Science and Public Policy, 34(8), October 2007, pages 575–583
DOI: 10.3152/030234207X255179; http://www.ingentaconnect.com/content/beech/spp

The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review

Henk F Moed

The paper discusses the strengths and limitations of 'metrics' and peer review in large-scale evaluations of scholarly research performance. A real challenge is to combine the two methodologies in such a way that the strength of the first compensates for the limitations of the second, and vice versa. It underlines the need to systematically take into account the unintended effects of the use of metrics. It proposes a set of general criteria for the proper use of bibliometric indicators within peer-review processes, and applies these to a particular case: the UK Research Assessment Exercise (RAE).

DURING THE PAST FEW DECADES, there has been increasing emphasis on the effectiveness and efficiency of government-supported research in most Organisation for Economic Co-operation and Development (OECD) countries. Governments need systematic evaluations for optimising their research allocations, re-orienting their research support, rationalising research organisations, restructuring research in particular fields, or augmenting research quality and productivity. Research policies were developed, aiming to stimulate research excellence and develop funding schemes in which the amount of funding of research groups or departments is dependent on their research quality.

Research evaluation plays a key role in these policies. It is carried out in various policy contexts and at various organisational levels: science policy of a national government by ministers responsible for scholarly research; research policy at the level of research organisations or institutions responsible for quality control and the allocation of research funds; and research management, carried out by directors of research groups or departments. A recent phenomenon is the installation by policy agencies of peer-review committees aiming to evaluate the past or expected future performance of research departments in scholarly institutions or disciplines. Typical examples are periodical evaluations by scholarly discipline of academic research in the UK and the Netherlands.

Such evaluations serve distinct objectives. They may primarily aim to provide departments subjected to evaluation with information that may enable them to improve their research performance. A second aim is to provide tools in decision-making processes about the allocation of research funds. A third objective is making research quality or scientific excellence manifest to the 'outside' world, that is, to scholars from other disciplines, potential external users of research results and the general public.

This paper focuses on basic research, defined as primarily being carried out to increase scholarly knowledge. Following Salter and Martin (2001), it includes both 'curiosity-driven' (sometimes denoted as 'pure') and 'strategic' or 'application-oriented' research. The latter is undertaken in a quest for a particular application, even though its precise details are not yet known.

Research evaluation may focus on a variety of aspects of research performance. This paper is primarily concerned with the assessment of the contributions scholars make in their research publications

Dr Henk F Moed is at the Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 52, Postbus 9555, 2300 RB Leiden, The Netherlands; Email: moed@cwts.leidenuniv.nl; Tel: +31 71 527 3940.

This paper summarises and further develops basic notions and conclusions outlined in Moed (2005).

 0302-3427/07/080575-9 US$08.00

Dr Henk F Moed is senior staff member at the Centre for Science and Technology Studies (CWTS) at Leiden University, the Netherlands. He has been active in quantitative science and technology studies since 1981. He obtained a masters degree in mathematics at the University of Amsterdam in 1978 and a PhD in science studies at the University of Leiden in 1989. He has published over 50 research articles and letters in international, peer-reviewed journals. He has been associate editor of *Scientometrics* since 1990 and an editorial board member of the *Journal of Informetrics* since 2007. He was winner of the Derek de Solla Price Award in 1999. He has been programme chair of several international conferences in the field. In 2005, he published a monograph, *Citation Analysis in Research Evaluation* (Springer), which is one of the very few textbooks in the field.

to the advancement of valid scholarly knowledge. Although it recognises the crucial importance of basic science for technological innovation, economic progress and social welfare, it does not deal with assessment of research activities from this perspective. Society supports university research because it expects benefits to rebound to society as a whole. A basic notion underlying exercises assessing the quality of scientific research is that better quality science is more likely to contribute effectively to desired social outcomes than science that is of a somewhat lower quality.

This paper focuses on the use of the Web of Science, a publication database published by Thomson Scientific (formerly the Institute for Scientific Information), and covering some 7,500 journals from all domains of science and scholarship. The general question addressed is: how should evaluation processes of the contribution basic research groups make to the advancement of scientific/scholarly knowledge be organised? What should be the role of peer review and that of bibliometric indicators in such processes?

