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

V.A. Traag[a)] and L. Waltman
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

(Dated: 13 August 2018)

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 quite well with peer review and may offer an alternative to peer review.

I. INTRODUCTION

Many countries have some form of a national research assessment exercise 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 research assessment exercise 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, but we do not consider these arguments further.

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, D'Este, and Neely (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.

[a)]Electronic mail: v.a.traag@cwts.leidenuniv.nl

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.

II. 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/)

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. Finally, 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 Section IV B.

To provide an indication of the importance of the REF 2014, we briefly look at the funding of UK higher education in 2017–2018[1]. 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 £2 000 per 3* publication per year on average[2]. The total amount of about £685M allocated through the evaluation of research output represented about 20% of the total budget of HEFCE of £3 602M and about 40% of the total research budget of HEFCE of £1 606M. As such, it is a sizeable proportion of the total budget.

III. LITERATURE REVIEW

We review previous literature on how metrics compare to peer review in previous research assessment exercises in

―――――――

[1] http://www.hefce.ac.uk/funding/annallocns/1718/
[2] 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.

the UK. We then briefly review literature that analyses how metrics and peer review compared in the REF 2014.

A.   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, Taylor, and Francis (1993)).

The 1992 exercise—then called the Research Assessment Exercise (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[3]

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, Dhillon, and Coulthard (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, Pashardes, and Polycarpou (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, D'Este, and Neely (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, Gurney, and Jackson (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,

---

[3] 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.

$R \approx 0.7$–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[4]. 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 Section IV B.

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".

B. 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 citations correlate with peer review. The report found correlations[5] 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 Section IV A.

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

---

[4] Data on the results and submissions are provided at www.rae.ac.uk

[5] 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.

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 Section IV B.

IV. DATA AND METHODS

The REF 2014 provides a well-documented dataset of both the evaluation results and the submitted publications that have been evaluated[6]. 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 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 distinguishes 36 UoAs. Every publication is submitted on behalf of a particular institution. Some publications can be submitted multiple times if they are co-authored by several authors. We consider only the unique publications within an institution. In other words, we count a publication only once, even if it is submitted twice by the same institution. We use full counting, similar to the assessment in the REF. Publications that are co-authored and submitted by multiple institutions may thus be counted multiple times. Similarly, some publications are submitted in multiple UoAs, and we take them into account in each UoA.

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 use a citation window 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 4 000 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.

Before discussing our results, we first discuss four important methodological considerations. First, we discuss the level of aggregation at which agreement between metrics and peer review should be analysed. We then discuss 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.

A. 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

---

[6] All data can be retrieved at www.ref.ac.uk/2014.

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 evaluate individual publications but rather to evaluate institutions. 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.

B. 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 (1000 samples). In this example, the
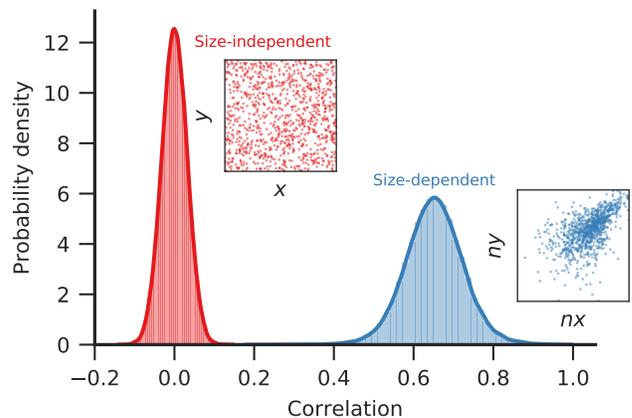

FIG. 1. The correlation between two size-dependent metrics can be quite high even if the corresponding size-independent metrics are completely uncorrelated. The insets show the scatter plots of size-dependent and size-independent metrics. For the size-independent scatter plot logarithmic scales are used.

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. 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 publica-

tions in the WoS. Both the P(4*) and the P(top 10%) are clearly size-dependent.

C. 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 Appendix 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 Appendix A for details).

D. 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. Evidence from the Italian research assessment exercise, known as VQR, suggests that internal peer review agreement is quite low (Bertocchi *et al.*, 2015). Unfortunately, detailed peer review results of the REF at the publication level are not available because of privacy reasons. 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. 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 Appendix 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 Appendix B) and one with high uncertainty ($\sigma_\epsilon^2 = 1$). The results presented in the next section are based on 1 000 bootstrap samples. We report both the median outcome obtained from 1 000 samples and the interval that covers 95% of the outcomes.

V. RESULTS

We now describe the results from our analysis.[7] 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.

A. 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 Table C.1). A number of other fields show
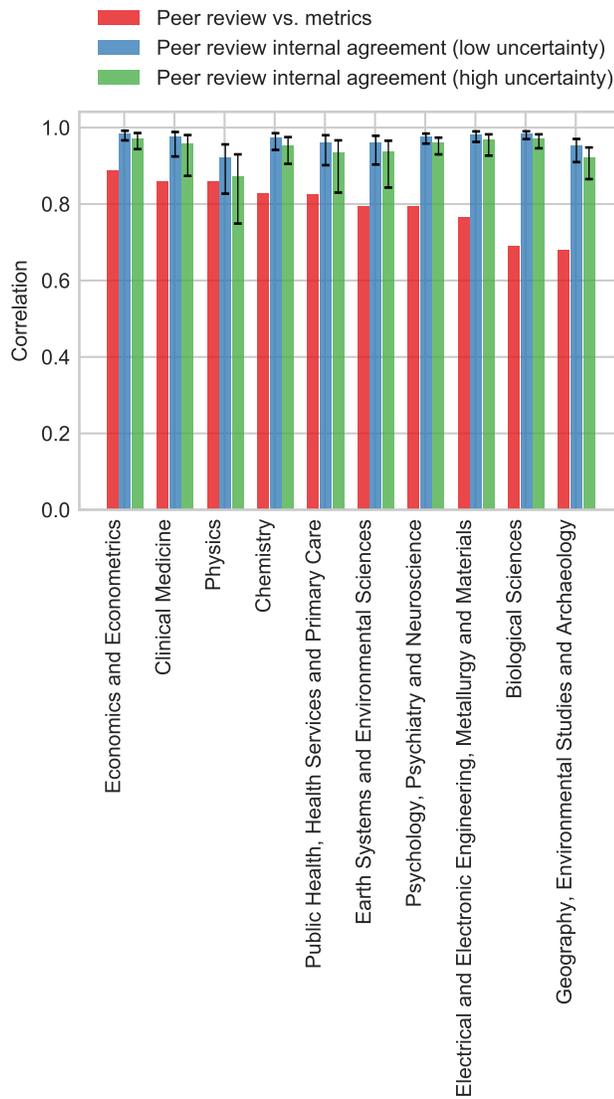

FIG. 2. Size-independent correlation between PP(top 10%) and PP(4*) compared with correlations based on a model of peer review uncertainty. Results are shown only for the 10 units of assessment with the highest correlation between metrics and peer review.
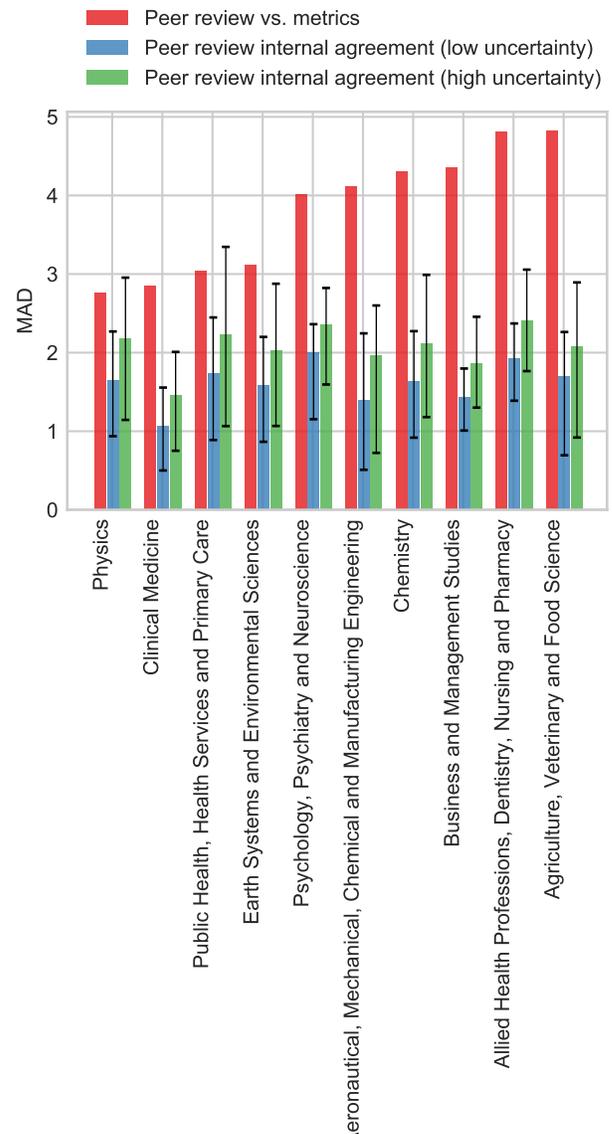

FIG. 3. Size-independent median absolute difference (MAD) between PP(top 10%) and PP(4*) compared with the MAD based on a model of peer review uncertainty. Results are shown only for the 10 units of assessment with the lowest MAD between metrics and peer review.

correlations on the order of 0.7, which is in line with previous 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).

