Scientometrics (2013) 97:767–777
DOI 10.1007/s11192-013-1058-9

Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence

O. Mryglod · R. Kenna · Yu. Holovatch · B. Berche

Received: 13 November 2012 / Published online: 18 June 2013

Abstract Many different measures are used to assess academic research excellence and these are subject to ongoing discussion and debate within the scientometric, university-management and policy-making communities internationally. One topic of continued importance is the extent to which citation-based indicators compare with peer-review-based evaluation. Here we analyse the correlations between values of a particular citation-based impact indicator and peer-review scores in several academic disciplines, from natural to social sciences and humanities. We perform the comparison for research *groups* rather than for individuals. We make comparisons on two levels. At an absolute level, we compare total impact and overall *strength* of the group as a whole. At a specific level, we compare academic impact and *quality*, normalised by the size of the group. We find very high correlations at the former level for some disciplines and poor correlations at the latter level for all disciplines. This means that, although the citation-based scores could help to describe research-group strength, in particular for the so-called hard sciences, they should not be used as a proxy for ranking or comparison of research groups. Moreover, the correlation between peer-evaluated and citation-based scores is weaker for soft sciences.

Keywords Peer review · Citations · Research assessment exercise (RAE) · Research excellence framework (REF)

O. Mryglod (✉) · Yu. Holovatch
Institute for Condensed Matter Physics of the National Academy of Sciences of Ukraine,
1 Svientsitskii Str., Lviv 79011, Ukraine
e-mail: olesya@icmp.lviv.ua

R. Kenna
Applied Mathematics Research Centre, Coventry University, Coventry CV1 5FB, UK

B. Berche
Université de Lorraine, Campus de Nancy, B. P. 70239, 54506 Vandoeuvre lès Nancy Cedex, France

Introduction

Although it is not without critics, peer-review is mostly considered, amongst the broad academic community, to be the most reliable approach to assess the quality of academic research (van Raan 2005; Derrick et al. 2001). However because it is expensive, time-consuming and dependent on different circumstances (the so-called Hawthorne effect, see Bornmann (2012)), it is tempting to seek other approaches and citation-based indicators offer an obvious alternative (Nature 2010; Warner 2003). Numerous scientometric indicators based on the numbers of citations as well as the general number of publication and other aspects were proposed during the past half-century (e.g., see Garfield 1955, 1973; Hirsch 2005; Egghe 2006). The real challenge is to invent a simple but reliable way to assess individual or collective scientific performance. Sophisticated normalization procedures and different approaches have been designed to overcome the well-known nuances of citation counts (Vinkler 2001, 2003; Moed 2005). But this remains a problem of current importance today. For over half a century, scientists and research managers have discussed the merits and drawbacks of each approach. For practicing academics the accuracy and reliability of peer review broadly wins out (see, e.g., Derrick et al. 2001; Donovan 2007; Bornmann et al. 2008 and references therein). University managers, policy makers and the media, however, are attracted to the simplicity and economy of citation-based methodologies. Each approach is beset by ambiguities and problems and it is frequently argued that a combination may be needed to minimize the shortcomings of each. To achieve this, the technical and methodological limitations need to be clear (van Raan 2005). Here we address the question of whether a set of automated, scientometric or bibliometric indicators is a suitable substitute for, or component of, peer-review *at the level of the research group or department*.

The importance of evaluation of research quality at institutional level is exemplified by the growing number of reports produced by private companies and governmental bodies which document research performance of Higher Education Institutions within nations and worldwide (e.g., Butler 2010; Williams 2012; Evidence 2012). The Research Assessment Exercise (RAE) and Research Excellence Framework (REF) are examples of such processes nationally in the UK, and the Shanghai Academic Ranking is a famous example on an international scale (ARWU 2012). The Shanghai Ranking, in particular, is widely known but heavily criticised by the scientometric community (Florian 2007; Billaut 2010; Ioannidis et al. 2007).Despite well-known weaknesses of different systems for ranking universities, these are of increasing importance in many developed countries, which seek to organize national assessments of research. Many aspects of the UK's RAE, in particular, have been imitated in other countries (Macilwain 2010).

