# Commentary II
# The journal impact factor in the evaluation of research quality: villain, scapegoat or innocent bystander?

*Dr. Marcel Zwahlen is an epidemiologist and biostatistician and Dr. Christoph Junker is a clinical epidemiologist at the Department of Social and Preventive Medicine, University of Berne. Dr. Matthias Egger is the head of the latter Department and also a professor of clinical epidemiology at University of Bristol, UK*

Decker and colleagues (2004), in this issue of the Journal, express their reservations about the role the journal impact factor in the evaluation of the quality of scientific research. The journal impact factor is a measure of the frequency with which the "average" article in a journal has been cited (see Box 1), which was first used in the Institute of Scientific Information's (ISI) *Science Citation Index* (SCI) in 1961 (Garfield 1972; 1996). According to Decker et al. the impact factor has assumed a particularly prominent role in Germany, not only in the evaluation of institutions but also of individual researchers. They do not describe the situation in Germany in any detail, but go on to evaluate the "psychometric properties" of the journal impact factor "as an assessment procedure", and conclude that the impact factor is of limited use and that we have no other choice but to read scientific papers in order to assess their quality (Decker et al. 2004).

In this commentary we will discuss some of the potentials of citation analysis and argue that a wider discussion of the objectives, criteria and procedures of the evaluation of research is needed. We describe the 2001 Research Assessment Exercise (RAE 2001) in the United Kingdom as one exemplar of how this could be done. We believe that such evaluation exercises should also be seen as an opportunity, and not, as Decker et al. (2004) imply, merely as a threat. Indeed, the experience in the UK has been positive: universities and medical schools have developed strategies to streamline their research portfolios and concentrate on existing areas of strength, with an increase in the quality of their output (Tomlinson 2000).

**Box 1** Useful definitions in bibliometric research

**Scientometrics**
The quantitative study of the disciplines of science based on published literature and communications. This could include identifying emerging areas of scientific research, examining the development of research over time, or geographic and organizational distributions of research.

**Bibliometrics**
Study of the quantitative data of the publication patterns of individual articles, journals, and books in order to analyse trends and make comparisons within a body of literature.

**Citation index**
Compilation of all referenced or cited source items published in a given time span. Useful for tracking the historical development of an idea or given topic within the literature published in a wide selection of journals.

**Journal impact factor**
The number of citations in one year (for example 2002) to articles published in a specific journal in the two preceding years (2000 and 2001) divided by the total number of citable articles (original articles and reviews, but not other items) published in this two-year period.

**Cited half-life**
Measure of the long-term impact of source items in a single journal publication. It is the number of years, going back from the current year, that account for 50% of the total citations received by the cited journal in the current year.

**Journal immediacy index**
The average number of times that an article published in a specific year within a specific journal is cited over the course of that same year.

Adapted from the pages of the Institute for Scientific Information (ISI)

Soz.- Präventivmed. 49 (2004) 19–22
0303-8408/04/010019–4
DOI 10.1007/s00038-003-0033-2

### Citation analysis

Publication is a crucial stage in the research process, but not the final step. The critical appraisal of the published findings, for example in letters to the editors, and their use and citation by other researchers are also important. Research is a cumulative process; every published piece is part of a growing body of evidence. Findings that either remain unpublished, or are published but not noticed, make little contribution, and may introduce bias in the published and cited evidence base (Egger & Smith 1998; Egger et al. 2001a). Evidently, there is no way to quantify accurately the usefulness of a given piece of research, but the frequency with which an article is cited is a useful proxy measure of its impact on other researchers. Although the motivations for citing an article range from decoration to showing up-to-datedness and knowledge (Brooks 1985), citation analysis is a useful tool for identifying the more important building blocks: many scientific articles are never cited, and those who are cited frequently tend to be published in a limited number of journals (Garfield 1996; Seglen 1997). These journals have a high impact factor, but this does not mean that each article they publish is in fact widely cited: the distribution of citation rates is skewed and their impact factor is driven by a relatively small number of highly cited articles (Seglen 1992). In other words, the impact factor of the journal that publishes a given article is a poor predictor of the number of citations the article is likely to receive. Furthermore, the impact factor of a journal can be manipulated, for example by publishing a controversial article (like the article by Decker et al. 2004), which counts as a citable item, and commissioning commentaries (like this one) which gain citations for the journal but do not count as citable items (Adams 2002). It has long been recognised that the journal impact factor should not be used to evaluate the research output of faculty members or junior researchers (Garfield 1984), although citation rates of individual articles, relative to the typical citation impact in that field, may be useful in this context (Schubert & Braun 1996). Citation analysis is of course an important research tool to address broader questions in bibliometric and scientometric research (see Box 1).

