palgrave communications

HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

https://doi.org/10.1057/s41599-019-0233-x OPEN

Systematic analysis of agreement between metrics and peer review in the UK REF

V.A. Traag [1] & L. Waltman[1]

ABSTRACT When performing a national research assessment, some countries rely on citation metrics whereas others, such as the UK, primarily use peer review. In the influential *Metric Tide* report, a low agreement between metrics and peer review in the UK Research Excellence Framework (REF) was found. However, earlier studies observed much higher agreement between metrics and peer review in the REF and argued in favour of using metrics. This shows that there is considerable ambiguity in the discussion on agreement between metrics and peer review. We provide clarity in this discussion by considering four important points: (1) the level of aggregation of the analysis; (2) the use of either a size-dependent or a size-independent perspective; (3) the suitability of different measures of agreement; and (4) the uncertainty in peer review. In the context of the REF, we argue that agreement between metrics and peer review should be assessed at the institutional level rather than at the publication level. Both a size-dependent and a size-independent perspective are relevant in the REF. The interpretation of correlations may be problematic and as an alternative we therefore use measures of agreement that are based on the absolute or relative differences between metrics and peer review. To get an idea of the uncertainty in peer review, we rely on a model to bootstrap peer review outcomes. We conclude that particularly in Physics, Clinical Medicine, and Public Health, metrics agree relatively well with peer review and may offer an alternative to peer review.

[1] Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands. Correspondence and requests for materials should be addressed to V.A.T. (email: v.a.traag@cwts.leidenuniv.nl)

Introduction

Many countries have some form of a national Research Assessment Exercise (RAE) in which universities and other research institutions are evaluated (Hicks, 2012). In part, such assessments aim to account for the expenses of public funds, but sometimes they also function to distribute funds based on performance. Scientific quality or scientific impact plays a central role in many assessment exercises (Zacharewicz et al., 2018), but institutions may also be evaluated on other performance dimensions, such as their societal, cultural, and economic impact. Here, we restrict ourselves to scientific quality or scientific impact determined based on the publication output of an institution. However, we acknowledge that other dimensions may also play a critical role.

How the quality or impact of publications is assessed differs from country to country. Some countries have a national RAE that is driven by citation metrics, whereas others rely on peer review (Hicks, 2012). In particular, the United Kingdom (UK) has a long tradition of research assessment that relies on peer review, starting with the first assessment exercise in 1986. The latest assessment exercise, referred to as the Research Excellence Framework (REF), took place in 2014. It was followed by a detailed report, known as the *Metric Tide* report (Wilsdon et al., 2015), that critically examined the possible role of citation metrics in the REF. It concluded that "[m]etrics should support, not supplant, expert judgement" (Wilsdon et al., 2015, p. viii). To support this conclusion, the report provided statistical evidence of the lack of agreement between metrics and peer review. Here, we re-examine the statistical evidence for this conclusion. The *Metric Tide* report also offered other arguments to support the above conclusion. It argued that metrics are contested among academics, and should therefore not be used, whereas peer review commands widespread support. Moreover, metrics may create perverse incentives. We do not consider these arguments further in this paper, and restrict ourselves to the statistical argument presented in the *Metric Tide*. Of course, the other arguments should play a role in the broader discussion on the relative merits of peer review and metrics.

The various assessment exercises carried out in the UK during the past decades have all been accompanied by papers that compare citation metrics and peer review. Although the results vary from field to field, most studies found correlations of about 0.7. Some authors obtained higher correlations, on the order of 0.9. However, the *Metric Tide* report found significantly lower correlations in the range of about 0.2–0.4. Interestingly, even when authors obtained similar correlations, they did not always draw the same conclusion. Some, such as Mryglod et al. (2015b) and Mahdi et al. (2008), argued that a correlation of 0.7 is too low to consider using metrics, while others, such as Thomas and Watkins (1998) and Taylor (2011), argued that a correlation of 0.7 is sufficiently high.

We try to provide clarity in this debate by considering four important points:

1. The agreement between metrics and peer review depends on the level of aggregation. The level of individual publications constitutes the lowest level of aggregation. The level of researchers and the level of research institutions represent higher levels of aggregation.
2. At aggregate levels, metrics and peer review may take a size-dependent perspective—scaling with the size of an institution—or a size-independent perspective—being independent of the size of an institution. This distinction is particularly relevant when reporting correlations.
3. Correlations between metrics and peer review may not be the most informative measure of agreement. Other measures may be more appropriate.
4. Peer review is subject to uncertainty. This should be taken into consideration when interpreting the agreement between metrics and peer review.

We first briefly discuss the REF and consider its objectives. This is followed by a review of the literature on comparing metrics and peer review in the context of the REF and its precursors. We argue that in the REF context, proper comparisons between metrics and peer review should be made at the institutional level, not at the level of individual publications. We also briefly discuss how a size-dependent perspective relates to a size-independent perspective. As we show, size-dependent correlations can be high even if the corresponding size-independent correlations are low. We then introduce two measures of agreement that we consider to be more informative than correlations. One measure is especially suitable for the size-dependent perspective, while the other measure is more suitable for the size-independent perspective. To get some idea of the uncertainty in peer review, we introduce a simple model of peer review.

Based on our analysis, we conclude that for some fields, the agreement between metrics and peer review is similar to the internal agreement of peer review. This is the case for three fields in particular: Clinical Medicine, Physics, and Public Health, Health Services & Primary Care. Finally, we discuss the implications of our findings for the REF 2021 that is currently in preparation.

UK Research Excellence Framework

The UK REF has three objectives:

1. To provide accountability for public investment in research and produce evidence of the benefits of this investment.
2. To provide benchmarking information and establish reputational yardsticks, for use within the [Higher Education] sector and for public information.
3. To inform the selective allocation of funding for research.

From http://www.ref.ac.uk/about/whatref/ for REF 2021.[1]

In addition, three further roles that the REF fulfills were identified:

4. To provide a rich evidence base to inform strategic decisions about national research priorities.
5. To create a strong performance incentive for HEIs and individual researchers.
6. To inform decisions on resource allocation by individual HEIs and other bodies.

From https://www.ref.ac.uk/media/1050/ref2017_01.pdf

To meet these objectives, the REF assesses institutions in terms of (1) research output, (2) societal impact of the research, and (3) the environment supporting the research. Here, we are concerned only with the assessment of research output. In the REF 2014, the assessment of research output accounted for 65% of the overall assessment of institutions. Each output evaluated in the REF 2014 was awarded a certain number of stars: four stars indicates world-leading research, three stars indicates internationally excellent research, two stars indicates internationally recognised research, and one star indicates nationally recognised research.

The three above-stated objectives are each addressed in a different way. The overall *proportion* of high-quality research that has been produced is relevant for the first objective. Indeed, the REF 2014 website boasts that 30% of the submitted UK research was world-leading four-star research: public investment results in high-quality science. The *proportion* of research outputs awarded a certain number of stars also provides a reputational yardstick for institutions and thereby serves the second objective. Indicators based on these proportions feature in various league tables

constructed by news outlets such as The Guardian and Times Higher Education. Such indicators may influence the choice of students and researchers regarding where to study and perform research. The *total number* of publications that were awarded four or three stars influences the distribution of funding, which is relevant for the third objective of the REF.

Hence, the objective of establishing a reputational yardstick corresponds to a size-independent perspective, while the objective of funding allocation corresponds to a size-dependent perspective. This means that agreement between metrics and peer review is relevant from both perspectives. We will comment in more detail on the distinction between the two perspectives in the section "Size-dependent and size-independent perspectives".

To provide an indication of the importance of the REF 2014, we briefly look at the funding of UK higher education in 2017–2018[2]. In 2017–2018, REF results based on research output were used by the Higher Education Funding Council for England (HEFCE) to allocate £685M to institutions. Although many details are involved (e.g. extra funding for the London region, weighing cost-intensive fields), this was based largely on 4* and 3* publications, which were awarded roughly 80% and 20% of the money, respectively. This amounted to about £10,000 per 4* publication and about £2000 per 3* publication per year on average[3]. The total amount of about £685M allocated through the evaluation of research output represented about 20% of the total budget of HEFCE of £3602M and about 40% of the total research budget of HEFCE of £1606M. As such, it is a sizeable proportion of the total budget.