In a recent paper (Mryglod et al. 2012) we compared a citation-based indicator developed by Thomson Reuters Research Analytics (previously known as *Evidence*, see Evidence web-page 2012) to the peer-review-based RAE which was conducted in the UK in 2008. *Thomson Reuters* is one of the world's leading providers of scientometric information and performance measures for academic and research institutions, governments, not-for-profit organisations, funding agencies, and others with a stake in research. E.g., *Thomson Reuters*' (formerly the *Institute for Scientific Information*) *Web of Knowledge* is an important platform for information on citations in the sciences, social sciences, arts, and humanities. Using biology research institutions as a test case, we examined the correlations between results from both approaches at an amalgamated, research-group or department level. We made the comparison at two levels which we termed "absolute" and "specific". "Absolute" measures refer to the totality of group strength—the research

performance of the group as a whole. "Specific" quantities are normalised per head—the average strength, taken per group member. In this sense, "absolute strength" is the "volume of quality". E.g., the *absolute* citation count for a department in a given period is the total number of citations to the department's work, irrespective of how many researchers that department contains. The corresponding *specific* citation count is then the average number of citations per head (see, for example, Vinkler 2001, 2003).

Thus, the estimates of research "quality" and research "strength" introduced in Kenna and Berche (2010, 2011) are specific and absolute notions, respectively. We showed that the citation-based *specific* measure i provided by Thomson Reuters Research Analytics is not a good proxy for the peer-review *specific* measure s provided by RAE, in that these two measures are rather poorly correlated. However, when scaled up to the actual size N of a department (here and below N means the number of researchers in group), the *absolute* citation impact $\mathcal{I} = iN$ is very strongly correlated with the overall strength $\mathcal{S} = sN$ as measured by peer review. This is important because funding in the UK is determined on the base of strength $\mathcal{S}$ rather than quality s.

Another important feature of our previous analyses was that they focused on the research quality and strength of *groups* rather than individuals (Mryglod et al. 2012; Kenna and Berche 2010, 2011). In particular, the notion of two characteristic group sizes or "critical masses" was introduced in Kenna and Berche (2010, 2011). According to this concept, research performance is strongly dependent on group size up to a so-called upper critical mass N_c. Groups larger than N_c have either a reduced dependency of quality on quantity or no such dependency. A lower critical value N_k was also introduced in Kenna and Berche (2010, 2011) and interpreted as the minimum size a research department should achieve to be stable in the long term. These two critical masses, the values of which are strongly dependent on the research discipline, allow research groups and departments to be categorised as being *small* if they have size $N \leq N_k$, *medium* if $N_k \leq N \leq N_c$ or *large* if $N > N_c$. E.g., for the biological sciences analysed in the pilot study of Mryglod et al. (2012), the estimates for critical masses are $N_k = 10.4$ and $N_c = 20.8$ (Kenna and Berche 2010, 2011). (Fractions of staff are a feature of RAE (2008) in that Higher Education Institutes can include part-time researchers in their submissions and are counted as a proportion of full time equivalence). However, since small and medium research groups have the same linear dependency of quality on quantity (Kenna and Berche 2010) it is sensible to combine them in the correlation analysis. The strongest correlations between citation- and peer-review based measures of institutional strength for the biological sciences was observed for the large groups.

The implication of our previous analysis, therefore, is that citations, if used in an informed manner, could possibly be used as a proxy for departmental or group *strength* (and thus feed into funding requirements), provided that the departments are large. For smaller departments, however, peer review remains essential to determine *strength*. Moreover, citation-based indicators should not be used *in isolation* to estimate research *quality* for large, medium or for small groups.

It is natural to ask to what extent these conclusions cover other disciplines. Is there a difference between so-called hard and soft sciences or between the natural and social sciences and humanities? One might expect to observe differences due to different citation behaviour in different disciplines (Moed 2005; Stauffer 2012) and due to technical restrictions such as a smaller coverage by the *Web of Knowledge*. E.g., in the humanities, dissemination of original research through books is more common than in the natural sciences, and these are usually ignored in citation counting. These are the questions we address in this paper. We present quantitative results from comparisons of peer review and

citation-based indicators for several disciplines ranging from hard sciences to humanities. In particular, we consider chemistry; physics; mechanical, aeronautical and manufacturing engineering; geography and environmental studies; sociology and history.

