research active staff in the Institution and UoA concerned, according to data from the Higher Education Statistics Agency

In the following two subsections we discuss, in turn, the statistical issues around the choice of quality summary (w) and volume adjustment (m).

3. Summarising the quality profile

Construction of an ordering of institutions on the basis of REF 2014 outcomes requires a method for summarising a multidimensional quality profile in a single dimension. For summaries of the form of (1), this is achieved by the choice of $w = (w_4, w_3, w_2, w_1, w_0)$. There are two important statistical points which should be made about this. The first is that any projection onto a single dimension results in a loss of information. The second is that any choice of w which satisfies

$$w_4 \geq w_3 \geq w_2 \geq w_1 \geq w_0 \quad (2)$$

provides a coherent summary. The precise choice of w will depend on an, essentially subjective, decision about the relative importance of differences between the various quality grades. In particular, we note that, while the weighting $w = (4,3,2,1,0)$ may seem natural, there is no *statistical* basis to prefer this or any other weighting satisfying (2).

For REF outcomes, there *are* objective weightings, and they are the ones which funding councils use to allocate QR funding. While the formula for future allocations was not known at the time of publication of REF results, we propose that using the QR funding weightings that were in use during the year of the REF census date provides an objective summary, and one that has a

genuine (financial) interpretation in terms of expected QR funding. Different funding councils do not necessarily use exactly the same weightings, but they are typically very similar, and as the large majority of institutions submitting to REF receive funding from HEFCE, we propose to use the HEFCE funding weightings. For REF 2014 results, this implies the use of the weighting $w = (3,1,0,0,0)$, based on the HEFCE formula that was in place in 2013 to allocate QR funding based on the results of RAE2008. Aligning the quality weighting with QR funding goes part of the way towards aligning institutional goals of funding and league-table performance.

The initial (2015-16) QR funding formula was subsequently published by HEFCE in February 2015 [6], and uses the weighting $w = (4,1,0,0,0)$. We recommend that this weighting, or whatever has replaced it by the time of the date of publication of the results of the next REF, is used to construct league tables at that time.

The above leads to Recommendation R1.1.

4. Volume adjustment

Summarising research quality by weighted quality profile alone, i.e. with $m = 1$ in (1), produces a measure of "weighted average submission quality" (a generic term which we will abbreviate in the following as "WASQ"). The most prevalent such measure, using weights$w = (4,3,2,1,0)$, is the "grade point average" (GPA). The issue with using any WASQ measure, such as GPA, to compare institutions is that an institution can increase its standing by being more selective in terms of the staff it submits. Different institutions have had substantially different policies on how selective they have been when submitting staff to REF. So it is impossible to conclude from a WASQ ranking of two institutions whether the higher ranked institution has higher quality research or has just been more selective.

To account for this, rankings based on so-called "research power" have been proposed, which multiply a WASQ measure by $m =$ *submission volume*. Research power with the HEFCE funding weightings provides a measure of total funding volume. There is no benefit of a selective submission strategy to optimisation of research power. The difficulty with power-based rankings is that they do not provide a meaningful comparison of research quality between institutions with a large group in a particular UoA and those with a small group. Furthermore, as boundaries between UoAs are not perfectly defined, research power is also affected by how different institutions interpret those boundaries.

4.1 Using data on submission rates

An approach which aims to address the issues associated with GPA and Power is so-called "research intensity", whereby research power is normalised by a measure of the FTE volume of research-active staff in the institution in the UoA, the latter being estimated using data from HESA.

HESA provides an estimate of the FTE volume of research-active staff within each institution for each UoA. Hence research intensity sets $m =$ *submission volume* $\div$ *HESA volume* in (1), and is equivalent to multiplying a WASQ measure, such as GPA, by an estimate of the proportion of staff submitted. As with research power, there is no benefit of a selective submission strategy to optimisation of research intensity. Research intensity can also provide a meaningful comparison between institutions of different size.