The central thesis of this paper is expressed in its title: the future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. It argues that metrics, especially a sophisticated type of citation analysis, may provide tools to keep the peer-review process honest and transparent. Both metrics and peer review have their strengths and limits. A real challenge is to combine the two methodologies in such a way that the strength of the first compensates for the limitations of the second, and vice versa.

The structure of this paper is as follows. The next part focuses on metrics, especially on citation analysis. First, it addresses the question of what citations 'measure' and discusses the implications for their use in research evaluation. Then it briefly highlights a number of sophisticated indicators, but for more technical information it refers to other publications. There follows a brief discussion on the extent to which the Web of Science covers the various domains of science and scholarship, giving special attention to its coverage of social sciences and humanities. The choice of the aggregation level (groups versus individuals) is then discussed. The final section in this part underlines the need to systematically take into account the unintended effects that the application of various types of indicators may have on scientists' research and publication practices.

The following part deals with peer review, first by highlighting a number of general characteristics of peer-review processes. Most studies of these processes relate to the evaluation of submitted journal manuscripts and grant proposals. Then there is a brief presentation of the outcomes of some empirical studies on peer reviews of the research performance of research groups and departments.

The final part discusses how bibliometric indicators and peer review can be properly combined in research-assessment exercises. First, it draws conclusions from the two previous parts and proposes a set of general criteria for a proper use of bibliometric indicators within peer-review processes. The next section focuses on a particular case: the UK Research Assessment Exercise (RAE) and the current debate on how this should be organised in the future. It proposes elements that could be implemented in such exercises.

Strengths/limitations of bibliometric indicators

What do citations measure?

Referencing behaviour and citations can be studied and interpreted from various disciplinary viewpoints — for instance, information-scientific, sociological, physical, psychological or historical — and within a discipline from various perspectives or 'paradigms' — for instance, normative and micro-sociological (Cronin, 1984). In principle, all these perspectives illuminate referencing practices. It is therefore extremely difficult, if not impossible, to express what citations measure in a single theoretical concept that covers all the interpretations covered by the various approaches. Citation counts can be conceived as manifestations of intellectual influence (Zuckerman, 1987), but the concepts of citation impact and intellectual influence do not coincide.

Citation impact is a quantitative concept that can be operationalised in elementary, or in more sophisticated, ways, for instance, through crude citation counts or an advanced, normalised measure. These indicators are further discussed in later. Citation-impact indicators may be denoted as objective in the sense that they reflect properties of the cited documents themselves, they are replicable and based on the practices and perceptions of large numbers of (citing) scientists rather than on those of a single individual scientist (White, 1990).

Concepts such as 'intellectual influence' and 'contribution to scholarly progress' are essentially theoretical concepts of a qualitative nature, and can

Outcomes of citation analysis must be valued in terms of a qualitative, evaluative framework that takes into account the substantive content of the works under evaluation: this can be done by peers only

be assessed only by taking into account the cognitive contents of the work under evaluation. Distinct notions of the concept of intellectual influence may exist, and evaluators assessing scholarly work may have different views on which are the most crucial aspects to be taken into account.

Outcomes of citation analysis must be valued in terms of a qualitative, evaluative framework that takes into account the substantive content of the works under evaluation. This can be done by peers only. The conditions for proper use of bibliometric indicators at the level of individual scholars, research groups or departments tend to be more readily satisfied in a peer-review context than in a policy context. It can therefore be argued that bibliometric analyses at such lower aggregation levels normally best find their way to the policy arena through peer assessments.

Advanced bibliometric indicators

Citation analysis is more than generating raw publication and citation counts, or using so called 'journal impact factors'. Measurements of citation impact should always be interpreted as a function of the universe of citing publications, that is, the database in which they took place. Moreover, they have a comparative nature: the outcomes for a particular entity (such as a research group) make sense only if they are in some way related to those of other, similar entities.