Literature review

We review previous literature on how metrics compare to peer review in previous RAEs in the UK. We then briefly review literature that analyses how metrics and peer review compared in the REF 2014.

Research assessment exercise. In 1986, the University Grants Committee (UGC) undertook the first nationwide assessment of universities in the UK, called the research selectivity exercise. Its primary objective was to establish a more transparent way of allocating funding, especially in the face of budget cuts (Jump, 2014). Only two years later, Crewe (1988) undertook the first bibliometric comparison of the results for Politics departments in the first 1986 exercise. The results of the 1986 exercise were announced per cost centre (resembling somewhat a discipline or field) of a university in terms of four categories: outstanding, above average, about average, and below average. This limited the possibilities for bibliometric analysis somewhat, and Crewe (1988) only made some basic comparisons based on the number of publications. He concluded that "there is a close but far from perfect relationship between the UGC's assessment and rankings based on publication records" (Crewe, 1988, p. 246). Indeed, later exercises also showed that higher ranked institutions are typically larger (in terms of either staff or publications). In the same year, Carpenter et al. (1988) analysed Physics and Chemistry outcomes of the UGC exercise. They compared the outcomes to a total influence score, a type of metric similar to the Eigenfactor (Bergstrom, 2007), and found a correlation of 0.63 for Physics and 0.77 for Chemistry. The total influence score used by Carpenter et al. (1988) is size-dependent, and the average influence per paper showed a correlation of only 0.22 and 0.34 for Physics and Chemistry, respectively. It is not clear whether the 1986 UGC results themselves were size-dependent or size-independent.

The next research selectivity exercise in 1989 was undertaken by the Universities Funding Council (UFC). The exercise made some changes and allowed universities to submit up to two publications per research staff (Jump, 2014). As an exception to the rule, the 1989 exercise was never used in any bibliometric study that compared the peer review results to metrics (although there were other analyses; see, for example, Johnes et al., 1993).

The 1992 exercise—then called the RAE—sparked more bibliometric interest. In addition to allowing two publications to be nominated for assessment by the institutions, the exercise also collected information on the total number of publications (Bence and Oppenheim, 2005). No less than seven studies appeared that compared the outcomes of the 1992 RAE to bibliometrics. Taylor (1994) analysed Business & Management and found a clear correlation[4] based on journal publications ($R^2 \approx 0.8$, $R \approx 0.9$). Oppenheim (1995) analysed Library & Information Management, and two years later, Oppenheim (1997) considered Anatomy, Archaeology, and Genetics. These two studies used both total citation counts and average citation counts per staff and found clear correlations on the order of 0.7–0.8 for both size-dependent and size-independent metrics and all analysed fields. Only for Anatomy, the size-independent metric was less clearly aligned with peer review outcomes, with a correlation of $R \approx 0.5$. Lim Ban Seng and Willett (1995) also analysed Library & Information Management and found even higher correlations on the order of 0.9 using both average citations and total citations. The correlation found by Colman et al. (1995) for Politics was lower, at only 0.5, where they used the number of publications in high impact journals per staff as a metric. Finally, Thomas and Watkins (1998) analysed Business & Management Studies using a journal-based score and found a correlation of 0.68. For the 1992 exercise, overall, both size-dependent and size-independent metrics correlated reasonably well with peer review in quite a number of fields. Most authors recommended that the RAE should take metrics into account, for example, as an initial suggestion, which can then be revised based on peer review.

In the 1996 RAE, full publication lists were no longer collected (Bence and Oppenheim, 2005). In 2001, results were announced as rankings, and institutions also received an overall score of 1–5*. Smith et al. (2002) analysed both the 1996 and the 2001 RAE and found a correlation on the order of 0.9 for the average number of citations in Psychology for both exercises. Clerides et al. (2011) also analysed both the 1996 and the 2001 RAE and found a correlation of about $R \approx 0.7$ ($R^2 \approx 0.5$) using the total number of high impact journal articles. Norris and Oppenheim (2003) analysed Archaeology and found correlations of about 0.8 for both size-dependent and size-independent metrics. Mahdi et al. (2008) analysed all units of assessments (UoAs; i.e. fields) and found that a number of fields showed substantial correlations on the order of 0.7–0.8 (e.g. Clinical Lab. Sciences, Psychology, Biological Sciences, Chemistry, Earth Sciences, and Business & Management) using the average number of citations per paper. Adams et al. (2008) also analysed the 2001 RAE results, although their focus was on which granularity of field-normalised citations works best. They found a reasonably high correlation of about 0.7 for Psychology, 0.6 for Physics, and only 0.5 for Biological Sciences. Finally, Butler and McAllister (2009) found a reasonable correlation ($R^2 \approx 0.5$–0.6, $R \approx 0.7$–0.8) for Political Science using the average number of citations.

In 2008, the results of the RAE were more structured. Rather than providing overall scores for institutions per UoA, a so-called quality profile was provided[5]. The quality profile offered more detailed information on the proportion of outputs that were awarded 1–4 stars. This enabled a more detailed analysis, since the measure was much more fine grained than the overall outcome. In addition, it allowed a clear distinction between size-dependent and size-independent results. Previously, only the overall results were announced, and the extent to which the

results were size-dependent or size-independent was unclear. Most studies found that larger institutions generally did better in RAEs, implying a certain type of size-dependent component, but this was never entirely clear. From 2008 onwards, the results were announced as a proportion of outputs that were awarded a certain number of stars, which was unambiguously size-independent.

Norris and Oppenheim (2010) examined Library & Information Science, Anthropology, and Pharmacy in the 2008 RAE using the *h*-index (and a variant thereof) and total citation counts. They compared this to a weighted average of the results multiplied by the number of staff, clearly a size-dependent metric. Norris and Oppenheim (2010) found a correlation of about 0.7 for Pharmacy, while Library & Information Science showed a correlation of only about 0.4, and Anthropology showed even a negative correlation. Taylor (2011) analysed Business & Management, Economics & Econometrics, and Accounting & Finance. They relied on a journal list from UK business schools to determine the proportion of publications in top journals and found a quite high correlation ($R^2 \approx 0.64$–0.78, $R \approx 0.80$–0.88) with the average rating. Kelly and Burrows (2011) found a clear correlation ($R^2 = 0.83$, $R = 0.91$) for Sociology. They also used the proportion of publications in top journals and compared it to a weighted average of RAE results. Mckay (2012) found that most scholars in the field of Social Work, Social Policy & Administration did not necessarily submit their most highly cited work for evaluation. This study did not explicitly report how well citations match peer review. Allen and Heath (2013) replicated the study of Butler and McAllister (2009) of Politics & International Studies and found a similar correlation ($R^2 \approx 0.7$, $R \approx 0.85$). They correlated the proportion of publications in top journals with the proportion of publications that obtained four stars, which are both clearly size-independent measures.

In two publications, Mryglod et al. (2013a, b) explicitly studied size-dependent correlations versus size-independent correlations in seven fields (Biology, Physics, Chemistry, Engineering, Geography & Environmental Science, Sociology, and History). They studied the average normalised citation score and the total normalised citation score and examined how they correlate with the RAE Grade (a weighted average of scores) and the RAE Score (the RAE Grade times the number of staff), respectively. They found size-independent correlations of only about 0.34 for History and Engineering and up to about 0.6 for Biology and Chemistry. The size-dependent correlations were substantially higher and reached about 0.9 for all fields. We discuss this in more detail in the section "Size-dependent and size-independent perspectives".

In conclusion, most studies in the literature have found correlations on the order of 0.6–0.7 for fields that seem to be amenable to bibliometric analysis. The conclusions that were drawn from such results nonetheless differed. Three types of conclusions can be distinguished. First, some authors concluded simply that the observed correlation was sufficiently high to replace peer review by metrics. Others concluded that peer review should be supported by citation analysis. Finally, some concluded that peer review should not be replaced by metrics, even though they found relatively high correlations. This indicates that different researchers draw different conclusions, despite finding similar correlations. One problem is that none of the correlations are assessed against the same yardstick; thus, it is unclear when a correlation should be considered "high" and when it should be considered "low".