Intensity-based ranking is not without its issues. Firstly these relate to its incorporation of the HESA staff volume measure, which is only an *estimate* of the actual FTE of research active staff. The limitations of the data are acknowledged by HESA [A]. This is particularly an issue for its use for rankings for specific UoAs, as there is not always a clear mapping between UoAs and research areas of staff. For staff who are not submitted, the choice of which UoA they are assigned to in the HESA data will then affect the estimate of proportion of staff submitted for those UoAs. As

submissions by institutions to individuals UoAs are often small, this can have a substantial effect on the estimate of proportion of staff submitted.

Even if we assume that the HESA data enables us to obtain an accurate estimate of the FTE volume of research-active staff within each UoA at each institution, there is also the question as to whether multiplying a WASQ by this proportion gives a meaningful measure of research quality. Ideally we should like it to give an estimate of, or at least be highly correlated with, the WASQ that would have been achieved had all research-active staff been submitted.

The research intensity measure used by [5] uses the GPA, a WASQ with $w = (4,3,2,1,0)$. As such, that research intensity measure is equivalent to assuming that non-submitted staff research is rated as *0** (or *Unclassified*) [B]. This is unrealistic, and has the effect of over-penalising institutions which submit a lower proportion of staff. The result is a quality measure that is overly dependent on the estimated proportion submitted, which is a concern given the potential uncertainty in the estimate of this proportion.

By choosing a different WASQ, which gives comparatively more weight to 4* and 3* research, one can obtain a more reliable measure. If we summarise the quality profile using the pre-REF 2014 QR formula, $w = (3,1,0,0,0)$, then the resultant research intensity measure effectively assumes that research of non-submitted staff is at most 2*. Whilst this will not be the case, it is reasonable to assume that institutions would be unlikely to choose a policy of selecting staff that would substantially reduce their QR income. The resulting intensity measure will be a reasonable measure of the relative QR income per FTE that each institution would have obtained if they had submitted all research-active staff. Hence it gives a sensible measure of research quality, by which institutions can be more meaningfully compared.

The use of such an intensity measure, or 'quality per researcher' measure as it might more transparently be described, is therefore recommended here, as a good way to make use of the available data to produce meaningful comparisons. The validity of such comparisons is limited mainly by the accuracy of the available administrative data (from HESA) on numbers of research-active staff and their most appropriate REF Units of Assessment. It should be recognised also that 'research active' status is defined in terms of the employment contracts of individual academics; as such, it is potentially open to manipulation through changes of contract to "teaching only" and suchlike. We suggest that, at institution level, league tables based on a measure of intensity should also publish the proportion of academic staff that are research active as contextual information to help the reader interpret the rankings.

The above leads to Recommendations R1.2 and R1.3.

5. Other issues

5.1 Effect of *Impact* on selectivity

The number of impact case studies required for a REF 2014 submission depended on the FTE-number of staff submitted to a given UoA. All submissions had to include two case studies, but those which submitted 15FTE or more had to include an additional case study for each additional 10FTE (or part thereof). So submissions of between 15FTE and 24.99FTE, required 3 case studies, those with between 25FTE and 34.99FTE required 4, and so on.

This threshold structure has had a clear effect on submission strategies; for example, 384 submissions were within 1FTE *below* a threshold for an extra case study, as compared to 52 being within 1FTE *above* a threshold. This is as to be expected due to cases where the potentially negative effect, on the overall profile, of including an additional (weaker) case study is viewed as a sufficiently high risk to negate the positive (QR) benefit of including a small number of additional staff. In situations where the extra case study is much weaker than those submitted, a smaller submission could even be better in terms of QR income.

Removing such a threshold effect is possible. A simple scheme would be to down-weight the influence of the weakest case study submitted, with the weight depending on how many FTE were submitted above the last threshold [C].

The above leads to Recommendation R2.1.

5.2 The REF 'Overall Quality Profiles'

The reported *overall quality profile* for each REF 2014 submission was calculated as a weighted average of the three quality *sub-profiles*, these sub-profiles representing the assessments of research *Outputs*, research *Impact* and research *Environment*. The assigned weights, which had been agreed and announced near the beginning of the REF 2014 development process, were 65% for *Outputs*, 20% for *Impact* and 15% for *Environment*.