Rather than relying on one single indicator, a series of indicators should be calculated and interpreted. Relative or normalised citation-impact measures play a special role. An important indicator calculates the ratio of the average citation impact of papers published by a research group and the world citation average in the subfields in which it is active (see, for instance, Narin, 1976; Vinkler, 1986; Braun *et al*, 1988). This indicator can be constructed in such a way that it takes into account not only differences in referencing practices among subfields, but also the type of publications subjected to a citation analysis (for instance, reviews versus original research articles) and their age distribution (Moed *et al*, 1995). It is known as a 'crown indicator' in bibliometric assessments of research performance.

Other types of indicator can be calculated as well, for instance, the number of a group's papers that are among the most highly cited in their field (see Narin, 1976; van Raan, 1996; 2004). Moreover, new, more sophisticated ones can be expected to become available in the future. Although journal-impact measures are useful tools in journal evaluation, there is a broad consensus among practitioners of information science and bibliometrics that these indicators cannot be used to predict 'actual' citation rates and therefore are no valid surrogates of the actual citation impact of a group's publications (see, for instance Garfield, 1996; Seglen, 1994).

Adequacy of coverage of Web of Science

Analysing the extent to which cited references in the Web of Science are themselves published in journals covered by this database, it can be shown that coverage of the Web of Science tends to be excellent in physics, chemistry, molecular biology and biochemistry, biological sciences related to humans and clinical medicine; good, yet not excellent, in applied physics and chemistry, engineering sciences, biological sciences related to animals and plants, geosciences, mathematics, psychology and psychiatry and other social sciences related to medicine and health. The coverage of the Web of Science was found to be moderate in other social sciences including sociology, political science, anthropology and educational sciences, and particularly in humanities (Moed, 2005).

A principal cause of non-excellent coverage is the importance of sources other than international journals, such as more nationally oriented journals, but also of non-journal sources, especially books and conference proceedings. In fields with a moderate ISI coverage, language or national barriers play a much greater role than they do in other domains of science and scholarship. In addition, research activities may be fragmented into distinct schools of thought, each with their own paradigms.

Therefore, it cannot be taken for granted that the Web of Science provides valid indicators of research performance in all subfields of these domains of scholarship. A challenge would be to systematically explore alternative data sources and methodologies. The expertise and perceptions of scholars active in the various subfields should play an important role in such an exploration. This paper focuses on fields for which the coverage of the Web of Science can be qualified as good or excellent. These fields are denoted by the term 'science'.

Research groups versus individuals

In science, the research group is the natural 'business' unit and therefore constitutes the most useful aggregation level in a citation analysis. Scientific research is the result of team work. A research group consists of a group leader, other senior scientists,

postdoctoral researchers and PhD students. Senior scientists may divide tasks among each other. Members of research groups tend to interact intensively one with another, and jointly carry out the group's research programme. In many areas of social sciences and humanities the organisational structure of research activities tends to be different from that in science. In these domains scholarly research tends to be more an individual activity.

Although research in science is teamwork, it does not follow that there are no differences in performance among individuals within a group. A crucial issue is the extent to which bibliometric analysis may be used to assess the performance of an individual working in a group, based on a citation analysis of the articles an individual publishes. In science, the publications (co-)authored by an individual researcher are often, if not always, the result of research to which other scientists have contributed as well, sometimes even dozens of them. The average number of authors of a paper is about four. Generally, performance of an individual on the one hand, and citation impact of the papers he or she (co-) authored on the other, relate to two distinct levels of aggregation.

Differences exist among groups with respect to authoring conventions of published papers. For instance, in some groups the group leader may as a rule co-author all papers emerging from the group, whereas in other groups PhD supervisors may not even be co-author of the articles published by their PhD students. Even if a list of authors in the byline of a paper properly reflects the list of contributors to the research, it does not directly reveal the size and nature of the contribution of each individual. The latter can be assessed properly only on the basis of sufficient background knowledge of the particular role of the scientist in the research presented in his/her publication oeuvre, for instance, whether this role has been leading, instrumental or technical.

Multi- and interdisciplinary research

New bibliometric methods provide dedicated approaches to multi- and interdisciplinary research (van Raan and van Leeuwen, 2002). A first creates and analyses a breakdown of the publication output of departments or groups into research fields, using a classification of research into about 200 scientific–scholarly subfields. Such a breakdown provides a clear impression of the research scope or 'profile' of the institute, department or research group. A measure of multi-disciplinarity can be derived from the distribution of the group's papers among fields. In this way groups with a multidisciplinary orientation can be distinguished from those following mono-disciplinary approaches. It needs emphasising that this breakdown of the publication output of an entity into fields does not necessarily coincide with a breakdown into its institutional parts.