Research Excellence Framework 2014. The REF 2014 was accompanied by an extensive study into the possibilities of using metrics instead of peer review, known as the *Metric Tide* report (Wilsdon et al., 2015). This report concluded that citations should only supplement, rather than supplant, peer review. One of the arguments for this conclusion was based on an analysis of how field-normalised citations based on Scopus data correlate with peer review. The report found correlations[6] in the range of about 0.2–0.4. This is quite low compared with most preceding studies, which found correlations of roughly 0.6–0.7. In contrast to preceding studies, Wilsdon et al. (2015, Supplementary Report II) analysed the correlation between metrics and peer review at the level of individual publications rather than at some aggregate level. This is an important difference that we revisit in the section "Level of aggregation".

The REF 2014 results were also analysed by Mryglod et al. (2015a, b) at the institutional level. They found that the departmental *h*-index was not sufficiently predictive, even though an earlier analysis suggested that the *h*-index might be predictive in Psychology (Bishop, 2014). An analysis by Elsevier found that metrics were reasonably predictive of peer review outcomes at an institutional level in some fields but not in others (Jump, 2015).

Both Pride and Knoth (2018) and Harzing (2017) compared the UK REF results with metrics using Microsoft Academic Graph (Harzing and Alakangas, 2017). Pride and Knoth (2018) compared the median number of citations with the REF GPA, which is a weighted average of the proportion of publications that have been awarded a certain number of stars for all UoAs, clearly taking a size-independent perspective. They found correlations on the order of 0.7–0.8 for the UoAs that showed the highest correlations. Harzing (2017) compared the total number of citations and a so-called REF power rating, taking a size-dependent perspective, and found a very high correlation of 0.97. This correlation was obtained at an even higher aggregate level, namely, the overall institutional level, without differentiating between different disciplines. She found similarly high correlations when studying Chemistry, Computer Science, and Business & Management separately. The high correlations can be partly explained by the use of a size-dependent perspective. We comment on this in the section "Size-dependent and size-independent perspectives".

Data and methods

The REF 2014 provides a well-documented dataset of both the evaluation results and the submitted publications that have been evaluated[7]. The REF 2014 has different scores for different profiles: "output", (societal) "impact", and "research environment". Only the "output" profile is based on an evaluation of the submitted publications. The others are based on case studies and other (textual) materials. We restrict ourselves to the REF 2014 scores in the output profile, and we compare them with citation metrics.

We match publications to the CWTS in-house version of the Web of Science (WoS) through their DOI. We use the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. Most publications are articles (type 'D' in the REF 2014 dataset), but the publications also include books, conference proceedings, and other materials. In total, 190,962 publications were submitted, of which 149,616 have an associated DOI, with 133,469 of these being matched to the WoS. Overall, the WoS covers about two-thirds of all submitted publications. Some fields are poorly covered in the WoS, such as the arts and humanities, having a coverage of only about 10–30% of submitted publications, whereas the natural sciences generally have a high coverage of 90–95% (see Supplemental Material, Table A.1 for an overview). In the calculation of citation metrics, we take into account only publications covered in the WoS. In the calculation of statistics based on peer review, all

publications submitted to the REF are considered, including those not covered in the WoS.

All matched publications are associated with a particular UoA, which roughly corresponds to a field or discipline. The REF 2014 distinguished 36 UoAs. Every publication was submitted on behalf of a particular institution. Some publications were submitted in multiple UoAs, and we take them into account in each UoA. Publications that were co-authored and submitted by multiple institutions may thus be counted multiple times. Publications co-authored by several authors from the same institution were sometimes submitted multiple times in the same UoA by the same institution[8]. We consider only the unique publications of an institution in a UoA. In other words, we count a publication only once, even if it was submitted multiple times in the same UoA by the same institution.

Some institutions can have separate submission headings in the same UoA to differentiate more fine-grained subjects. For example, Goldsmiths' College separately submits publications for Music and Theatre & Performance in the overall UoA of Music, Drama, Dance & Performing Arts. The results of such separate submissions are also announced separately, and we therefore also consider them to be separate submissions.

We consider citations coming from publications up to and including 2014, which is realistic if metrics had actually been used during the REF itself. For this reason, we exclude 365 publications that were officially published after 2014 (although they may have already been available online). We use about 4000 micro-level fields constructed algorithmically on the basis of citation data (Waltman and van Eck, 2012; Ruiz-Castillo and Waltman, 2015) to perform field normalisation. Citations are normalised on the basis of publication year and field, relative to all publications covered in the WoS.

We calculate how many 4^* publications correspond to how many top 10% publications per UoA (see Table A.1). This can differ quite substantially from one UoA to another. For example, Clinical Medicine shows 0.57 4^* publications per top 10% publication, whereas Mathematical Sciences show 1.18 4^* publications per top 10% publication. This suggests that what is considered as 4^* publication in peer review differs per field, where some fields seem to use more stringent conditions than others. Similarly, Wooding et al. (2015) found that peer review was less stringent in REF 2014 than in REF 2008, in what publications were considered worthy of 4^*.

Before presenting our results, we first address four important methodological considerations. We start by reflecting on the level of aggregation at which agreement between metrics and peer review should be analysed. We then examine both the size-dependent and size-independent perspectives, especially regarding correlations. This leads us to consider alternative measures of agreement. Finally, we discuss the matter of peer review uncertainty.

Level of aggregation. The *Metric Tide* report analysed agreement between metrics and peer review at the level of individual publications. We believe that this is not appropriate in the context of the REF, and it may explain the large differences between the *Metric Tide* report and preceding publications in which agreement between metrics and peer review was analysed. The institutional level is the appropriate level to use for the analysis. The analysis at the level of individual publications is very interesting. The low agreement at the level of individual publications supports the idea that metrics should generally not replace peer review in the evaluation of a single individual publication. However, the goal of the REF is not to assess the quality of individual publications but rather to assess "the quality of research in UK higher education institutions"[9]. Therefore, the question should not be whether the evaluation of individual publications by peer review can be replaced by the evaluation of individual publications by metrics but rather whether the evaluation of institutions by peer review can be replaced by the evaluation of institutions by metrics. Even if citations are not sufficiently accurate at the individual publication level, they could still be sufficiently accurate at the aggregate institutional level; the errors may 'cancel out'. For this reason, we perform our analysis at the institutional level. We calculate citation metrics per combination of an institution and a UoA.

Size-dependent and size-independent perspectives. As briefly discussed earlier, the REF has multiple objectives. It aims to provide a reputational yardstick, but it also aims to provide a basis for distributing funding. A reputational yardstick is usually related to the average scientific quality of the publications of an institution in a certain UoA. As such, a reputational yardstick is *size-independent*: it concerns an average or percentage, not a total, and it does not depend on the size of an institution. In the REF, funding is related to the total scientific quality of the publications of an institution in a certain UoA. As such, funding is *size-dependent*: institutions with more output or staff generally receive more funding. Of course, quality also affects funding: institutions that do well receive more funding than equally sized institutions that do less well. Both the size-dependent and size-independent perspectives are relevant to the REF. We therefore believe that both perspectives are important in deciding whether metrics can replace peer review.

Many studies of the REF and its predecessors have analysed either size-dependent or size-independent correlations. Size-dependent correlations are typically much higher than size-independent correlations. For example, Mryglod et al. (2013a, b) found size-dependent correlations on the order of 0.9 but much lower correlations for size-independent metrics. Similarly, Harzing (2017) found a very high size-dependent correlation.

Size-dependent correlations can be expected to be larger in general. Let us make this a bit more explicit. Suppose we have two size-independent metrics x and y (e.g. metrics and peer review), where n denotes the total size (e.g. number of publications or staff). The two size-dependent metrics would then be xn and yn. Then, even if x and y are completely independent from each other, and hence show a correlation of 0, the two size-dependent metrics xn and yn may show a quite high correlation. This is demonstrated in Fig. 1, where x and y are two independent uniform variables and n is a standard log-normal variable

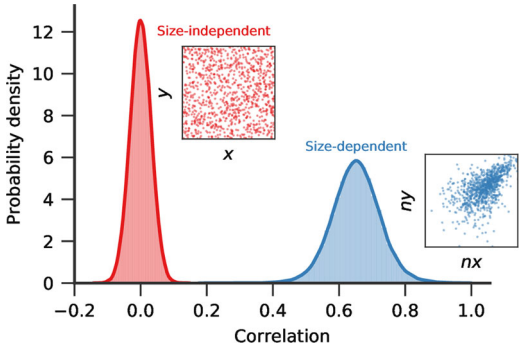

Fig. 1 The correlation between two size-dependent metrics can be quite high even if the corresponding size-independent metrics are completely uncorrelated. The insets show the scatter plots of size-dependent and size-independent metrics. For the size-independent scatter plot logarithmic scales are used

(1000 samples). In this example, the Pearson correlation between *xn* and *yn* may be as high as 0.7–0.8. In other words, the fact that *xn* and *yn* may show a high correlation may be completely explained by the common factor *n*. A similar observation has already been made before in bibliometrics (West and Bergstrom, 2010), and related concerns were already raised by Pearson as early as 1896.