Remarkably, our results strongly differ from the analysis by Elsevier of the REF results (Jump, 2015), even though the Elsevier analysis also found some relatively strong correlations. In particular, the Elsevier analysis found correlations for Physics and Clinical Medicine on the order of 0.3. Public Health did a little better in the Elsevier analysis, but still the correlation was only about 0.5. Finally, Biology had the single highest correlation of about 0.75 in the Elsevier analysis, whereas this cor-

relation is much lower in our results. Unfortunately, the results of the Elsevier analysis were described only briefly in a blog post, and we have not been able to obtain any further details. It may be of interest to compare the different results and to better understand why Elsevier's results differ from ours. There can be various reasons for this, such as the focus on all publications of an institution versus only the publications submitted to the REF, the use of different databases (Scopus vs. WoS), and the use of different field definitions in the field-normalised citation metrics.

The results of the peer review uncertainty may be surprising (see Fig. 2). Although the bootstrapped corre-
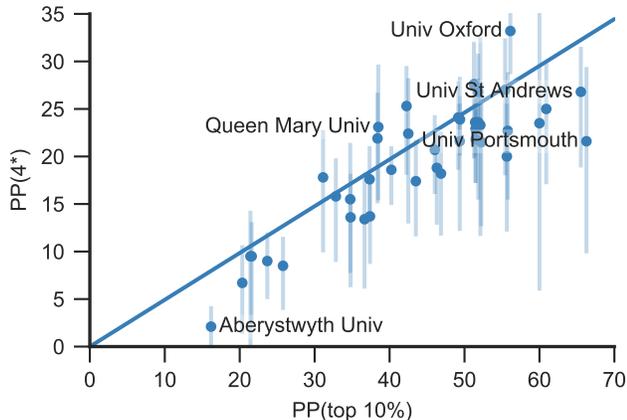
FIG. 4. Scatter plot of PP(top 10%) and PP(4*) at the institutional level for Physics. Error bars indicate the 95% interval of bootstrapped peer review results for low peer review uncertainty. The solid line indicates the proportion of 4* publications considered to be equivalent to a given proportion of top 10% publications (see Appendix A for details).

lations 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 Table C.1). In the interpretation of the MAD, it is important to keep in mind that overall about 30% of the 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 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 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.

B. Size-dependent perspective

As expected, the size-dependent correlations are much higher than the size-independent correlations (see 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 Section IV B. 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 Fig. C.3). The 10 UoAs with the lowest MAPDs all show size-dependent
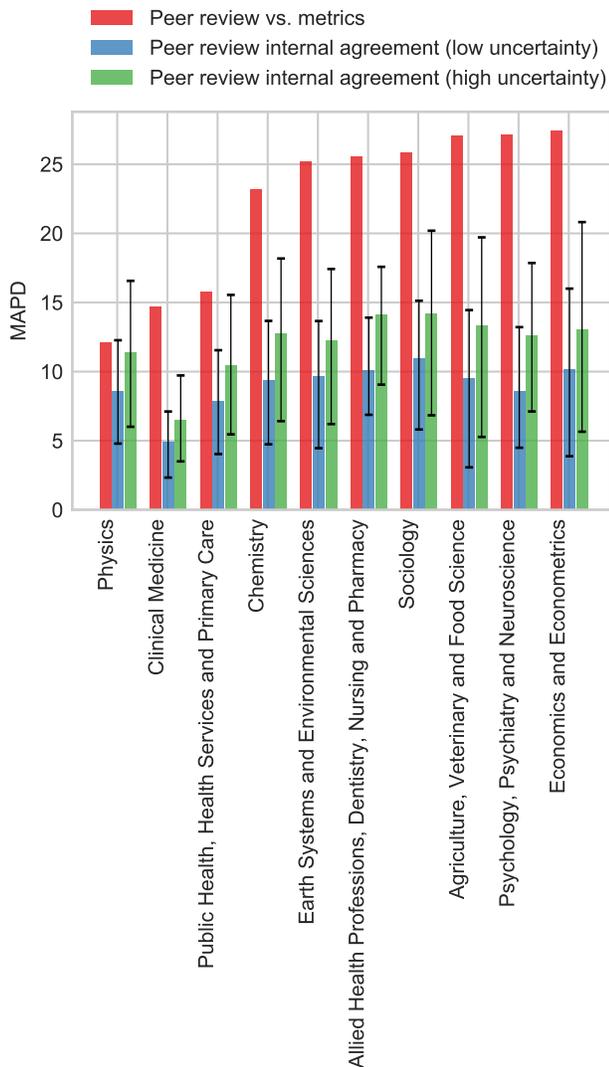
FIG. 5. Size-dependent median absolute percentage difference (MAPD) of P(top 10%) relative to P(4*) compared with the MAPD based on a model of peer review uncertainty. Results are shown only for the 10 units of assessment with the lowest MAPD of metrics relative to peer review.

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 Appendix A for details).

versity 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.

VI. DISCUSSION

National research assessment exercises 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 Research Excellence Framework (REF) 2014 in the UK, the influential *Metric Tide* report (Wilsdon *et al.*, 2015) concluded that metrics should only supplement, rather than supplant, peer

correlations close to or above 0.9, which illustrates how correlations and MAPDs may potentially lead to different conclusions. Biology is a 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 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. Uni-
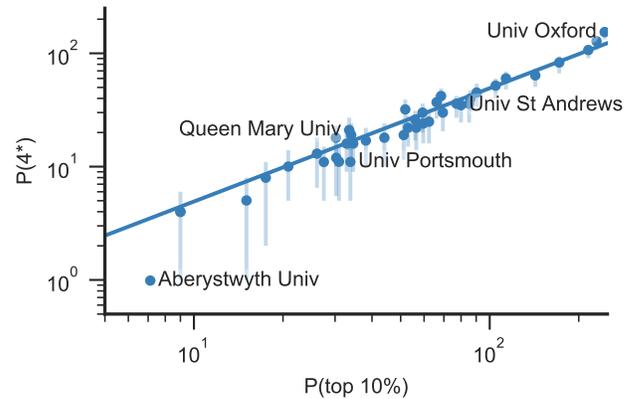
review. The report's conclusion was substantiated by its finding that metrics correlate poorly with peer review. In contrast, earlier studies have shown that metrics may correlate quite well with peer review.

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 median absolute difference for the size-independent perspective and the median absolute percentage difference 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 median absolute difference 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 median absolute percentage difference 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%.

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). 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 the use of a metric-driven approach in some fields, while maintaining a peer review approach in other fields, may complicate the overall evaluation exercise. A third 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. The above arguments should be carefully weighed in the discussion on whether metrics should or should not 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, Haunschild, and Adams, 2018) 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.

ACKNOWLEDGMENTS

Adams, J., Gurney, K., and Jackson, L., "Calibrating the zoom — a test of Zitt's hypothesis," Scientometrics **75**, 81–95 (2008).

Allen, N. and Heath, O., "Reputations and research quality in british political science: The importance of journal and publisher rankings in the 2008 RAE," Br. J. Polit. Int. Relations **15**, 147–162 (2013).