A second methodology takes into account the origin of the citations given to the work of a particular group. It focuses on the cognitive orientation of 'knowledge users' and enables one to identify and measure trans-disciplinary impact, that is, impact on research activities in other subfields. Citation analysis is an appropriate tool for measuring this type of impact. It can provide measures of the trans- or interdisciplinary nature of an entity's citation impact. The two methodologies enable one to dedicate special attention to departments or groups that carry out multi- or interdisciplinary research. Therefore, there is no intrinsic reason why the use of bibliometric indicators would undervalue this type of research.

Effects of use of bibliometric indicators on scholars

Bibliometric investigators, and also other members of the scholarly community and the research policy arena, are increasingly aware of the need to analyse the effects of the use of bibliometric indicators in research evaluation (ranging from crude publication counts to sophisticated citation-impact measures) on the scholarly community and scholarly progress in general (see, for instance, Warner, 2003; Woolgar, 1991). One important issue is the effect of the use of citation analysis on scholars' publication and referencing practices. Evidence of these effects is often rather informal, or even anecdotal (Watkins, 2005). However, recent studies focusing on a 'formulaic' type of use of bibliometric indicators in policies allocating research funds examine these effects in a systematic way (Butler, 2003; 2004).

They make a valuable contribution to a deeper understanding of the actual and future role of citation analysis in research evaluation. This understanding contributes to the further development of the 'critical' potential of citation analysis as a research evaluation tool. It needs emphasising, though, that the crucial issue at stake is not whether scholars' practices change under the influence of the use of bibliometric indicators, but rather whether the application of such measures as a research evaluation tool enhances research performance and scholarly progress in general.

The crucial issue at stake is not whether scholars' practices change under the influence of the use of bibliometric indicators, but rather whether the application of such measures as a research evaluation tool enhances research performance and scholarly progress in general

A longitudinal bibliometric analysis of UK science covering almost 20 years revealed in the years prior to a Research Assessment Exercise (undertaken in 1992, 1996 and 2001), three distinct patterns, which can be interpreted in terms of scientists' responses to the principal evaluation criteria applied (Moed, 2008). When, in the 1992 RAE, total publications counts were requested, UK scientists substantially increased their article production. When a shift in evaluation criteria in the 1996 RAE was announced from 'quantity' to 'quality', UK authors gradually increased their number of papers in journals with a relatively high impact factor. Along the way towards the 2001 RAE, evaluated units in a sense shifted back from quality to quantity, particularly by stimulating their members to collaborate or at least to co-author more intensively, and thus increase the number of active research staff.

Later, this paper discusses how the proper use of sophisticated citation-based indicators as quality markers can shift the assessment's focus from prestige journals and publication counts towards research quality, as argued, for instance, by Lipsett and Fazackerley (2005). In this way, the unintended effects of RAE assessment criteria, either formal or anticipated, on UK scientists' publication and authoring practices can be reduced.

Strengths/limitations of peer review

General characteristics

Assessment of the quality of research carried out by a group requires a detailed knowledge of the specific topics in which the group is active, in order to evaluate criteria such as the methodological soundness of the research and the (potential) significance of its contribution to scientific progress, both in the narrow speciality and from the wider disciplinary perspective. Only peers tend to have such knowledge; this is why peer review has always been such an important instrument in quality control in science. However, peer review also has its limitations and biases.

Ben Martin and John Irvine (1983) described peer evaluation as a method:

> based on individual scientists' perceptions of contributions by others to scientific progress, perceptions arrived at through a complicated series of intellectual and social processes, mediated by factors other than the quality, importance or impact of the research under evaluation.

They identified three major problems in using the outcomes of peer evaluation in a policy context. First, evaluators may be influenced by political and social pressures within the scientific community, such as the possible implications of their judgements for their own work and that of their colleagues. Secondly, peer reviewers tend to evaluate in terms of their own research interests, and may not possess all the knowledge that is needed to form a balanced judgement. Finally, peers tend to conform to conventionally accepted patterns of belief, and may, for instance, be influenced by a scientist's reputation rather than his or her actual contribution to scientific progress.