This complicates the interpretation of size-dependent correlations. A high size-dependent correlation may be due to *x* and *y* being strongly correlated, but it may also be due to *n* having a high variance. The higher the variance of *n*, the higher the size-dependent correlation. In fact, if *n* is distributed according to a log-normal distribution with a very large variance, the size-dependent correlation will be close to 1, regardless of the extent to which *x* and *y* are correlated. The strength of the size-dependent correlation then mainly reflects the variance of the size of institutions.

In our analysis, we consider both a size-dependent and a size-independent perspective. We calculate the proportion of publications that belong to the top 10% most highly cited publications in their field and year, which we call the PP (top 10%). In addition, we use PP(4^*) to denote the proportion of publications with a 4^* rating in the REF. The PP(top 10%) and PP(4^*) are similar in spirit[10]. They aim to identify whether publications have a high impact or are of high quality ("world leading"), respectively. Other citation metrics, such as those based on average normalised citation counts, are more difficult to translate into a 4^* rating system. Both the PP(4^*) and the PP (top 10%) are clearly size-independent. We calculate the total number of 4^* rated outputs, called the P(4^*), by multiplying the PP(4^*) by the number of submitted outputs. Similarly, we obtain the total number of top 10% outputs, called the P(top 10%), by multiplying the PP(top 10%) by the number of submitted publications in the WoS. Both the P(4^*) and the P(top 10%) are clearly size-dependent.

Measures of agreement. Agreement between metrics and peer review can be measured using a variety of measures. For example, the *Metric Tide* report employs measures, such as precision and sensitivity, which are well suited for the individual publication level. Most analyses of the REF and its predecessors employ correlation coefficients. As we argued in the previous section, correlations may be difficult to interpret when taking a size-dependent perspective. Moreover, correlations provide little intuition of the size of the differences between metrics and peer review. For this reason, we consider two different measures (see Supplemental Material, Section A for details): the median absolute difference (MAD) and the median absolute percentage difference (MAPD).

The MAD gives an indication of the absolute differences that we can expect when switching from peer review to metrics. We believe that this measure is especially informative when taking a size-independent perspective. For example, if an institution has a PP(4^*) of 30% and the MAD is 3 percentage points, then in half of the cases switching to metrics would yield an outcome equivalent to a PP(4^*) between 27% and 33%. The idea of the MAD is that an increase or decrease of 3 percentage points would likely be of similar interest to institutions with different PP(4^*) scores. That is, if one institution has a PP(4^*) of 50% and another has a PP(4^*) of 30%, a difference of 3 percentage points would be of similar interest to both.

This is quite different for the size-dependent perspective. The size of institutions varies much more than the proportion of 4^* publications of institutions. As such, a certain absolute difference will probably not be of the same interest to different institutions when taking a size-dependent perspective. For example, in terms of funding, if we report an absolute difference of £10,000, this would be of major interest to institutions receiving only £20,000, but probably not so much for institutions receiving £1,000,000. From this point of view, the MAPD can be considered more appropriate, as it gives an indication of the relative differences that we can expect when switching from peer review to metrics. The idea of MAPD is that an increase or decrease of 10% would likely be of similar interest to both small institutions that receive little funding and large institutions that receive much funding. The MAPD is the same for both size-dependent and size-independent metrics, since the common factor falls out in the calculation (see Supplemental Material, Section A for details).

Peer review uncertainty. Regardless of the measure of agreement, the perspective (i.e. size-independent or size-dependent), and the level of aggregation, it is important to acknowledge that peer review is subject to uncertainty. Hypothetically, if the REF peer review had been carried out twice, based on the same publications but with different experts, the outcomes would not have been identical. This is what we refer to as peer review uncertainty. It is sometimes also called internal peer review agreement. Evidence from the Italian research assessment exercise, known as the VQR, suggests that peer review uncertainty is quite high (Bertocchi et al., 2015). Unfortunately, detailed peer review results of the REF at the publication level are not available. Also, the *Metric Tide* report (Wilsdon et al., 2015) did not quantify internal peer review agreement, which could have served as a baseline for our study. Internal peer review agreement in the REF has not been investigated in other publications either, although peer review in the REF has been studied from other perspectives (e.g. Derrick, 2018).

To quantify peer review uncertainty and get an idea of the order of magnitude of the agreement that we can expect in peer review itself, we perform a type of bootstrap analysis (see Supplemental Material, Section B for details). Since we do not know exactly the degree of uncertainty in peer review, we consider two scenarios, one with low uncertainty ($\sigma_\epsilon^2 = 0.1$, see Supplemental Material, Section B) and one with high uncertainty $\left(\sigma_\epsilon^2 = 1\right)$. The results presented in the next section are based on 1000 bootstrap samples. We report both the median outcome obtained from 1000 samples and the interval that covers 95% of the outcomes.

Results

We now describe the results from our analysis. Our analysis compares the agreement between metrics and peer review with the internal agreement of peer review, based on a simple model of peer review. For simplicity, we consider only 4^* publications, as they are deemed four times more valuable than 3^* publications in the REF. We first describe our results from the size-independent perspective and then turn to the size-dependent perspective. All necessary replication materials have been deposited at Zenodo (Traag and Waltman, 2018) and can be accessed at https://github.com/vtraag/replication-uk-ref-2014.

Size-independent perspective. To facilitate comparison with earlier studies, we first discuss our results in terms of Pearson correlations. We find that Economics & Econometrics, Clinical Medicine, Physics, Chemistry, and Public Health show a high size-independent Pearson correlation between the percentage of 4^* rated submissions and the percentage of top 10% publications: Pearson correlations are higher than 0.8 (see Fig. 2 and Supplemental Material, Table C.1). A number of other fields show correlations on the order of 0.7, which is in line with previous

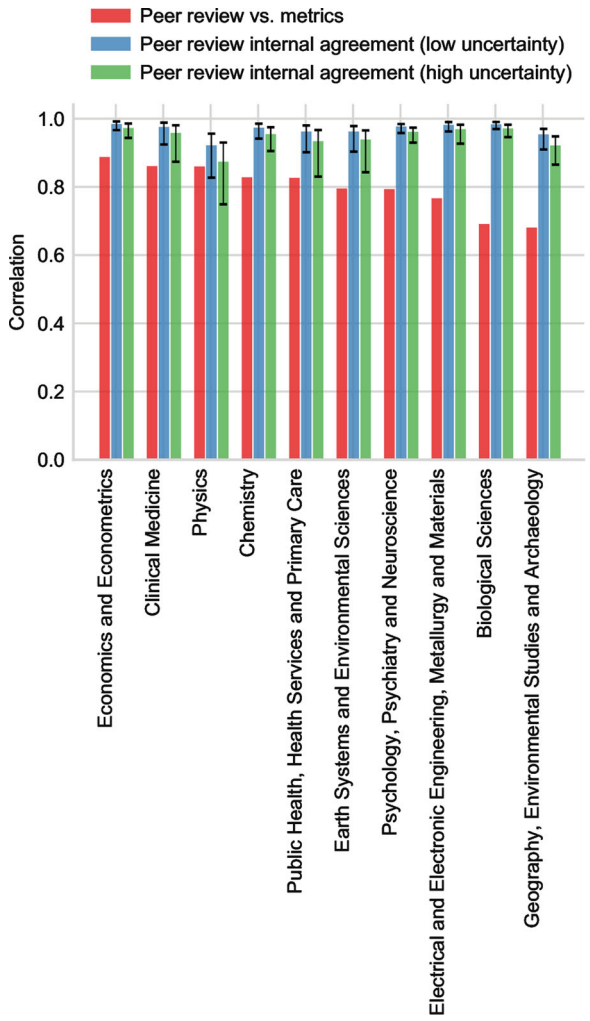

Fig. 2 Size-independent correlation between PP(top 10%) and PP(4*) compared with correlations based on a model of peer review uncertainty. Results are shown only for the 10 units of assessment with the highest correlation between metrics and peer review