Bence, V. and Oppenheim, C. T., "The evolution of the UK's Research Assessment Exercise: Publications, performance and perceptions," J. Educ. Adm. Hist. **37**, 137–155 (2005).

Bergstrom, C. T., "Eigenfactor: Measuring the value and prestige of scholarly journals," Coll. Res. Libr. News **68**, 314–316 (2007).

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., and Peracchi, F., "Bibliometric evaluation vs. informed peer review: evidence from Italy," Res. Policy **44**, 451–466 (2015).

Bishop, D., "BishopBlog: an alternative to REF2014?" Blog (2014).

Bornmann, L., Haunschild, R., and Adams, J., "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)," (2018), arXiv:1807.03977.

Butler, L. and McAllister, I., "Metrics or peer review? Evaluating the 2001 UK research assessment exercise in political science," Polit. Stud. Rev. **7**, 3–17 (2009).

Carpenter, M. P., Gibb, F., Harris, M., Irvine, J., Martin, B. R., and Narin, F., "Bibliometric profiles for British academic institutions: An experiment to develop research output indicators," Scientometrics **14**, 213–233 (1988).

Clerides, S., Pashardes, P., and Polycarpou, A., "Peer review vs metric-based assessment: testing for bias in the RAE ratings of UK economics departments," Economica **78**, 565–583 (2011).

Colman, A. M., Dhillon, D., and Coulthard, B., "A bibliometric evaluation of the research performance of British university politics departments: publications in leading journals," Scientometrics **32**, 49–66 (1995).

Crewe, I., "Reputation, research and reality: the publication records of UK departments of politics, 1978–1984," Scientometrics **14**, 235–250 (1988).

Harzing, A.-W., "Running the REF on a rainy Sunday afternoon: do metrics match peer review?" (2017), white paper.

Harzing, A.-W. and Alakangas, S., "Microsoft Academic: is the phoenix getting wings?" Scientometrics **110**, 371–383 (2017).

Hicks, D., "Performance-based university research funding systems," Res. Policy **41**, 251–261 (2012).

Johnes, J., Taylor, J., and Francis, B., "The research performance of UK universities: a statistical analysis of the results of the 1989 Research Selectivity Exercise," J. R. Statist. Soc. A **156**, 271–286 (1993).

Jump, P., "Evolution of the REF," Times Higher Education (2014).

Jump, P., "Can the research excellence framework run on metrics?" Times Higher Education (2015).

Kelly, A. and Burrows, R., "Measuring the value of sociology? Some notes on performative metricization in the contemporary academy," Sociol. Rev. **59**, 130–150 (2011).

Lim Ban Seng, and Willett, P., "The citedness of publications by United Kingdom library schools," J. Inf. Sci. **21**, 68–71 (1995).

Mahdi, S., D'Este, P., and Neely, A., "Are they good predictors of RAE scores?" Tech. Rep. February (Advanced Institute of Management Research, 2008).

Mckay, S., "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 (2012).

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B., "Absolute and specific measures of research group excellence," Scientometrics **95**, 115–127 (2013a).

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B., "Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence," Scientometrics **97**, 767–777 (2013b).

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B., "Predicting results of the Research Excellence Framework using departmental h-index," Scientometrics **102**, 2165–2180 (2015a).

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B., "Predicting results of the Research Excellence Framework using de-

partmental h-index: revisited," Scientometrics **104**, 1013–1017 (2015b).

Norris, M. and Oppenheim, C., "Citation counts and the Research Assessment Exercise v," J. Doc. **59**, 709–730 (2003).

Norris, M. and Oppenheim, C., "Peer review and the h-index: two studies," J. Informetr. **4**, 221–232 (2010).

Oppenheim, C., "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 (1995).

Oppenheim, C., "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 (1997).

Pearson, K., "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. London **60**, 489–498 (1896).

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

Ravenscroft, J., Liakata, M., Clare, A., Duma, D., Thirion, B., and Grisel, O., "Measuring scientific impact beyond academia: an assessment of existing impact metrics and proposed improvements," PLOS ONE **12**, e0173152 (2017).

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., and Hammarfelt, B., "Evaluation practices and effects of indicator use—a literature review," Res. Eval. **25**, 161–169 (2016).

Ruiz-Castillo, J. and Waltman, L., "Field-normalized citation impact indicators using algorithmically constructed classification systems of science," J. Informetr. **9**, 102–117 (2015).

Smaldino, P. E. and McElreath, R., "The natural selection of bad science," R. Soc. Open Sci. **3**, 160384 (2016).

Smith, D. A. T., Eysenck, P. M., Smith, A., and Eysenck, M., "The correlation between RAE ratings and citation counts in psychology," Tech. Rep. (University of London, 2002).

Taylor, J., "Measuring research performance in business and management studies in the United Kingdom: the 1992 Research Assessment Exercise," Br. J. Manag. **5**, 275–288 (1994).

Taylor, J., "The assessment of research quality in UK universities: peer review or metrics?" Br. J. Manag. **22**, 202–217 (2011).

Thomas, P. R. and Watkins, D. S., "Institutional research rankings via bibliometric analysis and direct peer review: a comparative case study with policy implications," Scientometrics **41**, 335–355 (1998).

Waltman, L. and van Eck, N. J., "A new methodology for constructing a publication-level classification system of science," J. Am. Soc. Inf. Sci. Technol. **63**, 2378–2392 (2012).

West, J. and Bergstrom, T., "Big Macs and Eigenfactor scores: don't let correlation coefficients fool you," J. Am. Soc. Inf. Sci. Technol. **61**, 1–25 (2010).

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., and Johnson, B., "Metric Tide: report of the independent review of the role of metrics in research assessment and management," Tech. Rep. (Higher Education Funding Council for England, 2015).

Zacharewicz, T., Lepori, B., Reale, E., and Jonkers, K., "Performance-based research funding in EU member states—a comparative assessment," Sci. Public Policy , scy041 (2018).

Appendix A: Measures of agreement

The Pearson correlation coefficient quantifies the extent to which a linear relationship of the form $\hat{y}_i = a + b x_i$ provides a good fit to the data. In this framework, both the intercept a and the slope b are estimated based on the least squares principle. According to this principle, we find a and b such that $\sum_i (\hat{y}_i - y_i)^2$ is minimal. The explained variance R^2 can then be expressed as

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}, \tag{A1}$$

where $\bar{y} = \frac{1}{n} \sum_i y_i$ is the average of $y_1, \dots, y_n$, with n the number of observations. The Pearson correlation coefficient is either the positive or the negative square root of R^2. In our context, correlations are usually positive (i.e. $b > 0$), which means that the Pearson correlation coefficient is the positive square root of R^2. When the Pearson correlation coefficient is high, the average of the squared differences, i.e. $\frac{1}{n} \sum_i (\hat{y}_i - y_i)^2$, is small relative to the variance, i.e. $\frac{1}{n} \sum_i (\bar{y} - y_i)^2$. However, if the variance is very large, a high correlation coefficient may be obtained even though the squared differences are still substantial. This illustrates the underlying problem of size-dependent correlations, as discussed in Section IV B.

Another problem of the Pearson correlation coefficient is that it allows for a non-zero intercept a. In our context, having no top 10% publications should correspond to having no 4* publications. This means that the intercept a should always be zero. We then work with the simple linear relationship $\hat{y}_i = b x_i$, and we need to estimate only the slope b. The Pearson correlation has the drawback that it allows for a non-zero intercept a, for which there is no proper conceptual justification in our context.

As stated in the main text, we use two measures of agreement in addition to correlations: the median absolute difference (MAD) and the median absolute percentage difference (MAPD). The MAD is defined as

$$\text{MAD} = \text{median}_i \, |\hat{y}_i - y_i|, \tag{A2}$$

and the MAPD is defined as

$$\text{MAPD} = \text{median}_i \, \frac{|\hat{y}_i - y_i|}{y_i}, \tag{A3}$$

where $\hat{y}_i = bx_i$. If $\hat{y}_i = y_i = 0$, we define the MAPD to be 0. The MAPD is independent of a multiplicative factor: if we multiply each y_i by a certain n_i, this does not affect the MAPD. In other words,

$$\text{median}_i \frac{|n_i\hat{y}_i - n_i y_i|}{n_i y_i} = \text{median}_i \frac{|\hat{y}_i - y_i|}{y_i}, \tag{A4}$$

provided that the estimate $\hat{y}_i$ remains unchanged (which it indeed does, given our estimation of b, as we discuss next).