Langfeldt reviewed numerous studies of peer-review processes. She identified one group of studies focusing on the degree of agreement among reviewers ('reliability') and the effects of possible biases, and a second group primarily analysing evaluation criteria applied by review panels (Langfeldt, 2001). Studies from the first group tended to report low degrees of agreement among reviewers and identified various kinds of bias, including the applicant's academic status and gender, and institutional and cognitive bias. The second group of studies revealed that reviewers tend to use a common set of evaluation criteria. She concluded that the combination of these findings indicate that "while there is a certain set of criteria that reviewers pay attention to — more or less explicitly —, these criteria are interpreted or operationalised differently by various reviewers" (Martin and Irvine, 1983: 821).

Empirical studies on peer review of research groups

Most studies of peer review processes relate to the evaluation of submitted journal manuscripts and grant proposals. Thus far, little research has been undertaken on the evaluation processes of the quality of research groups or departments. Peer-review processes are normally carried out without documentation of the bases for conclusions. It is therefore difficult to assess the extent to which citation and publication data are used in peer review. What we can do is analyse statistical relationships between peer judgements or ratings of departments on the one hand, and bibliometric indicators calculated for these departments on the other.

One study (Moed, 2005) analysed such relationships in three national research-assessment exercises in the Netherlands, in which peer-review committees of eight to ten international experts evaluated all research departments in a broad discipline (chemistry, biology and physics, respectively). In addition, it examined a peer review of all departments in a West European university. In all reviews, the research quality of departments was rated on a five-point scale (excellent; good; satisfactory; unsatisfactory; and poor).

It was found that the distributions of peer ratings among departments in the various exercises were statistically similar, whereas the departments' average citation impact differed substantially from one exercise to another. This finding suggests that a peer-rating system tends to generate a peer-quality distribution that depends on the rating system itself and that is to some extent independent of the overall level of quality of evaluated departments.

Analysis of the three national field reviews demonstrated that, if those responsible for the evaluation had not conducted a peer review at all, but had commissioned solely a bibliometric study, the outcomes of the latter, in terms of whether departments had a citation impact above or below world average, would correctly predict a peer rating, in terms of good or excellent versus less good (satisfactory or unsatisfactory), in about eight out of ten cases. In about one out of ten cases, the bibliometric study alone would rate a department higher than peers would have done. In another one out of ten cases, it would rate a department lower.

However, it was also found that, among the departments with a very high citation impact, the number of departments rated excellent was similar to that evaluated as good by the peers. This outcome suggests that the peer-review committees were able to identify good or valuable research meeting minimum quality standards, but they were not very successful in identifying genuinely excellent or top research.

Possibly, peers are more able to identify the bottom of the quality distribution (what is qualitatively less good) than the top (what is excellent or genuine top research), particularly when they are cognitively rather distant from most of the research activities they have to evaluate. This hypothesis is consistent with conclusions drawn in earlier studies on the evaluation of grant proposals and journal manuscripts in general or diffuse subfields, stating that referees tend to agree much more about what is unworthy of support than about what does have scientific value (Cicchetti, 1991).

Combining the two measures

General principles and considerations

As outlined previously, the use of citation analysis in research evaluation should be founded on the idea that citation impact, although a most useful and valuable aspect in its own right, does not fully coincide with notions such as intellectual influence, contribution to scientific progress or research quality. The outcomes of a citation analysis must be valued in terms of a qualitative, evaluative framework that takes into account the substantive contents of the works under evaluation (Moed, 2005). This can be done by peers only.

On the other hand, application of bibliometric indicators should be based on the notion that peer evaluators need to be provided with condensed, systematic, verified, objective information on the research performance of the groups to be evaluated, and that the grounds for their judgement, or the assumptions underlying it, should become more explicit, thus making the process more transparent.

It is argued that, in the policy domain, the use of citation analysis is more appropriate the more it is carried out openly according to transparent procedures with clear objectives; subjected entities are able to verify data and comment on results; potentialities and limitations, technical and validity issues are explicitly stated; its outcomes contribute to insight, pose problems or address particular questions that participants in the process seek to answer; and the process ensures the availability of expert knowledge on the entities involved and the fields in which they are active.