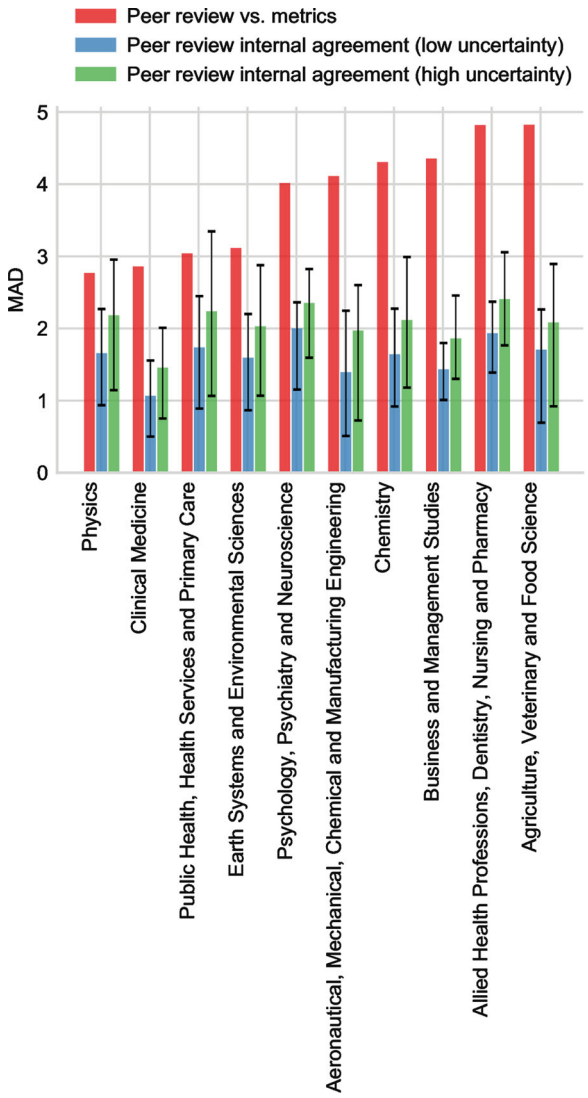

Fig. 3 Size-independent median absolute difference (MAD) between PP (top 10%) and PP(4*) compared with the MAD based on a model of peer review uncertainty. Results are shown only for the 10 units of assessment with the lowest MAD between metrics and peer review

studies on earlier rounds of the RAE/REF. These correlations are much higher than the correlations found by the *Metric Tide* report (Wilsdon et al., 2015).

Our results strongly differ from the analysis by Elsevier of the REF results (Jump, 2015), even though it also found some relatively strong correlations. In particular, the analysis found correlations for Physics and Clinical Medicine on the order of 0.3. Public Health did a little better, but still the correlation was only about 0.5. Finally, Biology had the single highest correlation of about 0.75, whereas this correlation is much lower in our results. It may be of interest to compare the different results in more detail and to better understand why Elsevier's results (Jump, 2015) differ from ours. The differences most likely stem from the use of all publications of an institution versus only the publications submitted to the REF. Another reason for the differences may be the use of different databases (Scopus vs. WoS) and the use of different field classification systems in the field-normalised citation metrics. The citation metrics of Jump (2015) were normalised on the basis of the journal-based classification system of Scopus, whereas we normalised on the basis of a detailed publication-based classification system (Ruiz-Castillo and Waltman, 2015).

The results of the peer review uncertainty may be surprising (see Fig. 2). Although the bootstrapped correlations are almost always higher than the correlations of the REF results with the PP(top 10%), the differences are sometimes small. Most notably, Physics shows a correlation between metrics and peer review of 0.86, which is on par with the bootstrapped correlations, especially for high peer review uncertainty. This indicates that for Physics, metrics work at least equally well as peer review, assuming some uncertainty in peer review. For Economics, Clinical Medicine, Chemistry, and Public Health, the correlations between metrics and peer review are lower than the bootstrapped correlations, but the differences are not very large. Hence, the metrics correlate quite well with peer review for these fields. Other UoAs show correlations between metrics and peer review that are substantially lower than the correlations obtained using the bootstrapping procedure.

The MAD provides a more intuitive picture of what these correlations mean in practice (see Fig. 3 and Supplemental Material, Table C.1). In the interpretation of the MAD, it is important to keep in mind that overall about 30% of the

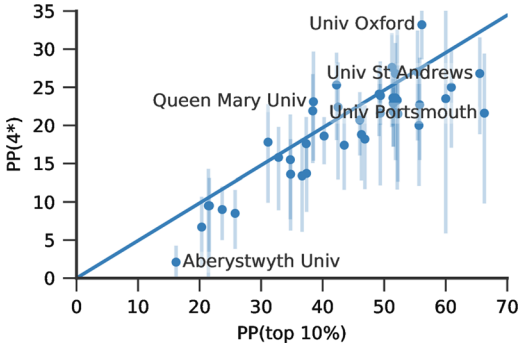

Fig. 4 Scatter plot of PP(top 10%) and PP(4*) at the institutional level for Physics. Error bars indicate the 95% interval of bootstrapped peer review results for low peer review uncertainty. The solid line indicates the proportion of 4* publications considered to be equivalent to a given proportion of top 10% publications (see Supplemental Material, Section A for details)

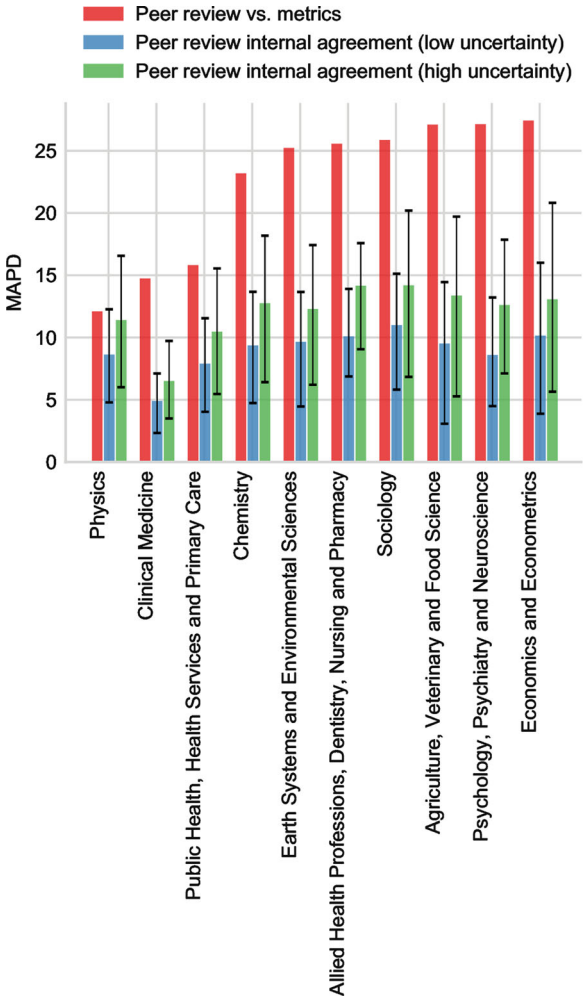

Fig. 5 Size-dependent median absolute percentage difference (MAPD) of P (top 10%) relative to P(4*) compared with the MAPD based on a model of peer review uncertainty. Results are shown only for the 10 units of assessment with the lowest MAPD of metrics relative to peer review

publications have been awarded 4* in the REF. The MAD in Physics reaches almost 3 percentage points in PP(4*) when switching from peer review to metrics. This is just somewhat more than 1 percentage point higher than the median bootstrapped MAD for low peer review uncertainty and less than 1 percentage point higher than the median bootstrapped MAD for high peer review uncertainty. Hence, in Physics, the difference between metrics and peer review seems to be just slightly larger than the difference between different peer review exercises. Moreover, for high peer review uncertainty, the difference between metrics and peer review still falls within the 95% interval of bootstrapped peer review results. This means that it is possible that the difference between metrics and peer review is of a similar magnitude as the difference between different peer review exercises. In Clinical Medicine, we also find an MAD of almost 3 percentage points in PP(4*) when switching to metrics, although in this UoA the difference with the bootstrapped MADs is more substantial. In Public Health, the MAD is slightly higher than 3 percentage points in PP(4*). The difference with the bootstrapped MADs is not very large, and for high peer review uncertainty, it falls within the 95% interval of bootstrapped peer review results. For other fields, we observe that the MAD when switching from peer review to metrics is higher than the bootstrapped MADs, but for many of these fields, the MAD may still be considered to be relatively small (e.g. < 5 percentage points). On the other hand, there are also fields for which the MAD is quite large (see Supplemental Material, Fig. C.1 for the MADs for all UoAs). These are especially fields that are not well covered in the WoS.