It is amply evident in the results of REF 2014 that the three components *Outputs*, *Impact* and *Environment* were assessed quite differently by most or all of the REF expert panels. Some informal analysis of this can be found in [7], which shows that because of the relatively low variation between sub-profiles for Outputs, as compared with the substantially higher variation seen across Impact and Environment sub-profiles, the *actual* effect of *Outputs* on (rankings based upon) the reported overall quality profiles was markedly less than the announced weight of 65% for *Outputs* might be taken to imply. The *Impact* and *Environment* assessments correspondingly had *more* actual effect on the reported overall quality profiles, and resultant rankings, than their stated weights of 20% and 15% (respectively) would suggest. It is quite natural from a statistical perspective to expect that the assessment of a relatively small number of evidential items on *Impact* and *Environment*, versus a much larger set of published *Output*s, will tend to yield more homogeneous sub-profiles for *Outputs* than for the other two components.

HEFCE recently has recognised these differences in assessment of the three components, at least in terms of the implications for QR funding in 2015-16, by deciding in February 2015 (see [6]) to *re-weight* the components for QR funding purposes, in such a way that 65% of total QR funding is allocated based on the Outputs sub-profiles, 20% on Impact and 15% on Environment This differs from the QR allocations made since RAE 2008 up to funding year 2014-15, which have been based on the published *overall quality profiles*. (But note that the same phenomenon, i.e. substantially lower actual weight for Outputs than was apparently intended, had been clear also in the results of RAE 2008: see for example [8].)

HEFCE's effective re-weighting of the funding formula accounts for the expert panels' being more or less generous in their quality assessments of the three components, but not for the higher variation seen across quality sub-profiles for Impact and for Environment [D]. As currently constructed, using a fixed-weight average of the sub profiles, the overall profiles are neither the basis of the HEFCE funding allocation, nor do they produce overall rankings in a way that accounts for the differential spread of scores for the different sub profiles.

The above leads to Recommendation R2.2.

6. Recommendations

The recommendations are of two types, addressed respectively to the compilers of league tables and to the national funding councils. A common theme is that information to the public about research quality is best served by elimination, as far as possible, of the effects of 'strategic' decisions made by universities about their REF submissions.

R1. Published rankings should be more closely aligned with the QR funding outcomes per researcher, and this should include the funding outcomes also for any non-submitted staff. This would eliminate a strategic dilemma otherwise faced by universities, as well as providing a more direct interpretation of the published rankings.

Specifically, for the construction of *more informative* REF-based league tables:

R1.1 Research quality levels should be weighted *according to the weights used in the QR funding formula* that is in use at the REF submission date.

R1.2 In league tables the REF results should be *normalized*, to account for any non-submitted eligible research staff, by using the best available administrative data.

Taken together, R1.1 and R1.2 imply that league tables should be based on a suitably defined 'quality per researcher' measure, which differs in detail from the similar 'intensity' measure that is already in circulation. R1.2 makes sense if applied together with R1.1, as it is reasonable to assume that staff would only be excluded from REF if their university believed that their exclusion would not substantially reduce QR funding. Note that R1.1 is based on the principle of aligning decisions about maximising QR funding with those of maximising league table position, and *do not* imply any *statistical* justification for the HEFCE choice of QR formula.

R1.3 Institution-level league tables that are based on such a 'quality per researcher' measure should also publish the number of eligible research staff as a proportion of total academic staff, as contextual information to aid interpretation of the results.

Related recommendations to HEFCE and the other national funding councils are as follows.

R2. Consideration should be given to adjustments to the REF submission and assessment rules, to enhance the public-information value of REF outcomes.

Specifically:

R2.1 The distorting 'threshold' effect of REF 2014's Impact Case Study requirements should be eliminated, as far as possible. It is evident that this aspect of the REF 2014 rules had a substantial effect on universities' decisions about the number of eligible research staff to include in each department's REF submission.