Again, as in Mryglod et al. (2012), we used data from Thomson Reuters Research Analytics and the UK's 2008 version of the Research Assessment Exercise (called RAE 2008). As in the pilot study for biology, here we provide evidence that correlations between *specific* citation indicators and peer-measured group qualities for all the disciplines are very weak, even in the case of ranked values. However, when scaled up to the actual size of the department N, the absolute citation impact is strongly correlated with the overall group strength as measured by peer review. The correlation is very strong (above 95 %) for the hard sciences, less strong for geography and engineering, and weakest for social sciences (below 90 %). Although the correlations of measures are statistically strong for all the disciplines examined. Since national assessment is linked to funding distribution, even small differences can involve large financial impact. Thus, the threshold of reliability of results should be very high. This means that our previous conclusions (Mryglod et al. 2012) indeed extend to the hard sciences, physics and chemistry. But they do not extend to beyond the natural sciences. The social sciences and humanities, in particular, require peer-evaluated measurements of both quality and strength.

Peer review and the normalised citation impact for research institutes

The RAE and the REF

Quality related funding forms one element of the UK's dual research-support system. Until now, this has been based on the RAE (2008) and the annual distribution of quality-related funding is over 2 billion euro. In the future it will be based on the REF (2012). The evaluation of the quality of academic research output forms the major component of each of these schemes. Using published criteria, RAE 2008 assessed submissions in each of 67 different subject areas (units of assessment) and awarded a profile for each of them. All submissions are related to the assessment period which is from 1 January 2001 to 31 July 2007 (RAE 2008). Co-authored publications were not counted *pro-rata* and how they were dealt with depended upon whether co-authors belong to the same, or different, submitting groups or departments. In the case of collaborations between different universities, two or more co-authors from different institutions may submit the same output. However, RAE had extensive guidelines on this matter and there was considerable variation on what was allowed in different disciplines. In physics, for example, where the number of co-authors of a given publication is fewer than five, it was normally assumed that each author made a significant contribution to the work and the overall quality of that output was credited to the quality profile. Where the number of authors was five or more, the physics sub-panel asked for evidence of the extent of an individual's contribution to the research output. In the case of two or more co-authors from a given submitted group, on the other hand, a co-authored publication was not normally ascribed more than once in a submission to this UOA from a particular department. A given paper could normally be submitted only once for a given institution in a given unit of assessment. E.g., in physics the equivalent of 1685.57 full-time scientists were submitted to the RAE. Each researcher could submit up to 4 publications for assessment. There was an average of 40.13 full-time equivalent authors per submission, which translates into of the order of 160 papers per group.

RAE experts seek to quantify the proportion of a department's or research centre's submitted work which falls into each of five quality bands. The highest band is denoted as 4* and represents world-leading research. Remaining bands are graded through 3*, 2* and 1* to the lowest quality level which is called "Unclassified" (RAE 2008). The RAE quality profile assigned to a given research group is represented by a set of values p_{n*}, which represent the percentage of a team's research which was rated n*. For example, the profile $p_{4*} = 25$, $p_{3*} = 20$, $p_{2*} = 35$, $p_{1*} = 15$, $p_U = 5$ would indicate that 25 % of a groups research is of world-leading quality; 20 % is of 3* (internationally excellent); 35 % is of 2* quality (recognised internationally) and 15 % is 1* (recognised nationally).

Governmental funding post RAE is determined by a formula which combines the quality scores in a weighted manner. While the formula is subject to regional and temporal variation (the latter often due to the influence of lobby groups) the one introduced by the Higher Education Funding Council for England (2009) immediately following RAE 2008 rated 4* and 3* research as being seven and three times the value of 2* work, while lower quality research was unrewarded. In Mryglod et al. (2012), we denoted the strength of a given research group by $\mathcal{S}$. This is defined as the volume of quality,

$$\mathcal{S} = sN, \tag{1}$$

where N is the size of the group of quality s. The amount of quality-related funding distributed by the *Higher Education Funding Council for England* to a given university after RAE is a function of its strength $\mathcal{S}$. While strength determines future funding, it is, of course, not sensible to rank groups or universities according to their $\mathcal{S}$ values because different group shave different sizes. However many media and managers readily rank according to the quality measures s (although this also neglects very strong size effects as pointed out in Kenna and Berche (2010, 2011).