Looking at the result for Physics in more detail, we see that most institutions have bootstrapped peer review results that agree reasonably well with metrics (see Fig. 4, see Supplemental Material, Fig. C.2 for all UoAs). However, some larger differences remain. University of Oxford and Queen Mary University are systematically valued more highly by peer review than by metrics. Conversely, University of St. Andrews, University of Portsmouth, and Aberystwyth University are systematically valued less highly by peer review than by metrics.

Size-dependent perspective. As expected, the size-dependent correlations are much higher than the size-independent correlations (see Supplemental Material, Table C.1). Half of all UoAs reach correlations higher than 0.9. Some have very high size-dependent correlations, even when the size-independent correlations are low, as previously explained in the section "Size-dependent and size-independent perspectives". For example, Mathematical Sciences shows a size-dependent correlation of 0.96, whereas the size-independent correlation is only 0.39. As discussed above, we believe the correlations are not so informative for the size-dependent perspective, and we therefore focus on the MAPD.

Peer review uncertainty leads to MAPDs of somewhere between 10% and 15% for many fields (see Fig. 5). Hence, peer review uncertainty may have a substantial effect on the amount of funding allocated to institutions. Comparing peer review with metrics, we find that Physics has an MAPD of 12%, which is similar to what can be expected from peer review uncertainty. Clinical Medicine has an MAPD of almost 15%, which is substantially higher than the MAPD resulting from peer review uncertainty. Likewise, Public Health has an MAPD of about 16%, which is higher than the expectation from peer review uncertainty. Other fields show MAPDs between metrics and peer review that are above 20%, especially fields that are not well covered in the WoS (see Supplemental Material, Fig. C.3). The 10 UoAs with the lowest MAPDs all show size-dependent correlations close to or above 0.9, which illustrates how correlations and MAPDs may potentially lead to different conclusions. Biology is a

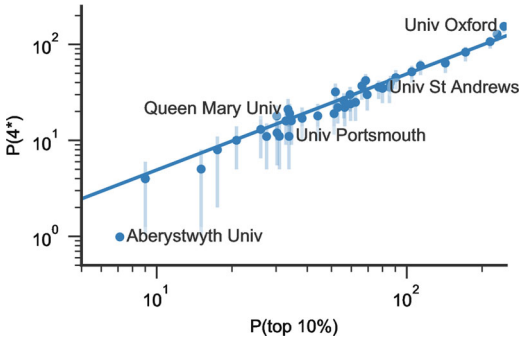

Fig. 6 Logarithmic scatter plot of P(top 10%) and P(4*) at the institutional level for Physics. Error bars indicate the 95% interval of bootstrapped peer review results for low peer review uncertainty. The solid line indicates the number of 4* publications considered to be equivalent to a given number of top 10% publications (see Supplemental Material, Section A for details)

clear example: it has a size-dependent correlation of 0.98, yet it has an MAPD of 32%.

The MAPD summarises the overall differences, but for individual institutions, the differences can be substantially larger or smaller. We again consider Physics in somewhat more detail (see Fig. 6, see Supplemental Material, Fig. C.4 for all UoAs). Although the absolute differences are sometimes difficult to discern in Fig. 6, some of the institutions that we already encountered when taking the size-independent perspective (see Fig. 4) still show clear differences. University of Oxford would have 22% fewer 4* publications based on metrics than based on peer review, while the difference varies between −14% and +9% based on low peer review uncertainty. Likewise, Queen Mary University would have 22% fewer 4* publications based on metrics than based on peer review. Based on low peer review uncertainty, the difference varies between −33% and +28%. The University of Portsmouth would have 51% more 4* publications based on metrics than based on peer review, while the difference varies between −55% and +36% based on low peer review uncertainty. The University of St. Andrews would have 13% more 4* publications based on metrics. This is within the range of −30% to +18% obtained based on low peer review uncertainty. Finally, Aberystwyth University would have 255% more 4* publications based on metrics, and it would have about ±100% 4* publications based on low peer review uncertainty. Although other institutions also show differences between metrics and peer review, these are not much larger or smaller than what could be expected based on peer review uncertainty.

Discussion

National RAEs evaluate the scientific performance of universities and other research institutions. To a large extent, this is often based on scientific publications. The role of citation metrics is regularly discussed in the literature, and the extent to which they correlate with peer review has been repeatedly analysed. Recently, in the context of the REF 2014 in the UK, the influential *Metric Tide* report (Wilsdon et al., 2015) concluded that metrics should only supplement, rather than supplant, peer review. The report's conclusion was substantiated by its finding that metrics correlate poorly with peer review. In contrast, earlier studies have shown that metrics may correlate quite well with peer review.

The discussion on metrics and peer review is characterised by a variety of correlations and an even larger variety of interpretations of these correlations. Correlations between metrics and peer review in the *Metric Tide* report are generally on the order of 0.4. Most previous studies have found correlations on the order of 0.7, but some have even reported correlations up to 0.9. Conclusions vary, even if the correlations are the same: some argue that a correlation of 0.7 is too low to consider replacing peer review by metrics, whereas others argue that a correlation of 0.7 is sufficiently high to do so.

We identify four points that need careful consideration in discussions on the agreement between metrics and peer review: (1) the level of aggregation; (2) whether a size-dependent perspective or a size-independent perspective is taken; (3) appropriate measures of agreement; and (4) uncertainty in peer review.

Most previous studies have analysed the agreement between metrics and peer review at the institutional level, whereas the recent *Metric Tide* report analysed the agreement at the level of individual publications. For the purpose of deciding between the use of metrics or peer review in the REF, the value of such a publication-level analysis is limited. The REF results are made available at the institutional level, which is therefore the most appropriate level of analysis. If correlations at the publication level are low, this does not necessarily mean that correlations at the institutional level will be low as well. Indeed, we find correlations at the institutional level that are substantially higher than the correlations at the publication level reported in the *Metric Tide* report. In line with previous results, we obtain size-independent correlations above 0.8 for a number of fields.

The REF has multiple objectives. It aims to provide a reputational yardstick, which is, for example, visible in the various league tables that are produced on the basis of the REF. It also aims to provide a basis for distributing funding. The objective of a reputational yardstick corresponds to a size-independent perspective, while the objective of funding allocation corresponds to a size-dependent perspective. Both perspectives are important in deciding whether metrics can replace peer review.

Some authors have found high size-dependent correlations, on the order of 0.9. We indeed find similar size-dependent correlations for many fields. It is important to realise that size-dependent correlations tend to reach high levels because metrics and peer review share a common factor, namely the size of an institution. This explains why size-dependent correlations may be as high as 0.9 while the corresponding size-independent correlations may be much lower. For example, we find a size-dependent correlation of 0.96 for Mathematical Sciences, whereas the size-independent correlation is only 0.39.

Measures of agreement should quantify agreement in a way that is most relevant in the specific context in which the measures are used. From this point of view, correlations are not necessarily the most appropriate measure of agreement. To compare metrics and peer review, we therefore use two other measures of agreement: the MAD for the size-independent perspective and the MAPD for the size-dependent perspective. In the REF, about 30% of the publications have been awarded 4*. From the size-independent perspective, we find that a number of fields in the REF show a MAD of about 3 percentage points between metrics and peer review. In these fields, when switching from peer review to metrics, the percentage of 4* publications of an institution will typically increase or decrease by about 3 percentage points. The MAPD between metrics and peer review from the size-dependent perspective is about 15% for these fields. This essentially means that the amount of funding allocated to an institution will typically increase or decrease by about 15%.