R2.2 It should be considered whether the reported 'overall profiles', derived from the detailed REF assessments of individual departments, can be made more informative. As currently constructed, the overall profiles often yield rankings that are influenced more by the REF assessments of research *Impact* and research *Environment* than by the assessment of research *Outputs*.

References

[1] http://www.ref.ac.uk

[2] http://www.ref.ac.uk/about/

[3] Research Fortnight 4470 (Dec 18, 2014)
http://www.researchprofessional.com/0/dms/Paper-publication-PDFs/Research-Fortnight/2014/RF4470

[4] Times Higher Education (Dec 18, 2014) http://www.timeshighereducation.co.uk/features/ref-2014-results-by-subject/2017594.article

[5] Times Higher Education (Jan 1, 2015)
http://www.timeshighereducation.co.uk/features/ref-2014-rerun-who-are-the-game-players/2017670.article

[6] HEFCE funding letter, February 2015. http://www.hefce.ac.uk/pubs/year/2015/CL,032015/

[7] "The Impact of Impact." Blog post by S Oliver, University of Sussex, January 2015.
https://sebboyd.wordpress.com/2015/02/12/the-impact-of-impact/

[8] "RAE 2008: How much weight did research outputs actually get?" Blog post by D Firth, University of Warwick, February 2010. https://statgeek.wordpress.com/2010/02/07/rae-how

Notes

[A] The HESA data is available from https://www.hesa.ac.uk/ref2014. There it is made clear that definitions used for HESA do not match exactly that for REF. For example that research assistants are not included in the HESA data, although some research assistants are eligible for REF; that there are "known to be staff employed by the Colleges of Oxford, Cambridge and the University of the Highlands and Islands, and the Institute of Zoology who are eligible for submission to REF but are not included in the HESA Staff Record". The fact that the HESA data is only an estimate of the number of eligible staff can be seen by a number of cases where institutions submitted more staff to REF for a given UoA then the HESA staff numbers for that UoA. Note that the HESA data counts only those academic staff on research-only or research and teaching contracts, and institutions vary considerably in the proportion of academic staff on teaching-only contracts.

[B] The limitation of the research intensity measure is acknowledged in the article where is is presented:
"This method is not perfect. The major flaw is that it in effect gives a zero score to the research of anyone not submitted to the REF – which, in many cases, will clearly not reflect reality."
REF 2014 rerun: who are the 'game players'? by Paul Jump
http://www.timeshighereducation.co.uk/features/ref-2014-rerun-who-are-the-game-players/2017670.article

[C] For an example of such a scheme, assume that the FTE of a submission was *x* above the nearest threshold for adding an extra case study. The weakest one would be given a weight that is *x/10* of each of the other case studies. Thus for a submission just over the threshold, the penalty of submitting an extra case study would be proportionately smaller.

[D] The differences in both mean and spread for Outputs, Environment and Impact differ across UoAs. Here we will just demonstrate the issues for a single UoA: UoA10 Mathematical Sciences. The FTE-weighted means (based on the new QR formula) for the three components are 1.50

(*Outputs*), 1.90 (*Impact*) and 2.24 (*Environment*). The FTE-weighted standard deviations are 0.35 (*Outputs*), 0.93 (I*mpact*) and 1.30 (*Environment*).

To take account of the differences in standard deviation, we could standardise the scores for each component by dividing the scores for each institution for each component by the standard deviation of that component. We can then define an *effective weighting* of the three components as the weighting for these standardised scores that would give the same ranking as the actual weights when used for the original scores. This indicates the actual influence each component has on rankings after we take account for the differential spread.

Current league tables use a weight of 65%, 20% and 15% for the three components. When you take account of their difference in spread, this relates to an *effective weighting* of 37%, 31% and 32%.

The formula for QR funding, which takes account of the differential means of the three components, would use weights of 72% (*Outputs*), 17% (*Impact*) and 11% (*Environment*). These correspond to an effective weighting of 45% (*Outputs*) 29% (*Impact*) and 26% (*Environment*).

REF League Tables Working Group report
Published by the Royal Statistical Society, 11 May 2015