We rely on a simple estimation of b that has a straightforward interpretation. We determine how many 4* publications are worth a single top 10% publication. To do so, we calculate for each UoA the ratio of the total number of 4* publications and the total number of top 10% publications. This ratio then provides our estimate of b, and it provides a straightforward way to transform a certain number of top 10% publications into a corresponding number of 4* publications.

The estimates of b for each UoA are reported in Table A.1. The number of 4* publications per top 10% publication varies quite substantially over fields. In some fields, such as Clinical Medicine and Physics, each top 10% publication is worth about 0.5 4* publications. In other fields, such as Economics & Econometrics, each top 10% publication is worth about 1.5 4* publications. There are also fields, such as Law and Classics, in which the number of 4* publications per top 10% publication is very high, even above 10. To some extent, this is caused by the fact that the WoS coverage in these fields is low. In addition, the criteria for awarding 4* may not be the same across all UoAs, at least not compared with metrics.

TABLE A.1: Publication numbers per unit of assessment.

	Unit of assessement	Nb. submissions	Nb. 4* pub.	Nb. top 10% pub.	Ratio $\frac{4*}{\text{top }10\%}$	WoS coverage
1	Clinical Medicine	13 400	3 107	5 417	0.57	94.8
2	Public Health, Health Services and Primary Care	4 881	1 093	2 008	0.54	90.0
3	Allied Health Professions, Dentistry, Nursing and Pharmacy	10 358	2 185	2 008	1.09	91.3
4	Psychology, Psychiatry and Neuroscience	9 126	2 361	2 866	0.82	93.2
5	Biological Sciences	8 608	2 511	3 126	0.80	97.3
6	Agriculture, Veterinary and Food Science	3 919	708	1 003	0.71	95.5
7	Earth Systems and Environmental Sciences	5 249	951	2 011	0.47	94.8
8	Chemistry	4 698	1 026	1 646	0.62	98.9
9	Physics	6 446	1 363	2 760	0.49	95.0
10	Mathematical Sciences	6 994	1 562	1 326	1.18	84.7
11	Computer Science and Informatics	7 651	1 693	880	1.92	61.1
12	Aeronautical, Mechanical, Chemical and Manufacturing Engineering	4 143	752	706	1.07	93.3
13	Electrical and Electronic Engineering, Metallurgy and Materials	4 025	800	926	0.86	92.2
14	Civil and Construction Engineering	1 384	246	256	0.96	89.5
15	General Engineering	8 679	1 486	1 656	0.90	91.6
16	Architecture, Built Environment and Planning	3 781	840	321	2.61	50.0
17	Geography, Environmental Studies and Archaeology	6 017	1 326	1 453	0.91	72.9
18	Economics and Econometrics	2 600	715	470	1.52	78.3
19	Business and Management Studies	12 202	2 500	1 809	1.38	77.7
20	Law	5 522	1 104	110	10.05	18.0
21	Politics and International Studies	4 365	910	487	1.87	50.4
22	Social Work and Social Policy	4 784	917	371	2.47	54.8
23	Sociology	2 630	514	300	1.71	56.5
24	Anthropology and Development Studies	2 013	385	243	1.59	49.0
25	Education	5 519	1 205	430	2.80	48.2
26	Sport and Exercise Sciences, Leisure and Tourism	2 757	541	447	1.21	76.7
27	Area Studies	1 724	408	84	4.86	32.3
28	Modern Languages and Linguistics	4 932	1 204	155	7.76	23.6

Unit of assessement	Nb. submissions	Nb. 4^* pub.	Nb. top 10% pub.	Ratio $\frac{4^*}{\text{top }10\%}$	WoS coverage
29 English Language and Literature	6 923	1 950	157	12.44	15.3
30 History	6 431	1 765	379	4.66	26.1
31 Classics	1 386	410	35	11.59	8.8
32 Philosophy	2 173	569	210	2.71	41.5
33 Theology and Religious Studies	1 558	356	24	14.68	14.1
34 Art and Design: History, Practice and Theory	6 321	1 130	89	12.74	12.5
35 Music, Drama, Dance and Performing Arts	4 246	1 071	100	10.71	14.6
36 Communication, Cultural and Media Studies, Library and Information Management	3 517	818	158	5.17	27.0

Appendix B: Model of REF peer review

To analyse uncertainty in peer review in the REF, we use a simple mathematical model. Below, we discuss the details of this model and the way in which the model is used to analyse uncertainty in REF peer review.

1. Model

Each institution k submits a certain number of publications n_{ku} in a certain UoA u. We assume that each publication i has an intrinsic "value" v_i. This value cannot be observed directly, but it does influence the peer review judgement that determines whether a publication is awarded four stars or not. We assume that the value v_i of publication i is log-normally distributed as

$$v_i \sim \text{LogNormal}(\mu_{ku}, 1). \tag{B1}$$

The parameter μ_{ku} represents the capability of institution k in UoA u to produce publications of a high value. The higher μ_{ku}, the higher on average the values of publications of institution k in UoA u.

Reviewers cannot directly observe the value v_i of publication i. Instead, they need to estimate this value. When reviewers estimate the value v_i of publication i, they may make a certain "error". In other words, the value estimated by reviewers may differ from the true value. We assume that the "perceived value" p_i of publication i is given by

$$p_i = v_i \epsilon_i, \tag{B2}$$

where ϵ_i represents the error made in estimating the value of the publication. We assume ϵ_i to be log-normally distributed as

$$\epsilon_i \sim \text{LogNormal}\left(-\frac{\sigma_\epsilon^2}{2}, \sigma_\epsilon^2\right). \tag{B3}$$

On average, ϵ_i equals 1. If $\epsilon_i > 1$, reviewers overestimate the value v_i. On the other hand, if $\epsilon_i < 1$, they underestimate the value v_i. However, on average, the perceived value p_i equals the true value v_i.

The inaccuracy of peer review is determined by the parameter σ_ϵ^2. If $\sigma_\epsilon^2 = 0$, peer review is perfectly accurate, and the perceived value p_i always equals the true value v_i. Higher values of σ_ϵ^2 correspond to less accurate peer review, so that the perceived value p_i may differ from the true value v_i.