Differences between metrics and peer review can be interpreted in various ways. In this paper, we take peer review as the "gold standard" that should be matched as closely as possible by metrics. In the context of the REF this seems the most relevant perspective, because the REF currently relies on peer review and because the use of peer review in the REF seems to be widely

accepted. However, it is also possible that differences between metrics and peer review indicate that metrics better reflect the "true" scientific quality of publications than peer review. Without an independent third measure that can serve as the "gold standard", there is no way of establishing whether metrics or peer review offer a better reflection of scientific quality.

Regardless of the level of aggregation at which agreement between metrics and peer review is analysed and regardless of whether a size-dependent or a size-independent perspective is taken, agreement between metrics and peer review should be placed in an appropriate context. To determine whether agreement between metrics and peer review should be regarded as high or low, it is essential to make a comparison with internal peer review agreement. Unfortunately, there are currently no data available to quantify peer review uncertainty in the REF. Ideally, one needs to have an independent replication of the peer review process in the REF to determine the degree to which peer review is subject to uncertainty and to quantify internal peer review agreement. We recommend that uncertainty in peer review is analysed in the next round of the REF in 2021 to clarify this important point.

Given the lack of empirical data, we rely on a simple model to get an idea of the degree of uncertainty in peer review. For some fields, our model suggests that agreement between metrics and peer review is quite close to internal peer review agreement. In particular, this is the case for Physics, Clinical Medicine, and Public Health, Health Services & Primary Care. For these fields, the differences between metrics and peer review are relatively minor, from both a reputational (size-independent) and a funding (size-dependent) perspective. From the viewpoint of agreement between metrics and peer review, in these fields one may consider switching from peer review to metrics.

In some fields, metrics were used to inform the REF peer review. Even in fields in which metrics were not used in a formal way, reviewers may still have informally been influenced by metrics. It could be argued that this explains the high agreement between metrics and peer review. This may suggest that peer review should be organised differently. For example, peer reviewers should have sufficient time to properly evaluate each publication without the need to rely on metrics. Still, it may be difficult to limit the influence of metrics. Peer reviewers may have a strong tendency to echo what metrics tell them. The added benefit of peer review then seems questionable, especially considering the time and money it requires.

Importantly, we do not suggest that metrics *should* replace peer review in the REF. As shown in this paper, the argument that metrics should not be used because of their low agreement with peer review does not stand up to closer scrutiny for at least some fields. However, other arguments against the use of metrics may be provided, even for fields in which metrics and peer review agree strongly. Foremost, by relying on a metric, the goal of fostering "high quality" science may become displaced by the goal of obtaining a high metric. Metrics may invite gaming of citations and strategic behaviour that has unintended and undesirable consequences (de Rijcke et al., 2016). For example, evaluation on the basis of certain metrics may unjustly favour problematic research methods, which may lead to the "evolution of bad science" (Smaldino and McElreath, 2016). The use of a metric-driven approach in some fields, while maintaining a peer review approach in other fields, may also complicate the evaluation exercise and amplify disciplinary differences. Other arguments against replacing peer review by metrics are of a more pragmatic or more practical nature. One argument is that citation analysis may wield insufficient support and confidence in the scientific community (Wilsdon et al., 2015). Another argument is that there will always be some outputs that are not covered in bibliographic databases and for which it is not possible to obtain metrics. Of course, there are also other arguments in favour of metrics. For example, the total costs of the recent REF 2014 have been estimated at £246 million (Farla and Simmonds, 2015). By relying on metrics instead of peer review these costs could be reduced. First of all, the costs of panellists' time (£19 million) could be saved. However, the bulk of the costs (£212 million) were born by the institutions themselves in preparing the submissions to the REF. To reduce these costs, it has been suggested to simply consider all publications of institutions rather than only a selection (Harzing, 2017). All above arguments for and against metrics and peer review should be carefully weighed in the discussion on whether metrics should (partly) replace peer review in the REF.

Finally, as a limitation of our work, we emphasise that we do not consider the broader societal, cultural, and economic impact that is also evaluated in the REF. Such a broader evaluation cannot be done on the basis of metrics (Ravenscroft et al., 2017; Bornmann et al., 2018; Pollitt et al., 2016) and should therefore be carried out using peer review. Outputs that are not covered in bibliographic databases, such as the WoS, Scopus, Dimensions, and Microsoft Academic also need to be assessed by peer review.

Data availability

The datasets generated during and/or analysed during the current study are available in the Zenodo repository replication and can be accessed at https://github.com/vtraag/replication-uk-ref-2014 and https://doi.org/10.5281/zenodo.2564797.

Received: 17 August 2018 Accepted: 30 January 2019
Published online: 12 March 2019

Notes

1 Interestingly, the order of these objectives for REF 2014 are different, see https://www.ref.ac.uk/2014/about/.
2 http://www.hefce.ac.uk/funding/annallocns/1718/
3 In the REF 2014, in total 42,481 publications were awarded 4^* and 94,153 publications were awarded 3^*. In reality, calculations are more complex, as they involve the number of staff in FTE, subject cost weights, and specific weights for the London area.
4 Various studies have employed a multiple regression framework, and they have typically reported R^2 values. R^2 simply corresponds to the square of the (multiple) correlation. To provide unified results, we converted all R^2 values to their square root and report R values. To be clear, we also provide the originally reported R^2 values.
5 Data on the results and submissions are provided at https://www.rae.ac.ukwww.rae.ac.uk
6 The report also used precision and specificity, which are more appropriate than correlations for the individual publication level, but for comparability, we here focus on the reported correlations.
7 All data can be retrieved at http://www.ref.ac.uk/2014/.
8 Occasionally, incorrect DOIs were provided, resulting in seemingly duplicate publications for the same UoA and institution.
9 https://www.ref.ac.uk/about/
10 Note that PP (top 10%) concerns the proportion of publications that have been matched in the WoS, whereas PP (4^*) concerns the proportion of all submitted outputs.

References

Adams J, Gurney K, Jackson L (2008) Calibrating the zoom—a test of Zitt's hypothesis. Scientometrics 75:81–95. https://doi.org/10.1007/s11192-007-1832-7

Allen N, Heath O (2013) Reputations and research quality in British political science: the importance of journal and publisher rankings in the 2008 RAE. Br J Polit Int Relat 15:147–162. https://doi.org/10.1111/1467-856X.12006

Bence V, Oppenheim CT (2005) The evolution of the UK's Research Assessment Exercise: publications, performance and perceptions. J Educ Adm Hist 37:137–155. https://doi.org/10.1080/00220620500211189

Bergstrom CT (2007) Eigenfactor: measuring the value and prestige of scholarly journals. Coll Res Libr News 68:314–316. https://doi.org/10.5860/crln.68.5.7804

Bertocchi G, Gambardella A, Jappelli T, Nappi CA, Peracchi F (2015) Bibliometric evaluation vs. informed peer review: evidence from Italy. Res Policy 44:451–466. https://doi.org/10.1016/j.respol.2014.08.004

Bishop D (2014) BishopBlog: an alternative to REF2014? Blog. http://deevybee.blogspot.nl/2013/01/an-alternative-to-ref2014.html

Bornmann L, Haunschild R, Adams J (2018) Do altmetrics assess societal impact in the same way as case studies? An empirical analysis testing the convergent validity of altmetrics based on data from the UK Research Excellence Framework (REF). http://arxiv.org/abs/1807.03977arXiv:1807.03977

Butler L, McAllister I (2009) Metrics or peer review? Evaluating the 2001 UK research assessment exercise in political science. Polit Stud Rev 7:3–17. https://doi.org/10.1111/j.1478-9299.2008.00167.x

Carpenter MP, Gibb F, Harris M, Irvine J, Martin BR, Narin F (1988) Bibliometric profiles for British academic institutions: an experiment to develop research output indicators. Scientometrics 14:213–233. https://doi.org/10.1007/BF02020076

Clerides S, Pashardes P, Polycarpou A (2011) Peer review vs metric-based assessment: testing for bias in the RAE ratings of UK economics departments. Economica 78:565–583. https://doi.org/10.1111/j.1468-0335.2009.00837.x

Colman AM, Dhillon D, Coulthard B (1995) A bibliometric evaluation of the research performance of British university politics departments: publications in leading journals. Scientometrics 32:49–66. https://doi.org/10.1007/BF02020188

Crewe I (1988) Reputation, research and reality: the publication records of UK departments of politics, 1978–1984. Scientometrics 14:235–250. https://doi.org/10.1007/BF02020077

Derrick, G (2018) The Evaluators' Eye. Palgrave Macmillan, Cham, pp. 1–230. https://doi.org/10.1007/978-3-319-63627-6

Farla K, Simmonds P (2015) REF accountability review: costs, benefits and burden —report by Technopolis to the four UK higher education funding bodies, Technopolis

Harzing A-W (2017) Running the REF on a rainy Sunday afternoon: do metrics match peer review? https://harzing.com/publications/white-papers/running-the-ref-on-a-rainy-sunday-afternoon-do-metrics-match-peer-review Accessed 21 Nov 2018.