At RAE, the overall quality profile is constructed by summing sub-profiles for three separate elements (quality of "outputs", quality of "environment" and quality of "esteem"), of which outputs play the strongest role. In the future, the REF will replace the RAE for peer-review, institutional research assessment (REF 2012). The main difference is that overall quality profile will consist of "outputs", "impact" and "environment" instead of "outputs", "esteem" and "environment". Here "impact" refers to non-academic impact (thus, not e.g., citations). This new element is one of the major innovations of REF. But obviously, the very question about applicability of scientific results since long ago has been considered as one of the aspects of scientific productivity. Nevertheless, the "outputs" sub-profile remains the most important component of research assessment within REF providing the 65 % of Overall score. (The remaining 35 % is distributed between "impact" (20 %) and "environment" (15 %) (REF 2012)). To summarise, peer-review measures of research outputs will continue to dominate the UK's assessment of institutional research quality and strength in the years to come, and will be the main factor upon which billions of euros worth of funding will be allocated.

Although they may be influenced by non-academic impact and environment (e.g., visibility), citation counts refer only to outputs. Therefore it is sensible to compare citation-based measures with the "outputs" category of RAE. These are readily available on the official RAE (2008) web-page and we will henceforth confine our attention to these measures. To maintain consistency of notation with respect to Mryglod et al. (2012), we denote by s_1 the peer-review measure of quality coming from the "outputs" category of RAE 2008. The corresponding absolute measure is denoted by $\mathcal{S}_1 = s_1 N$.

Thomson Reuters Research Analytics citation indicator

As described in Mryglod et al. (2012), our citation-based measure of choice is that provided by Thomson Reuters Research Analytics (2012). This company offers a service analysing research performance tailored to individual client requirements. They have developed the so-called normalised citation impact (NCI) i as a coefficient of departmental performance in a given discipline.

Thomson Reuters Research Analytics calculate the NCI using data from *Web of Knowledge* databases (Evidence 2010, 2011). Similarly to relative citation rate (RCR) (i.e., (Schubert 1996)), the NCI is calculated by comparing to a mean or expected citation rate. It is a *specific* measure of academic citation impact because it is averaged over the entire research group. A non-trivial advantage of the NCI is that it takes account of different citation patterns between different academic disciplines. To achieve this, the total citation count for each paper is first normalised to an average number of citations per paper for the year of publication and either the field or journal in which the paper was published. This is called "rebasing" the citation count (Evidence 2011). To compare sensibly with the UK's peer-review mechanism, only the four papers per individual which were submitted to RAE 2008 were taken into account by Thomson Reuters Research Analytics in order to determine the average NCI for research groups (citation data till the end of 2009 were analysed, see (Evidence 2011), Appendix A).

Thus, the NCI may be considered as a citation-based *specific* measure of the academic impact of a department in a given field and we denote it by i. The corresponding *absolute* measure of impact (the total volume of academic impact of the department or group) is denoted by $\mathcal{I}$. The relationship between the two is

$$\mathcal{I} = iN. \tag{2}$$

Comparisons to be made

The objective of the remainder of this paper is to compare the peer-review and citation-based indicators for different disciplines. The *specific* indicators to compare are quality and citation impact s_1 and i as measures of the average strength and impact of the group or department *per individual* contained within it. We also compare the *absolute* indicators $\mathcal{S}_1$ and $\mathcal{I}$ as measures of the overall strength and total impact of the group as a whole.

Weak correlation between specific measures of quality and impact

A 100 % linear correlation between i and s_1 would indicate that the citation-based indicator (NCI) is a perfect proxy for RAE peer-review quality scores. The actual correlations for different disciplines are depicted in Fig. 1 and are far from perfect.

For the majority of disciplines one can observe some positive but weak correlation. This is quantified by a relatively small values of the Pearson coefficient r, the values of which are listed in Table 1. The conclusion is clear—the NCI indicators should not be used in place of peer-review measures of research-output quality.