Finally, we assume that publications with a perceived value higher than some threshold p^{4^*} are awarded four stars. Hence, publication i is awarded four stars if $p_i > p^{4^*}$. For any value v_i, there is some probability that publication i is awarded four stars. We denote this probability by $\Pr(i \text{ has } 4^* \mid v_i)$. For different values of σ_ϵ^2, Fig. B.1 shows how $\Pr(i \text{ has } 4^* \mid v_i)$ depends on the value v_i.
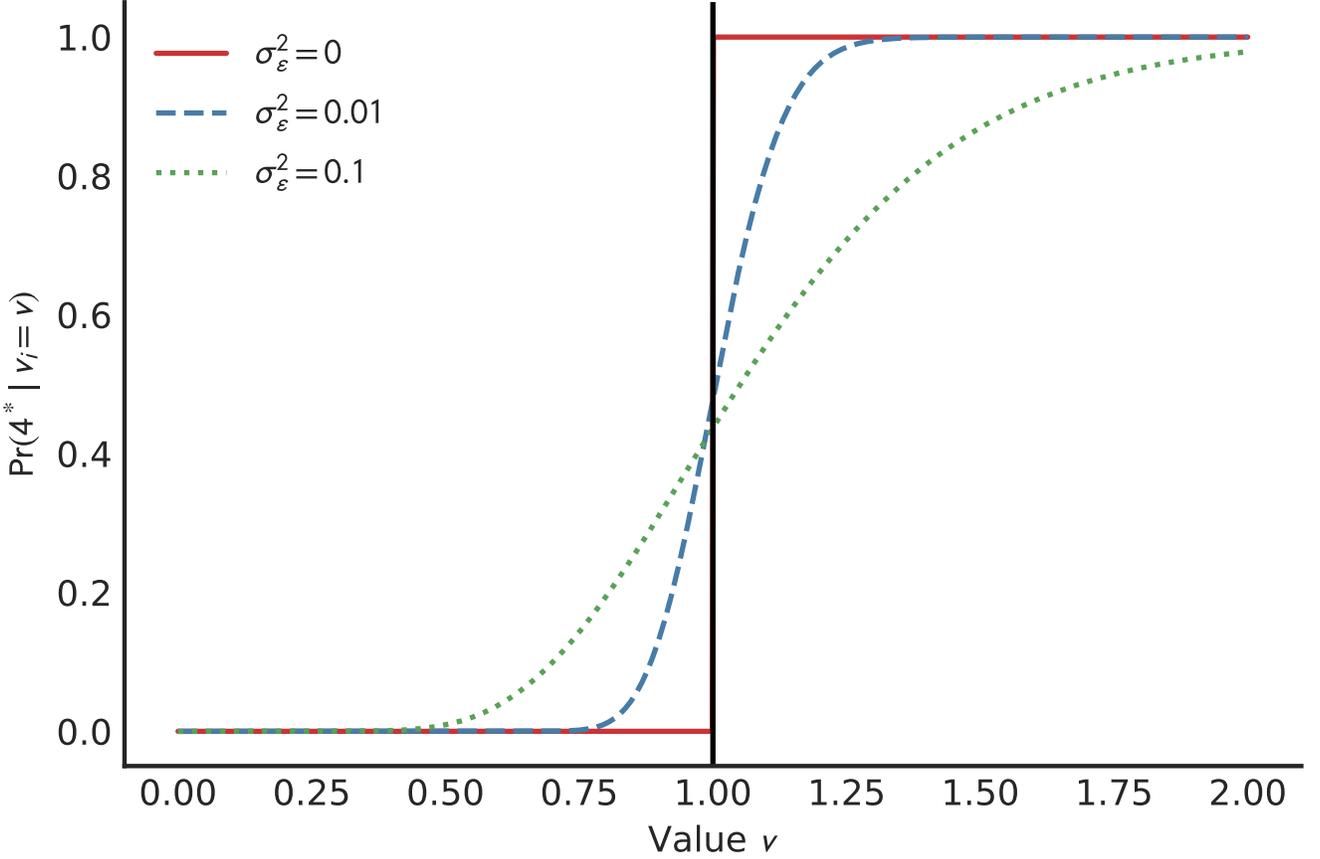
FIG. B.1. Conditional probability that publication i is awarded four stars given that it has value v_i. Three scenarios are considered for the inaccuracy of peer review σ_ϵ^2. The threshold for being awarded four stars equals $p^{4^*} = 1$.

2. Parameters

For each institution k and UoA u, we estimate μ_{ku} based on the observed proportion of 4^* publications of institution k in UoA u, denoted by $\mathrm{PP}_{ku}(4^*)$. It follows from Eqs. B1–B3 that the perceived value p_i is distributed as

$$p_i \sim \mathrm{LogNormal}\left(\mu_{ku} - \frac{\sigma_\epsilon^2}{2}, 1 + \sigma_\epsilon^2\right). \tag{B4}$$

Denote by F_{ku} the corresponding cumulative distribution. Hence, $F_{ku}(p)$ equals the probability that $p_i \leq p$. It follows that $1 - F_{ku}(p^{4^*})$ equals the probability that a publication is awarded four stars. For a given σ_ϵ^2 and p^{4^*}, the maximum likelihood estimate of μ_{ku} is obtained by choosing μ_{ku} such that $1 - F_{ku}(p^{4^*}) = \mathrm{PP}_{ku}(4^*)$.

Results obtained using our model seem to be largely independent of the threshold p^{4^*}. We therefore simply set this threshold to $p^{4^*} = 1$. We use values of 0.1 and 1.0 for σ_ϵ^2, corresponding respectively to a relatively high and a relatively low accuracy of peer review.

3. Estimating peer review uncertainty by resampling

In the REF peer review results, we observe only whether a publication has been awarded four stars or not. We do not observe the perceived value p_i of a publication i, and clearly we do not know a publication's true value v_i. To estimate the uncertainty of peer review in a certain field, we sample possible values for the publications of the institutions active in this field. Using Bayes' theorem, the probability that publication i has value v_i given that it has four stars
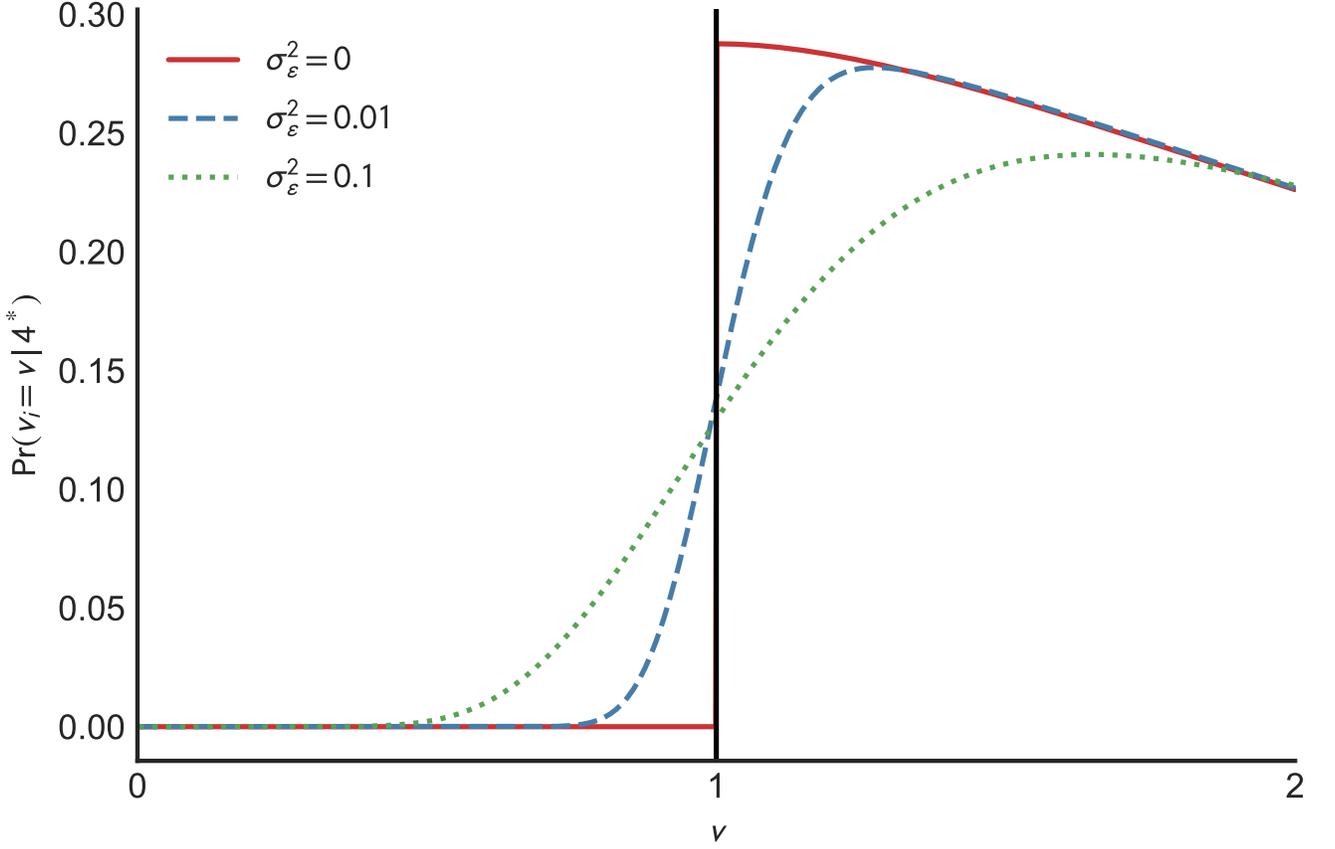
FIG. B.2. Conditional probability that publication i has value v_i given that it was awarded four stars. Three scenarios are considered for the inaccuracy of peer review σ_ϵ^2. The capability to produce high-value publications equals $\mu_{ku} = 1$. The threshold for being awarded four stars equals $p^{4^*} = 1$.

can be expressed as

$$\Pr(v_i = v \mid i \text{ has } 4^*) = \Pr(v_i = v \mid p_i > p^{4^*})$$

$$= \Pr(p_i > p^{4^*} \mid v_i = v)\frac{\Pr(v_i = v)}{\Pr(p_i > p^{4^*})}$$

$$= \Pr\left(\epsilon_i > \frac{p^{4^*}}{v}\right)\frac{\Pr(v_i = v)}{\Pr(p_i > p^{4^*})}.$$

As illustrated in Fig. B.2, the distribution of the value v_i conditional on publication i having four stars depends on the inaccuracy of peer review σ_ϵ^2. If peer review is perfectly accurate (i.e. $\sigma_\epsilon^2 = 0$), we have $p_i = v_i$. Consequently, if publication i has four stars, we have $p_i > p^{4^*}$, which then implies $v_i > p^{4^*}$. On the other hand, if peer review lacks perfect accuracy (i.e. $\sigma_\epsilon^2 > 0$), some publications have four stars even though $v_i \leq p^{4^*}$. These are publications for which reviewers have overestimated the value. For these publications, the error ϵ_i is sufficiently large so that $p_i > p^{4^*}$ even though $v_i \leq p^{4^*}$. Analogous considerations apply to $\Pr(v_i = v \mid i \text{ has no } 4^*)$.