A good example of an evaluation system in which bibliometric indicators play a formal role is the system installed in the early 1990s by the Association of Universities in the Netherlands (VSNU), aiming to assess periodically and by discipline all research departments located at Netherlands academic institutions. The system was essentially based on peer review, but in several disciplines, particularly physics, chemistry and biology, systematic bibliometric analyses of all departments involved constituted one of its inputs.

Although this assessment procedure was recently transformed into a system of self-evaluation, in the disciplines mentioned above, citation analysis still plays an important role (van Leeuwen, 2004). Publication data used in the citation analysis were verified by the departments themselves, and the evaluation protocol gave their leaders the opportunity to comment on the bibliometric outcomes.

Research Assessment Exercises in the UK

The organisation of the upcoming RAE in the UK (2008) was the subject of a consultation exercise, particularly the possible role of citation analysis therein. On the one hand, The RAE 2008 generic statement (Higher Education Funding Councils, 2007) states that **"**no panel will use journal impact factors as a proxy measure for assessing quality", but does not mention other, more advanced types of citation analysis (page 32). Although the criteria and working methods of the Review Panel for Chemistry allow institutes to highlight "their outputs with particularly high citation rates" as "relevant details", they state that "citation rates and journal impact factors will not be used as measures of quality" (page 26).

On the other hand, the UK Government favoured the approach of the 2008 RAE being replaced by metrics such as research grants obtained and publication or citation metrics, and scrapping peer review. The consultation has now ended, and the decision has been made. The humanities and social sciences (plus mathematics and statistics) will have a light-touch panel review and will be informed by metrics; all other fields will be assessed by metrics only (pages 57–58).

There are a number of good reasons why it is worthwhile considering the use of a sophisticated type of citation analysis, based on actual citation rates, as a tool in RAE panels' evaluation of institutes in science disciplines. Apart from the more general considerations outlined in the previous

sections, at least two specific reasons should be highlighted:

- Each RAE submission is to be assessed "against absolute standards and will not be ranked against other submissions" (Higher Education Funding Councils, 2007: 19). Yet, as outlined above, empirical studies suggest that panels covering an entire discipline may not identify top research properly, and their standards tend to be relative rather than absolute. Assessing against absolute standards requires special, objective information on research performance from a global perspective, which citation analysis could provide.
- Even if citation analysis formally does not play a role in an evaluation process, individual reviewers may collect citation data themselves on an informal basis, calculating indicators that may not be sufficiently sophisticated, and using them in forming their judgements. Evaluated institutions may anticipate the use of such indicators, regardless of their formal exclusion as information to be considered in an assessment. Therefore, it is sensible to give indicators a formal status, and to make sure they are sophisticated and accurate.

Any evaluation system of human endeavour should systematically take into account 'strategic' behaviour of units under evaluation. What could the implications of the quantitative analysis presented under "Multi- and interdisiciplinary research" (above) of this type of behaviour within earlier RAEs be for the construction of valid citation-impact indicators and the way they should be used within a future RAE?

1. We should apply sophisticated indicators measuring actual citation impact. Publication counts and journal-impact factors should not play an important role, and probably no role at all. As outlined above, there is a broad consensus among bibliometric researchers that journal-impact factors should not be used as surrogates of actual citation impact. It cannot be claimed that actual citation-impact indicators cannot be affected in any way by strategic behaviour. Evaluators, evaluated researchers and bibliometricians should keep an open eye for any unintended effects of their use. But they are far more informative of a group's research performance and less easily manipulated than crude publication counts and indicators based on the number of published papers in journals with a high citation-impact factor.
2. Rather than relying on a single bibliometric indicator, a series of indicators should be produced. This poses the problem of how a series of indicators should be further analysed and interpreted, and what weight should be given to each individual indicator. This can be done only within a framework that integrates a theoretical model of which aspects of research performance the various indicators measure on the one hand, and, on the other hand, an evaluation model expressing the role and importance of each aspect in the formation of a qualitative judgement.