As stated, normalized scores (be they RAE quality measurements or NCI citation-based indicators) are frequently used for ranking research groups. For this reason we also check the correlation between *ranks*. The ranks are constructed by listing ratings of research groups ascending order of their corresponding scores. Then each department is assigned an

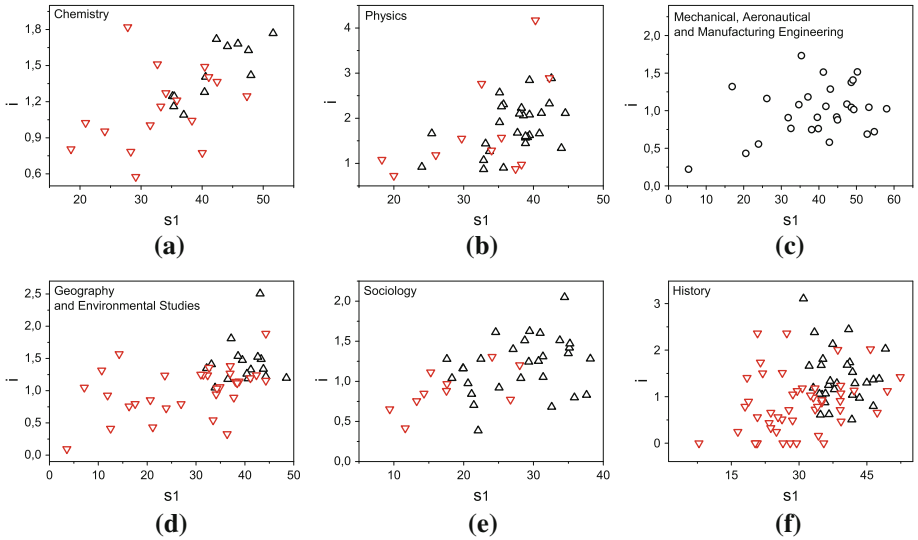

Fig. 1 Correlations between average quality of research groups according to RAE 2008 s_1 and average excellence of research groups according to normalised citation impact i for: **a** chemistry, **b** physics, **c** mechanical, aeronautical and manufacturing engineering, **d** geography and environmental studies, **e** sociology and **f** history. *Different symbols* represent large (*black up pointing triangle*) and medium/small (*red down pointing triangle*) groups. For engineering **c** information about group sizes is unavailable

Table 1 The approximate values of linear correlation coefficients between specific values s_1 and i calculated for several different disciplines

Description of the data sets	Pearson coefficient r			Spearman coefficient of ranked values ρ
	All groups	Large groups	Medium / small groups	
Biology (see Mryglod et al. (2012) (44 groups: 32 large, 7 medium, 5 small)	**0.60**	**0.57**	0.35	**0.53**
Chemistry (29 groups: 12 large, 14 medium, 3 small)	**0.60**	**0.82**	0.34	**0.62**
Physics (41 groups: 28 large, 9 medium, 4 small)	**0.48**	**0.45**	0.54	**0.53**
Mechanical, aeronautical and manufacturing engineering (30 groups)	0.34	–	–	0.18
Geography and environmental studies (41 groups: 28 large, 9 medium, 4 small)	**0.51**	0.13	**0.42**	**0.47**
Sociology (39 groups: 29 large, 8 medium, 2 small)	**0.49**	0.29	**0.64**	**0.47**
History (79 groups: 30 large, 24 medium, 25 small)	**0.34**	<0	0.27	**0.38**

Statistically significant values are highlighted in boldface (with significance level $\alpha = 0.05$)

ascending numerical rank (the average rank in the case of the equal scores). The linear correlation strength between ranked variables is expressed by Spearman coefficient ρ and these are also listed in Table 1.

Contrary to some earlier results which claimed high levels of correlation between the ranked RAE scores and citation counts (for example, $\rho = 0.80$ for music (Oppenheim and Summers 2008) and $\rho = 0.81$ for archaeology (Norris and Oppenheim 2003), our values of Spearman coefficient are low, varying from 0.18 to 0.62. This is perhaps unexpected, since the NCI i is a more sophisticated citation-based measure of academic impact compared to simple citation counting which was used in earlier analyses.