For each institution k in UoA u, we sample values v_i' for all n_{ku} publications. For 4^* publications, we sample according to $\Pr(v_i' = v \mid i \text{ has } 4^*)$. For publications with fewer stars, we sample according to $\Pr(v_i' = v \mid i \text{ has no } 4^*)$. For institutions with higher μ_{ku}, we are more likely to sample higher values, as illustrated in Fig. B.3. Next, for each publication i, we sample a perceived value p_i'. This perceived value is given by $p_i' = v_i'\epsilon_i'$, where the error ϵ_i' is sampled according to Eq. B3. We then calculate the proportion of publications for which $p_i' > p^{4^*}$. This yields the proportion of publications of institution k in UoA u that would be awarded four stars based on the sampled perceived values. We denote this proportion by $\mathrm{PP}_{ku}'(4^*)$.

Finally, for each field u, we calculate the agreement between $\mathrm{PP}_{ku}(4^*)$ and $\mathrm{PP}_{ku}'(4^*)$ for all institutions k. This offers an indication of the uncertainty of peer review in field u. For the results presented in Section V, we resample
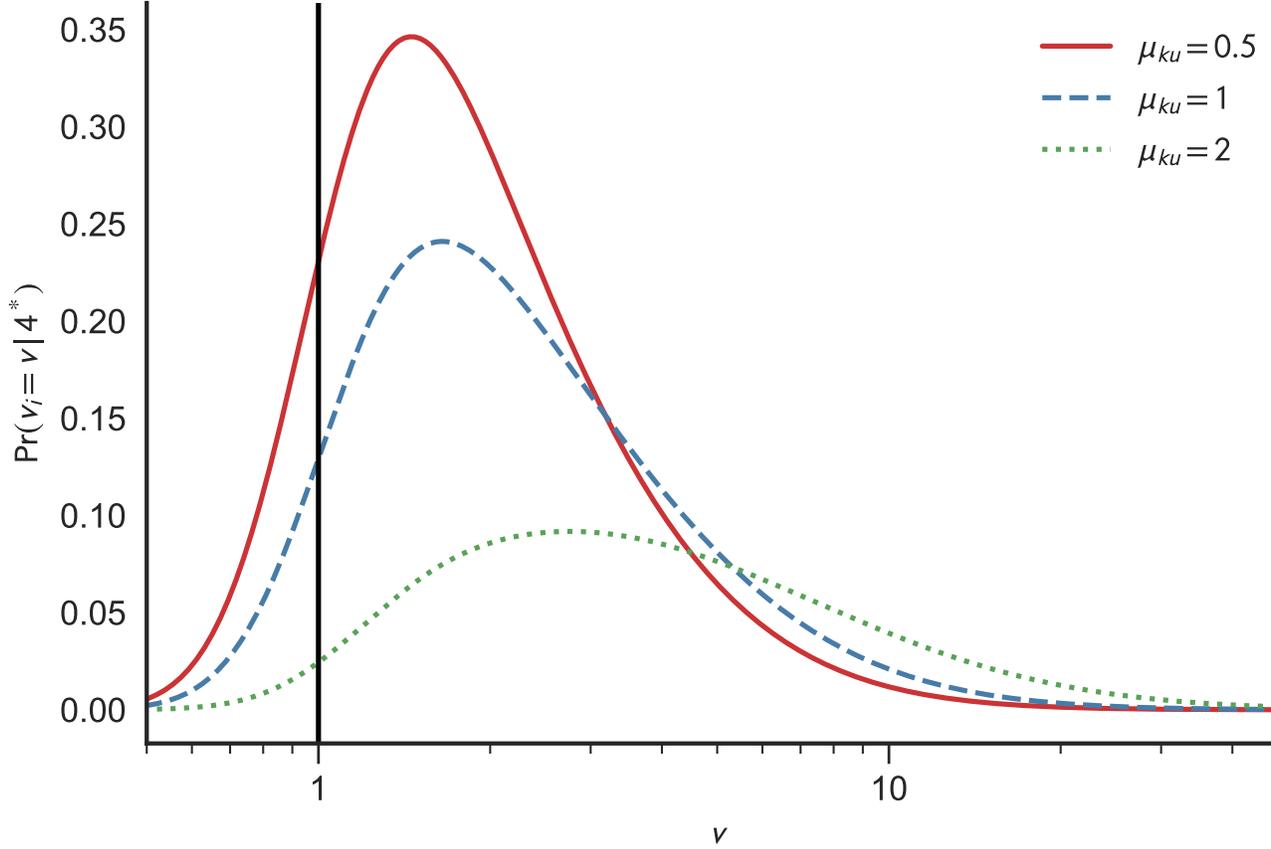
FIG. B.3. Conditional probability that publication i has value v_i given that it was awarded four stars. Three scenarios are considered for the capability to produce high-value publications μ_{ku}. The inaccuracy of peer review equals $\sigma_\epsilon^2 = 0.1$. The threshold for being awarded four stars equals $p^{4^*} = 1$.

$1\,000$ times. We report both the median outcome obtained from $1\,000$ samples and the interval that covers 95% of the outcomes.

Appendix C: Detailed results

TABLE C.1: Agreement per unit of assessment.

	Unit of assessment	Size-independent		Size-dependent	
		R	MAD	R	MAPD
1	Clinical Medicine	0.86	2.9	0.98	14.7
2	Public Health, Health Services and Primary Care	0.83	3.0	0.98	15.8
3	Allied Health Professions, Dentistry, Nursing and Pharmacy	0.54	4.8	0.95	25.6
4	Psychology, Psychiatry and Neuroscience	0.79	4.0	0.96	27.1
5	Biological Sciences	0.69	9.0	0.98	31.8
6	Agriculture, Veterinary and Food Science	0.68	4.8	0.96	27.1
7	Earth Systems and Environmental Sciences	0.80	3.1	0.93	25.2
8	Chemistry	0.83	4.3	0.94	23.2
9	Physics	0.86	2.8	0.98	12.1
10	Mathematical Sciences	0.39	5.9	0.96	33.4