   Developing such an integrated framework is not a technical but a theoretical activity. Forming a quality judgement is not a mathematical problem, but the use of a system of weighted indicators can be a useful tool. From a technical point of view, a peer-review committee should be provided with a flexible application tool (datasets and analysis software) in which they can enter weights to individual indicators, calculate composite indicators, insert quality categorisations, and produce quantitative results according to their own specifications.
3. Peer-review committees should carry out their work independently and, within the boundary conditions set by the commissioning agencies, design their own procedures. However, the author of this paper would suggest considering the use of bibliometric indicators in the following way. An initial, tentative bibliometric ranking or classification of groups to be evaluated into quality categories could be generated, based on a selection and weighting of indicators specified by the committee, using the application described under point 2 above. Adoption of such a role in the evaluation process was also recommended by Norris and Oppenheim (2003), who qualify it as "the primary procedure for the initial ranking of university departments".

   Such a ranking could function as a guideline in the formation of quality judgements. In the process of valuing the outcomes of the bibliometric analysis, these judgements may diverge from the initial rankings or classifications. In these cases, a committee could explicitly state its considerations as to why they do so. In this way, quality judgements may be given a more solid foundation, and the evaluation process becomes more transparent.
4. The assessment should focus on research groups rather than individuals. Researchers may move from one group to another during the time period of analysis, but a research group constitutes the base unit of research in science. It is useful to distinguish between the past performance of the research carried out within a group and that of the researchers that are currently active in a group (van Raan, 1996; Moed, 2005). The first is based on the papers published by members of a group reporting the outcomes of research carried out (at least partly) in that group. The papers of newly appointed researchers published prior to their appointment date in a group are not taken into account.

   The second relates to the set of researchers who have the task of shaping the future of the group and takes into account all their articles published during the time interval of analysis, regardless of whether the research was carried out in the group or not. In this way, newly appointed researchers bring in their complete publication oeuvre generated in their previous institutional settings. Discrepancies between the outcomes of these two

types of analysis are informative and ask for a further explanation and interpretation.

5. Even if research is teamwork, a research group consists of individuals and their role and importance may vary. On the other hand, we should not isolate researchers too easily from their team members and disregard the institutional context in which they carried out their research. It is therefore a delicate task to capture fully the effects of changes in the composition of a group on a group's past and future performance. The methodology described under point 4 can at least provide some insight but, as argued earlier, it is difficult, if not impossible, to draw conclusions on the performance of individuals merely on the basis of bibliometric indicators.

   Other indicators more directly reflecting personal achievements should be applied as well, among which the number of invited or 'keynote' lectures are important, as are international, peer-reviewed scientific conferences. Generally, the review process should be organised in such a way that the committee has sufficient valid background knowledge on individual researchers. This may not be an easy task, particularly for small review committees that have to evaluate large numbers of research groups, especially all research groups in a major country active in an entire discipline.

Concluding remarks

Regarding the effects, either negative or positive, intended or unintended, of the use of citation analysis or any other methodology in research evaluation, it is crucial to distinguish two points of view. We may focus on its consequences for an individual entity, such as an individual scholar, a research group or institution, or on the effects it has on scholarly activity and progress in general. Each methodology has its strengths and limitations, and is associated with a certain risk of arriving at invalid outcomes in individual cases. As Cole *et al* (1981) argued in their pioneering study, a methodology, even if it provides invalid outcomes in individual cases, may be beneficial to the scholarly system as a whole. This is true both for bibliometric analysis and for peer review.

It is primarily the task of members from the scientific/scholarly community and the domain of research policy. and not of bibliometric investigators, to decide whether or not these risks of using a particular method of citation analysis are acceptable and whether its benefits prevail. This task may also comprise an assessment of whether the extra costs of an advanced, sophisticated bibliometric analysis, compared to those of a less sophisticated one, match its surplus value in a research evaluation process. This paper aims to provide information about the potentialities and limits of the various types of citation analysis that help scholars and policy-makers to carry out such a delicate task.

A real challenge in the organisation of future research-assessment exercises in the UK is the development of a model that aims to enhance the research performance of a national research system by making the distribution of funds across institutions more dependent on performance criteria, and at the same time facilitates and strengthens quality control, research management and policy within the nation's research institutions.