As stated earlier, it was established in Kenna and Berche (2010, 2011) that the dependency of research quality on quantity of researchers differs depending on whether or not research groups exceed the upper critical mass N_c. For this reason, we also investigate these categories (large and medium/large groups) separately. The correlation coefficients between s_1- and i-values for large and for medium/small groups are also listed in Table 1. As one can see, the proportions of groups with $N > N_c$ and $N < N_c$ differ for different disciplines: whereas the sociological groups are mainly large, there is a high proportion of small/medium groups in the field of geography and environmental sciences. However, this division does not help and the correlation coefficients for the two *specific* measures of group-research performance are poor. (Further subdivision of the $N < N_c$ category into separate sets of small and medium sized groups does not ameliorate the situation.) We conclude that the NCI is a poor proxy for peer review measures of research quality in all subject areas analysed.

Strong correlation between absolute measures of strength and impact

A conspicuous feature of the above analysis is that all research groups are treated as contributing the same weight to the analysis. For example, the RAE-measured quality scores for the history-research groups at the Open University and the University of Glamorgan are almost equal: $s_1 \approx 34$. But, with 20.6 staff, the former is more than 3 times bigger than the latter which has only 6 researchers. This means that researchers in smaller groups contribute more weight to the analysis, and statistical inaccuracies in their scores are unduly amplified. This problem is remedied by multiplying the average quality of groups by their size, a process which also renders the specific measures absolute: quality becomes strength and the NCI is also scaled up to the volume of the group or department.

From Fig. 2, there are clear correlations between $\mathcal{S}_1$ and $\mathcal{I}$ for all disciplines studied. The corresponding values of Pearson coefficient are given in Table 2. The values of the correlation coefficients r for the six disciplines studied here vary from 0.87 to 0.96. For comparison, the equivalent statistic for the biology research groups studied in Evidence (2010) was 0.97. As in biology, the replacement of specific measures of quality and impact by their absolute counterparts has the effect of stretching the corresponding axes by amounts proportional to the quantity of the groups or departments, and this leads to improved correlations.

As observed previously for biology (Mryglod et al. 2012), the correlation between $\mathcal{S}_1$ and $\mathcal{I}$ is usually best for large groups. The only exception is geography: in this case medium and small groups exhibit a better correlation than large ones. One may speculate as to the reasons for this. One possibility is the highly interdisciplinary nature of the research, which includes "a wide range of enquiries into natural, environmental and human phenomena" (RAE 2008). Indeed, among the disciplines analysed in this paper, only the geographical unit of assessment was declared as highly interdisciplinary and this marks it out.

While the r-values are high for all the disciplines, there is a noticeable difference between the hard sciences (chemistry, physics and biology (Mryglod et al. 2012)) and

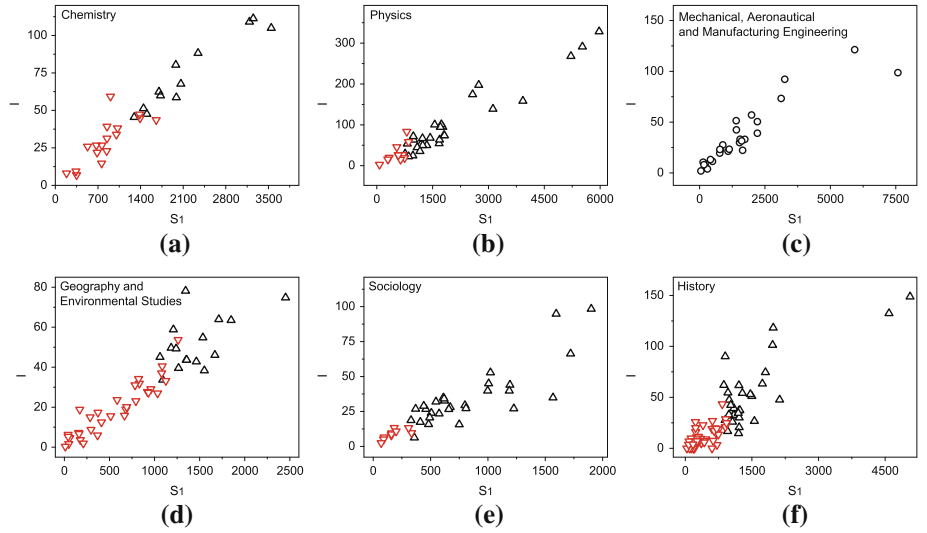

Fig. 2 Correlation between $\mathcal{S}_1$ (strength of research groups according to RAE 2008) and $\mathcal{I}$ (absolute citation impact) for: **a** chemistry, **b** physics, **c** mechanical, aeronautical and manufacturing engineering, **d** geography and environmental studies, **e** sociology and **f** history. The *symbols* are as in Fig. 1