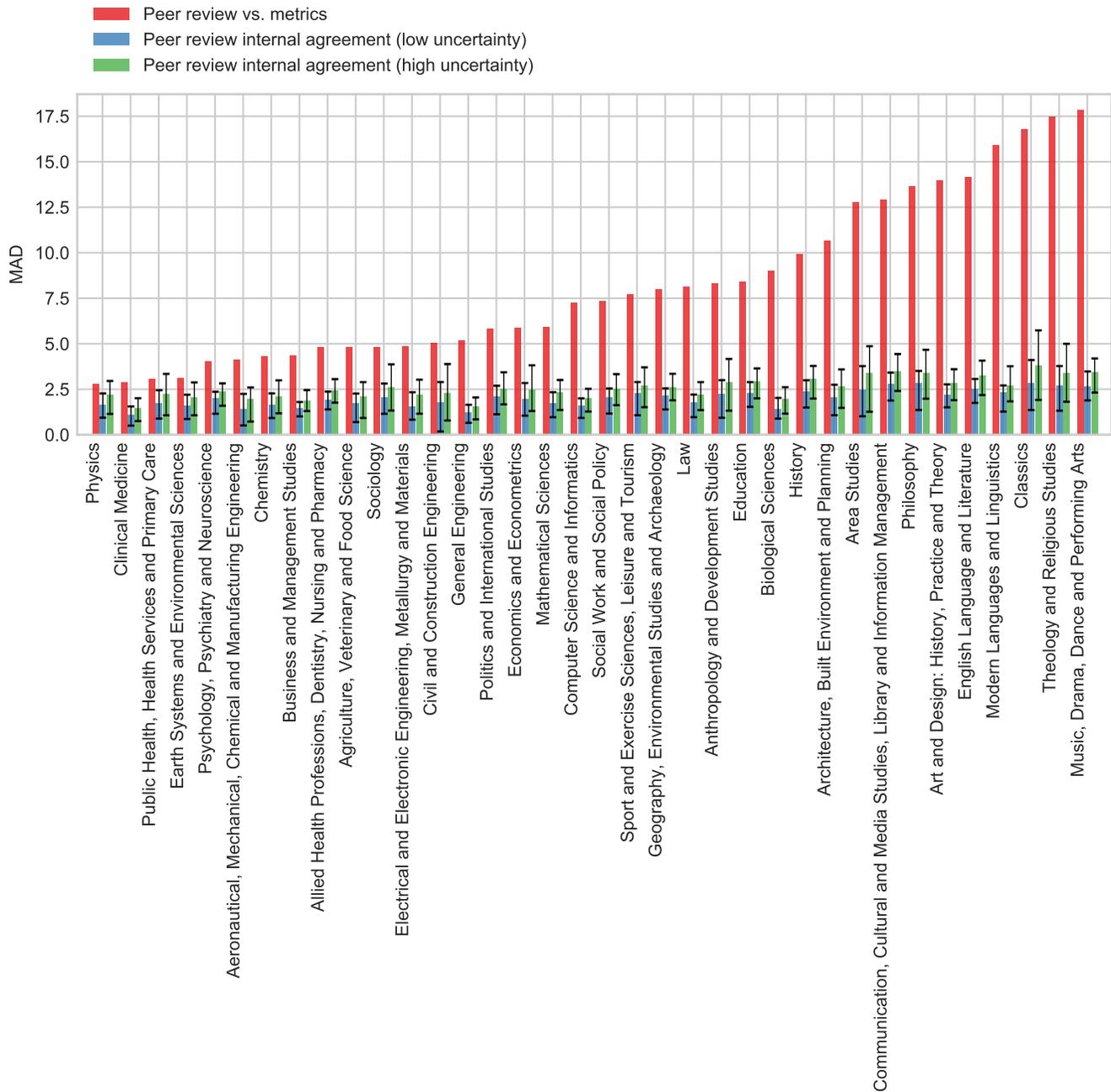|	Unit of assessment	Size-independent		Size-dependent	
		R	MAD	R	MAPD
11	Computer Science and Informatics	0.49	7.3	0.87	46.1
12	Aeronautical, Mechanical, Chemical and Manufacturing Engineering	0.39	4.1	0.98	29.8
13	Electrical and Electronic Engineering, Metallurgy and Materials	0.77	4.9	0.91	35.9
14	Civil and Construction Engineering	0.28	5.1	0.87	28.7
15	General Engineering	0.51	5.2	0.90	43.0
16	Architecture, Built Environment and Planning	0.29	10.6	0.83	78.5
17	Geography, Environmental Studies and Archaeology	0.68	8.0	0.85	45.1
18	Economics and Econometrics	0.89	5.9	0.95	27.4
19	Business and Management Studies	0.60	4.4	0.96	33.7
20	Law	0.40	8.1	0.87	85.2
21	Politics and International Studies	0.52	5.8	0.90	38.9
22	Social Work and Social Policy	0.36	7.3	0.88	44.6
23	Sociology	0.27	4.8	0.86	25.9
24	Anthropology and Development Studies	0.10	8.3	0.63	42.3
25	Education	0.41	8.4	0.95	45.9
26	Sport and Exercise Sciences, Leisure and Tourism	0.35	7.7	0.84	43.8
27	Area Studies	0.10	12.8	0.48	61.3
28	Modern Languages and Linguistics	0.35	15.9	0.58	80.2
29	English Language and Literature	0.34	14.2	0.56	62.2
30	History	0.40	9.9	0.90	56.1
31	Classics	0.21	16.8	0.49	63.3
32	Philosophy	0.42	13.6	0.83	52.3
33	Theology and Religious Studies	0.35	17.5	0.62	87.9
34	Art and Design: History, Practice and Theory	0.39	14.0	0.56	98.7
35	Music, Drama, Dance and Performing Arts	0.33	17.8	0.55	97.6
36	Communication, Cultural and Media Studies, Library and Information Management	0.42	12.9	0.68	72.4