Harzing A-W, Alakangas S (2017) Microsoft Academic: is the phoenix getting wings? Scientometrics 110:371–383. https://doi.org/10.1007/s11192-016-2185-x

Hicks D (2012) Performance-based university research funding systems. Res Policy 41:251–261. https://doi.org/10.1016/j.respol.2011.09.007

Johnes J, Taylor J, Francis B (1993) The research performance of UK universities: a statistical analysis of the results of the 1989 Research Selectivity Exercise. J R Stat Soc A156:271–286. https://doi.org/10.2307/2982732

Jump P (2014) Evolution of the REF. Times Higher Education. https://www.timeshighereducation.com/features/evolution-of-the-ref/2008100.article. Accessed 21 Nov 2018.

Jump P (2015) Can the Research Excellence Framework run on metrics? Times Higher Education. https://www.timeshighereducation.com/can-the-research-excellence-framework-ref-run-on-metrics. Accessed 21 Nov 2018.

Kelly A, Burrows R (2011) Measuring the value of sociology? Some notes on performative metricization in the contemporary academy. Sociol Rev 59:130–150. https://doi.org/10.1111/j.1467-954X.2012.02053.x

Lim Ban Seng, Willett P (1995) The citedness of publications by United Kingdom library schools. J Inf Sci 21:68–71. https://doi.org/10.1177/016555159502100109

Mahdi S, D'Este P, Neely A (2008) Are they good predictors of RAE scores? Technical Report February. Advanced Institute of Management Research. https://doi.org/10.2139/ssrn.1154053

Mckay S (2012) Social policy excellence—peer review or metrics? Analyzing the 2008 Research Assessment Exercise in social work and social policy and administration. Soc Policy Adm 46:526–543. https://doi.org/10.1111/j.1467-9515.2011.00824.x

Mryglod O, Kenna R, Holovatch Y, Berche B (2013a) Absolute and specific measures of research group excellence. Scientometrics 95:115–127. https://doi.org/10.1007/s11192-012-0874-7

Mryglod O, Kenna R, Holovatch Y, Berche B (2013b) Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. Scientometrics 97:767–777. https://doi.org/10.1007/s11192-013-1058-9

Mryglod O, Kenna R, Holovatch Y, Berche B (2015a) Predicting results of the Research Excellence Framework using departmental *h*-index. Scientometrics 102:2165–2180. https://doi.org/10.1007/s11192-014-1512-3

Mryglod O, Kenna R, Holovatch Y, Berche B (2015b) Predicting results of the Research Excellence Framework using departmental *h*-index: revisited. Scientometrics 104:1013–1017. https://doi.org/10.1007/s11192-015-1567-9

Norris M, Oppenheim C (2003) Citation counts and the Research Assessment Exercise v. J Doc 59:709–730. https://doi.org/10.1108/00220410310698734

Norris M, Oppenheim C (2010) Peer review and the *h*-index: two studies. J Informetr 4:221–232. https://doi.org/10.1016/j.joi.2009.11.001

Oppenheim C (1995) The correlation between citation counts and the 1992 Research Assessment Exercise ratings for British library and information science university departments. J Doc 51:18–27. https://doi.org/10.1108/eb026940

Oppenheim C (1997) The correlation between citation counts and the 1992 Research Assessment Exercise ratings for British research in genetics, anatomy and archaeology. J Doc 53:477–487. https://doi.org/10.1108/EUM0000000007207

Pearson K (1896) Mathematical contributions to the theory of evolution–-on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond 60:489–498. https://doi.org/10.1098/rspl.1896.0076

Pollitt A, Potoglou D, Patil S, Burge P, Guthrie S, King S, Wooding S, Wooding S, Grant J (2016) Understanding the relative valuation of research impact: a best–worst scaling experiment of the general public and biomedical and health researchers. BMJ Open 6:e010916. https://doi.org/10.1136/bmjopen-2015-010916

Pride D, Knoth P (2018) Peer review and citation data in predicting university rankings, a large-scale analysis. http://arxiv.org/abs/1805.08529arXiv:1805.08529

Ravenscroft J, Liakata M, Clare A, Duma D, Thirion B, Grisel O (2017) Measuring scientific impact beyond academia: an assessment of existing impact metrics and proposed improvements. PLoS ONE 12:e0173152. https://doi.org/10.1371/journal.pone.0173152

de Rijcke S, Wouters PF, Rushforth AD, Franssen TP, Hammarfelt B (2016) Evaluation practices and effects of indicator use—a literature review. Res Eval 25:161–169. https://doi.org/10.1093/reseval/rvv038

Ruiz-Castillo J, Waltman L (2015) Field-normalized citation impact indicators using algorithmically constructed classification systems of science. J Informetr 9:102–117. https://doi.org/10.1016/j.joi.2014.11.010

Smaldino PE, McElreath R (2016) The natural selection of bad science. R Soc Open Sci 3:160384. https://doi.org/10.1098/rsos.160384

Smith DAT, Eysenck PM, Smith A, Eysenck M (2002) The correlation between RAE ratings and citation counts in psychology. Technical Report. University of London. http://cogprints.org/2749/index.html

Taylor J (1994) Measuring research performance in business and management studies in the United Kingdom: the 1992 Research Assessment Exercise. Br J Manag 5:275–288. https://doi.org/10.1111/j.1467-8551.1994.tb00079.x

Taylor J (2011) The assessment of research quality in UK universities: peer review or metrics? Br J Manag 22:202–217. https://doi.org/10.1111/j.1467-8551.2010.00722.x

Thomas PR, Watkins DS (1998) Institutional research rankings via bibliometric analysis and direct peer review: a comparative case study with policy implications. Scientometrics 41:335–355. https://doi.org/10.1007/BF02459050

Traag VA, Waltman L (2018) Systematic analysis of agreement between metrics and peer review in the UK REF, Zenodo, replication material. https://doi.org/10.5281/zenodo.2564797

Waltman L, van Eck NJ (2012) A new methodology for constructing a publication-level classification system of science. J Am Soc Inf Sci Technol 63:2378–2392. https://doi.org/10.1002/asi.22748

West J, Bergstrom T (2010) Big Macs and Eigenfactor scores: don't let correlation coefficients fool you. J Am Soc Inf Sci Technol 61:1–25. https://doi.org/10.1002/ASI.V61:9

Wilsdon J, Allen L, Belfiore E, Campbell P, Curry S, Hill S, Jones R, Kain R, Kerridge S, Thelwall M, Tinkler J, Viney I, Wouters P, Hill J, Johnson B (2015) Metric Tide: report of the independent review of the role of metrics in research assessment and management. Technical Report. Higher Education Funding Council for England. https://doi.org/10.13140/RG.2.1.4929.1363

Wooding S, Van Leeuwen TN, Parks S, Kapur S, Grant J (2015) UK doubles its "World-Leading" research in life sciences and medicine in six years: testing the claim? PLoS ONE 10:e0132990. https://doi.org/10.1371/journal.pone.0132990

Zacharewicz T, Lepori B, Reale E, Jonkers K (2018) Performance-based research funding in EU member states—a comparative assessment. Sci Public Policy scy041. https://doi.org/10.1093/scipol/scy041

Acknowledgements

We thank Lutz Bornmann, Anne-Wil Harzing, Steven Hill, Sven Hug, and David Pride for their comments on an earlier version of this paper. We like to thank Jeroen Baas for discussion on the analysis by Elsevier.

Additional information

The online version of this article (https://doi.org/10.1057/s41599-019-0233-x) contains supplementary material, which is available to authorised users.

Competing interests: The authors act as bibliometric consultants to CWTS B.V., which provides commercial bibliometric services.

Reprints and permission information is available online at http://www.nature.com/reprints

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.