Table 2 The approximate values of the linear correlation coefficients between $\mathcal{S}_1$ and $\mathcal{I}$ for several disciplines

Description of the data sets	Pearson coefficient r		
	All groups	Large groups	Medium / small groups
Biology (44 groups: 32 large, 7 medium, 5 small)[a]	**0.97**	**0.96**	**0.90**
Chemistry (29 groups: 12 large, 14 medium, 3 small)	**0.96**	**0.96**	**0.79**
Physics (41 groups: 28 large, 9 medium, 4 small)	**0.96**	**0.96**	**0.67**
Mechanical, aeronautical and manufacturing engineering (30 groups)	**0.92**	–	–
Geography and environmental studies (41 groups: 28 large, 9 medium, 4 small)	**0.92**	**0.56**	**0.93**
Sociology (39 groups: 29 large, 8 medium, 2 small)	**0.88**	**0.82**	**0.73**
History (79 groups: 30 large, 24 medium, 25 small)	**0.88**	**0.79**	**0.66**

Statistically significant values (with significance level $\alpha = 0.05$) are highlighted in boldface

[a] The correlation coefficients for biology given in Mryglod et al. (2012) were based on the overall quality profiles $\mathcal{S}$. Here, to properly compare with the other subject areas and with $\mathcal{I}$, the output-based absolute scores $\mathcal{S}_1$ are used instead

"softer" disciplined (history and sociology). For the former set, the correlation coefficient between absolute measures exceeds 95 %. For the latter set of disciplines it is smaller than 90 %. The interdisciplinary area of geography and environmental studies with $r \approx 0.92$ is positioned somewhere between these two categories as is the engineering discipline studied.

Conclusions

Based on the above results, the following three main conclusions may be drawn.

- *Weak correlations between specific measures of research quality and impact* have been observed for the disciplines of chemistry; physics; mechanical, aeronautical and manufacturing engineering; geography and environmental studies; sociology and history. This signals that this citation-based measure is a poor proxy for peer-reviewed measures of the quality of research groups. Moreover, since rankings are based on normalized data, this indicates that citation-based indicators will provide quite different rankings to those based on peer review.
- *Strong correlation between absolute measures of research quality and impact* which was previously observed for biology (Mryglod et al. 2012), is seen to extend to various extents to the disciplines which were analysed here. Thus, citation-based measures may inform or serve as a proxy for peer-review measures of the strengths of research groups.
- Although the citation-based measures could be a reasonable proxy for, or may inform about, the strengths of research groups for all disciplines studied, the results for the hard sciences are superior than those of the softer disciplines. Specifically, Pearson coefficients exceeding 95 % were observed for physics, chemistry as well as for biology while the corresponding values for history and sociology are below 90 %. The interdisciplinary areas of geography and engineering are in between with a linear correlation of ≈ 92 % between absolute measures of scientific excellence.

Since quality-related funding is strength based, the use of citation-based indicators may offer a much cheaper, and less intrusive alternative to the system currently in use in the UK and some other countries for large research groups in the hard sciences. However, such a proxy would be far less reliable for the social sciences and humanities. Moreover, citation-based indicators should not be used in isolation to compare the average quality of Higher Education Institutions or separate research groups. Nor should they be used for rankings.

Acknowledgments This work was supported in part by the 7th FP, IRSES project No. 269139 "Dynamics and cooperative phenomena in complex physical and biological environments" and IRSES project No. 295302 "Statistical physics in diverse realizations". The authors thank Jonathan Adams from Thomson Reuters Research Analytics for the data and Ihor Mryglod for fruitful discussions.

References

Billaut, J. -C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shanghai ranking? *Scientometrics 84*, 237–263.

Bornmann, L. (2012). The Hawthorne effect in journal peer review, *Scientometrics 91*, 857–862.

Bornmann, L., Wallon, G., & Ledin, A. (2008). Is the *h* index related to (standard) bibliometric measures and to the assessments by peers? An investigation of the *h* index by using molecular life sciences data, *Research Evaluation 17*, 149-156.