References

Braun, T, W Glänzel and A Schubert 1988. World flash on basic research — the newest version of the facts and figures on publication output and relative citation impact of 100 countries 1981–1985. *Scientometrics*, **13**, 181–188.

Butler, L 2003. Modifying publication practices in response to funding formulas. *Research Evaluation*, **12**(1), April, 39–46.

Butler, L 2004. What happens when funding is linked to publication counts? In *Handbook of Quantitative Science and Technology Research: the Use of Publication and Patent Statistics in Studies of S&T Systems*, eds. H F Moed, W Glänzel and U Schmoch, pp. 389–340. Dordrecht/Boston/London: Kluwer Academic Publishers.

Cicchetti, D V 1991. The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. *Behavioral and Brain Sciences*, **14**, 119–186.

Cole, S, J R Cole and G A Simon 1981. Chance and consensus in peer review. *Science*, **214**, 881–886.

Cronin, B 1984. *The Citation Process: the Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.

Garfield, E 1996. How can impact factors be improved? *British Medical Journal*, **313**, 411–413.

Higher Education Funding Councils 2007. *RAE 2008 Panel Criteria and Working Methods*. Available at <http://www.rae.ac.uk/pubs/2006/01/ >, last accessed 14 September 2007.

Langfeldt, L 2001. The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science*, **31**, 820–841.

Lipsett, A and A Fazackerley 2005. RAE shifts focus from prestige journals. *Times Higher Education Supplement*, 22 July 2005.

Martin, B R and J Irvine 1983. Assessing basic research: some partial indicators of scientific progress in radio astronomy. *Research Policy*, **12**, 61–90.

Moed, H F 2005. *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

Moed, H F 2008. UK Research Assessment Exercises: informed judgments on research quality or quantity? *Scientometrics*, **74**, 141–149.

Moed, H F, R E de Bruin and T N van Leeuwen 1995. New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. *Scientometrics*, **33**, 381–442.

Narin, F 1976. *Evaluative Bibliometrics: the Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Washington DC: National Science Foundation.

Norris, M and C Oppenheim 2003. Citation counts and the Research Assessment Exercise: V: archaeology and the 2001 RAE. *Journal of Documentation*, **59**, 709–730.

Salter, A J and B R Martin 2001. The economic benefits of publicly funded basic research: a critical review. *Research Policy*, **30**, 509–532.

Seglen, P O 1994. Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, **45**, 1–11.

van Leeuwen, T N 2004. Descriptive versus evaluative bibliometrics. In *Handbook of Quantitative Science and Technology Research: the Use of Publication and Patent Statistics in Studies of S&T Systems*, eds. H F Moed, W Glänzel and U Schmoch, pp. 373–388. Dordrecht/Boston/London: Kluwer Academic Publishers.

van Raan, A F J 1996. Advanced bibliometric methods as

quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, **36**, 397–420.

van Raan, A F J 2004. Measuring science. In *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, eds. H F Moed, W Glänzel and U Schmoch, pp. 19–50. Dordrecht/Boston/London: Kluwer Academic Publishers.

van Raan, A F J and T N van Leeuwen 2002. Assessment of the scientific basis of interdisciplinary, applied research: application of bibliometric methods in nutrition and food research. *Research Policy*, **31**, 611–632.

Vinkler, P 1986. Evaluations of some methods for the relative assessment of scientific publications. *Scientometrics*, **10**, 157–178.

White, H D 1990. Author co-citation analysis: overview and defense. In *Scholarly Communication and Bibliometrics*, ed. C L Borgman, pp. 84–106. Newbury Park: Sage.

Warner, J 2003. Citation analysis and research assessment in the United Kingdom. *Bulletin of the American Society for Information Science and Technology*, **30**(1), October/November, 26–27.

Watkins, D 2005. Authors per paper. *Sigmetrics Digest*, 19–20 May, special issue (#2005-77).

Woolgar S 1991. Beyond the citation debate: towards a sociology of measurement technologies and their use in science policy. *Science and Public Policy*, **18**(5), October, 319–326.

Zuckerman, H 1987. Citation analysis and the complex problem of intellectual influence. *Scientometrics*, **12**, 329–338.