FIG. C.1. 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 for all units of assessment.
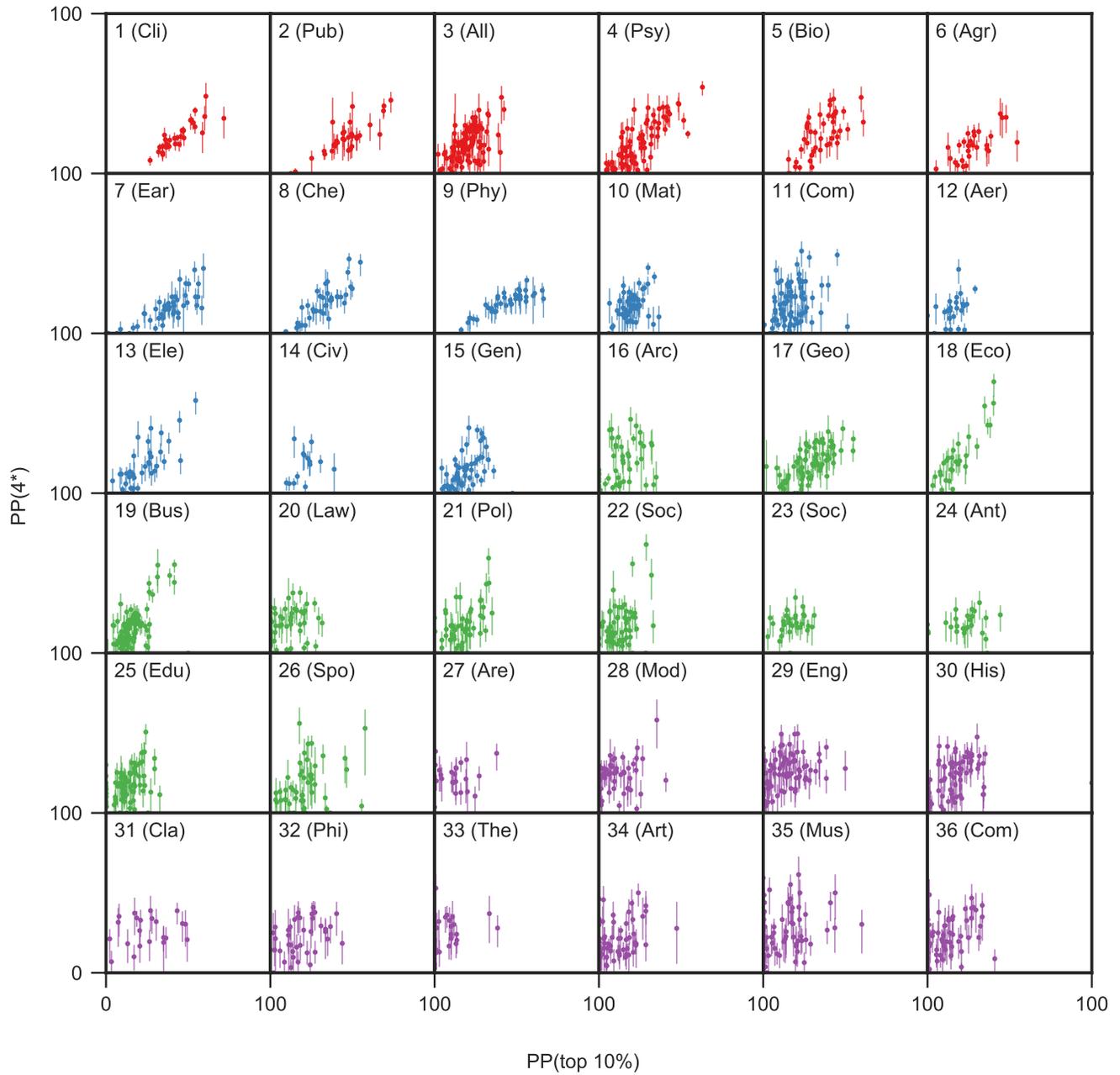
FIG. C.2. Scatter plots of PP(top 10%) and PP(4*) at the institutional level for all units of assessment. Error bars indicate the 95% interval of bootstrapped peer review results for low peer review uncertainty.
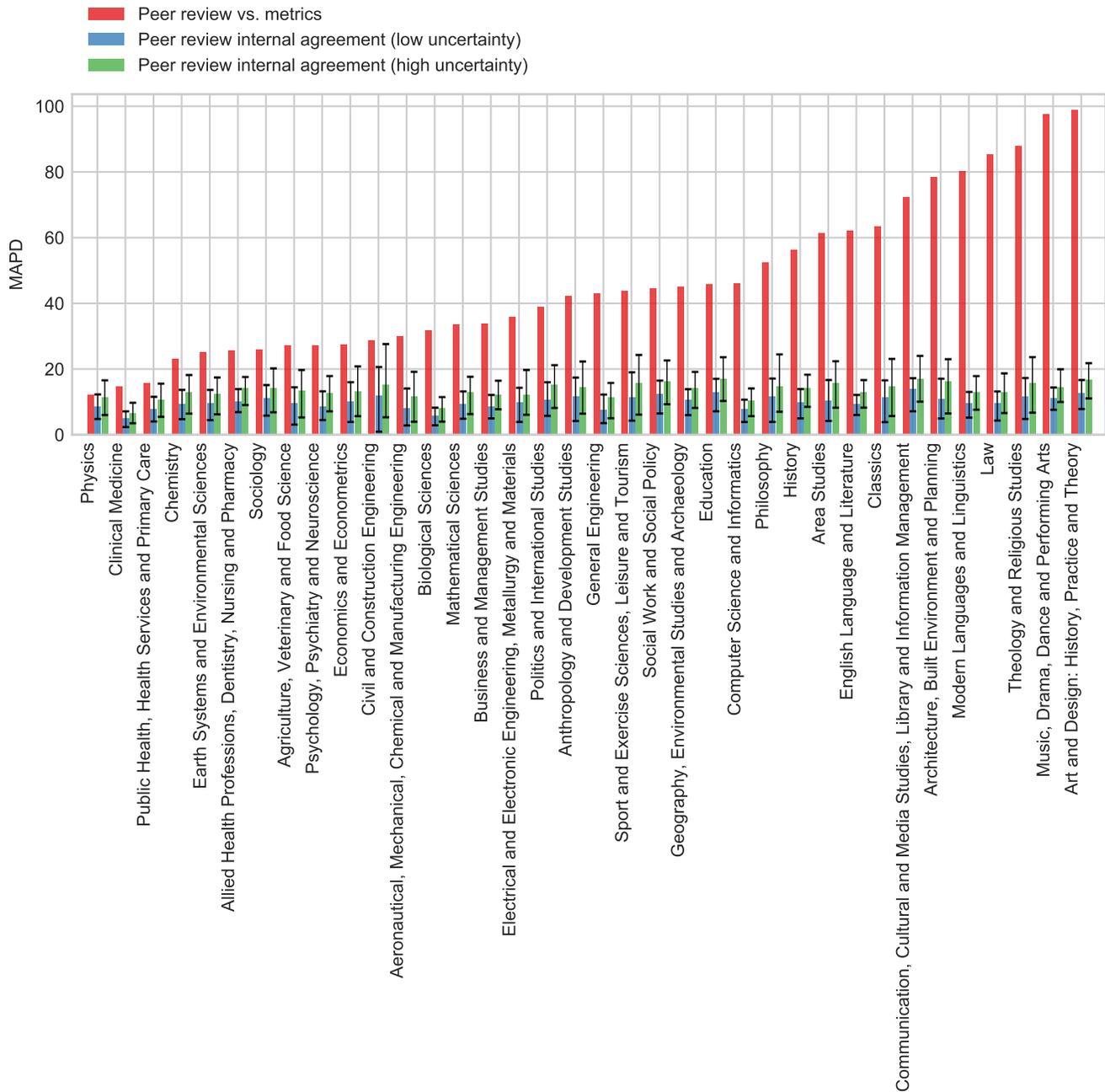
FIG. C.3. Size-dependent median absolute percentage difference (MAPD) of P(top 10%) relative to P(4*) compared with the MAPD based on a model of peer review uncertainty for all units of assessment.
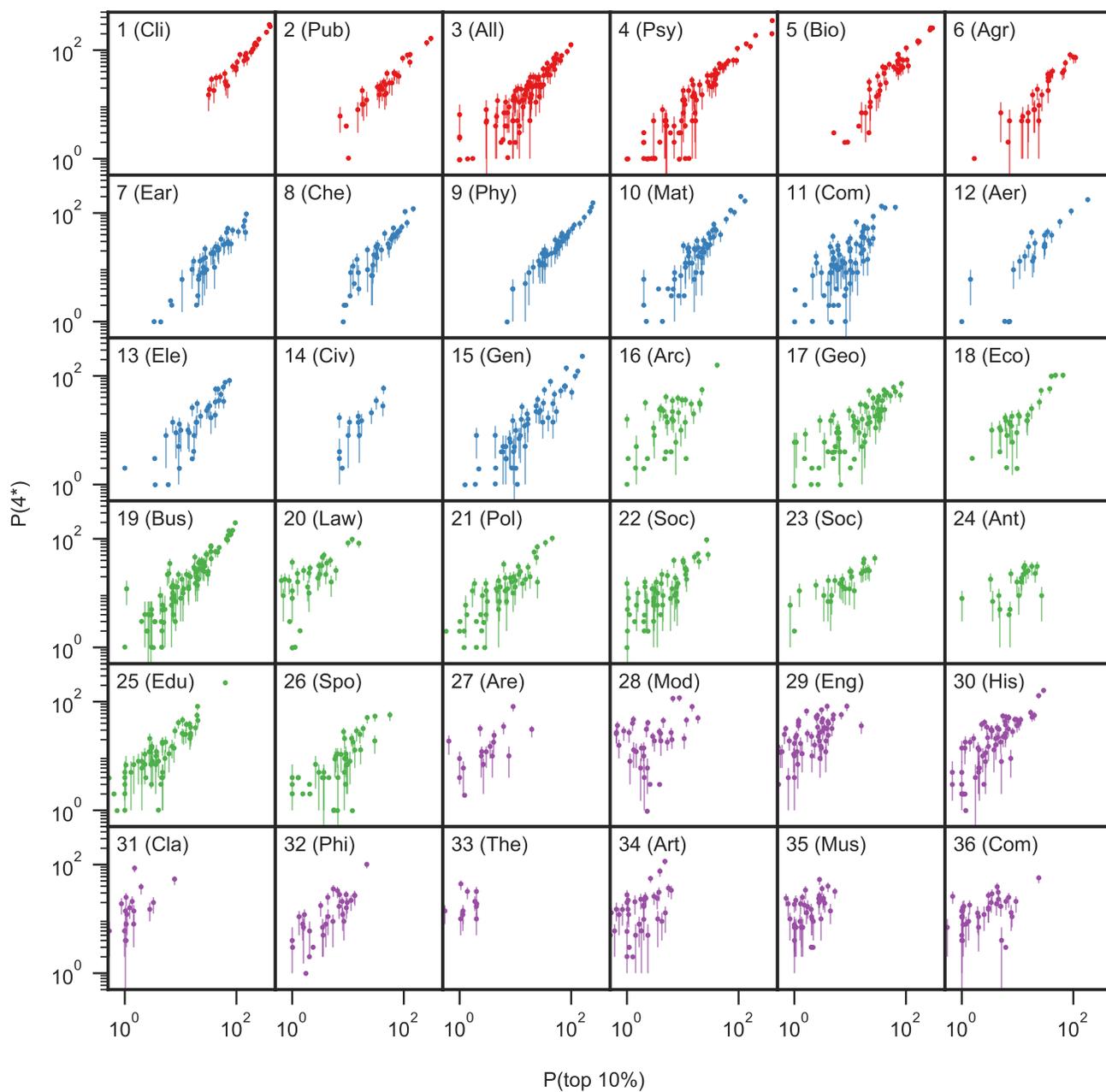
FIG. C.4. Logarithmic scatter plots of P(top 10%) and P(4*) at the institutional level for all units of assessment. Error bars indicate the 95% interval of bootstrapped peer review results for low peer review uncertainty.