Butler, D. (2010). University rankings smarten up, *Nature, 464*,16–17.

Derrick, G. E., Haynes, A., Chapman, S., & Hall, W. D. (2001). The association between four citation metrics and peer rankings of research influence of Australian researchers in six fields of public health, *PLoS One, 6*, e18521.

Donovan, C. (2007). Future pathways for science policy and research assessment: Metrics vs peer review, quality vs impact. *Science and Public Policy 34*, 538-542.

Egghe, L. (2006). Theory and practise of the g-index, *Scientometrics,69*(1), 131–152.

Evidence. (2010). Evidence (a Thomson Reuters business) report. The future of the UK university research base, July 2010.
Evidence. (2011). Funding research excellence: Research group size, critical mass and performance. A University Alliance report, July 2011.
Evidence. (2012). Bibliometric evaluation and international benchmarking of the UK's physics research, Summary report prepared for the Institute of Physics by Evidence, Thomson Reuters.
Florian, R. V. (2007). Irreproducibility of the results of the Shanghai academic ranking of world universities *Scientometrics 72*, 25–32.
Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of Ideas. *Science, 122*(3159), 108–111.
Garfield, E. (1973). Citation frequency as a measure of research activity and performance in essays of an information scientist, *Current Contents, 1*, 406–408.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS, 102*(46),16569–16572.
Ioannidis, J. P. A. et al. (2007). International ranking systems for universities and institutions: A critical appraisal *BMC Medicine 5*, 30.
Kenna, R., & Berche, B. (2010). Critical mass and the dependency of research quality on group size. *Scientometrics, 86*(2), 527–540.
Kenna, R., & Berche, B. (2011). Critical masses for academic research groups and consequences for higher education research policy and management. *Higher Education Management and Policy, 23*(3), 1–21
Macilwain, C. (2010). Wild goose chase. *Nature, 463*, 291.
Mryglod O., Kenna R., Holovatch Y., Berche B. (2012). Absolute and specific measures of research group excellence. *Scientometrics* doi:10.1007/s11192-012-0874-7.
Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.
Nature. (2010). Editorial, Metrics Special. *465*, p. 845. Retrieved April 2012, from http://www.nature.com/metrics.
Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise. V archaeology and the 2001 RAE, *Journal of Documentation, 59*(6), 709–730.
Oppenheim C., & Summers M. A. C. (2008). Citation counts and the Research Assessment Exercise, part VI. Unit of assessment 67 (music), *Information Research, 13*(2).
RAE. (2008). The panel criteria and working methods. Panel E. (2006). Retrieved October 19, 2012, from http://www.rae.ac.uk/pubs/2006/01/docs/eall.pdf.
Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics 36*(3), 311–324.
Stauffer D. (2012). A biased review of sociophysics. *Journal of Statistical Physics* doi:10.1007/s10955-012-0604-9.
The official web-page of the RAE. (2008). Retrieved October 18, 2012, from http://www.rae.ac.uk/.
The official web-page of the Higher Education Funding Council for England. Funding for universities and colleges in 2009–10 (2009). Electronic Publication 01/2009 in the ADMIN-HEFCE Archives. Retrieved October 19, 2012.
The official web-page of Academic Ranking of World Universities (ARWU). (2012). Retrieved October 19, 2012, from http://www.shanghairanking.com.
The official web-page of Evidence Thomson Reuters. (2012). Retrieved October 18, 2012, from http://www.evidence.co.uk.
The official web-page of the REF. (2012). Retrieved October 19, 2012, from http://www.ref.ac.uk/.
van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods, *Scientometrics, 62*(1), 133–143.
Vinkler, P. (2001). An attempt for defining some basic categories of scientometrics and classifying the indicators of evaluative scientometrics. *Scientometrics, 50*(3), 539–544
Vinkler, P. (2003). Relations of relative scientometric indicators. *Scientometrics,58*(3), 687–694.
Warner, J. (2003). Citation Analysis and Research Assessment in the United Kingdom, *American Society for Information Science and Technology, 30*(1), 26–27.
Williams R., de Rassenfosse G., Jensen P., & Marginson S. (2012). U21 Ranking of National Higher Education Systems, Report of the project sponsored by Universitas 21, University of Melbourne.