### Psychometric properties of the journal impact factor?

Measurement validity is defined as "the degree to which a measurement measures what it purports to measure" (Last 2001). By definition, the journal impact factor was never intended as a measure of, or test for, the quality of research output. The arguments against its use have been eloquently summarised in a widely cited article (121 citations as of September 2003) published by Seglen in 1997. Decker et al. (2004) revisit these arguments in the framework of an evaluation of the "psychometric properties" of the journal impact factor. Their justification for this exercise appears to be twofold. Firstly, that this measure is increasingly used inappropriately in this context, and secondly that it is affected by numerator-denominator and misclassification biases. This seems far fetched and hardly convincing: do we now need to assess the psychometric properties of the many other measures that are inappropriately used and suffer from these biases, for example, in epidemiology, prevalence and incidence? (Flanders & O'Brien 1989; Tapia Granados 1997). Decker et al. (2004) fail to realise that the impact factor is the "poor man's citation analysis" (Adams 2002), which by definition relates to the past two years only, and that long-term indicators of citation impact such as the cited half-life are available (see Box 1). Information scientists familiar with the ISI database can of course calculate indices for any time period, with our without lag times, and their assertion that in this journal *(Sozial und Präventivmedizin)* a larger proportion of citations refer to older articles could easily be examined formally. Finally, Decker et al. (2004) argue that there is no association between the quality of a study and the impact factor of the journal that published it. This important question would merit a systematic review of the literature (Egger et al. 2001b): one study recently showed that high citation rates of articles, journal impact factors, and low manuscript acceptance rates of journals predict higher methodological quality (Lee et al. 2002). Similarly, a study from our group found that controlled clinical trials published in languages other than English were of lower quality than trials published in English-language journals, which have higher impact factors than other journals (Juni et al. 2002).

We share the Decker et al.'s concern about the epidemic of impactitis, which appears to rage in Germany and elsewhere. One country which appears to be to some extent immune against this disease is the United Kingdom.

### The Research Assessment Exercise in the UK

Since 1986 a formal evaluation of publicly funded universities and higher education colleges has been conducted in the UK in order to provide ratings of the quality of their research (Research Assessment Exercise, RAE). A separate system has been put in place to assess the quality of learning and teaching in higher education. The purpose of the RAE is not only to inform the allocation of funds but also to promote high quality research. Indeed, there is evidence that the RAE has contributed to a substantial increase in the effectiveness and productivity of the UK

Soz.- Präventivmed. 49 (2004) 19–22

research base (Tomlinson 2000; United Kingdom Research Assessment Exercise).

The RAE operates through a process of peer review by experts of high standing. Each publicly funded university and higher education college is invited to submit information of their research activity for assessment. The research submitted is assessed against a benchmark of international excellence in the subject concerned, with ratings on a scale from five stars ("international excellence in more than half of research activity submitted and national excellence in the remainder) to one ("national excellence in virtually none of the research submitted").

In 2001, there were 68 units of assessment, with each unit covering a broad subject area. A large number of expert panels and sub-panels were responsible for assessing the submissions within these units. Each panel prepared a statement of the criteria and working methods that it would use during the assessment process. The criteria used by the expert panel on community based subjects, which includes public health and health services research, primary care, psychiatry and related subjects are summarised in Box 2. Of note, research output was restricted to four items per researcher and no quantitative bibliometric indices were used. All forms of research output (for example, journal articles and book chapters) were treated equally. Panels were concerned with quality, not quantity, and information on the total number of publications produced was not requested. Finally, although the greatest weight was afforded to the quality of research output, other measures such as external grant income and evidence of peer esteem were also considered. Detailed information on the RAE 2001 is available elsewhere (United Kingdom Research Assessment Exercise).

**Box 2** Criteria used by panel on community based subjects in the UK Research Assessment Exercise (RAE 2001)

**Quality of research output**
For each researcher up to four items of research output could be submitted (for example, papers in scientific journals, monographs and books). The evaluation of research papers did not use quantitative bibliometric indices, but was based on the professional judgement of the panel. The panel assessed the international and national impact of publications in the relevant field of interest. Order of authorship was not important, and the panel considered group authorship appropriate for the publication of collaborative research.

**Extent of research activities**
Indicators included the number of research assistants employed, the number of higher degrees awarded, and the number of studentships.

**External research income**
The panel did not adopt a mechanistic approach when assessing the research income per head of research active staff. It expected to see a dynamic research culture reflected in substantial research income, while accepting that different areas of research may require different levels of funding.

**Research strategy, structure and environment**
Institutions were asked to state the main objectives and activities in research over the next five years, to define research groups and to detail mechanisms and practices for promoting research and sustaining and developing an active and vital research culture. The procedures for developing research strategies had to be described.

**Evidence of peer esteem**
Indicators included, for example, Fellowships of the Royal Society, of the Academy of Medical Sciences, named lectures and invitations to give key-note addresses, senior awards from Research Councils, the Medical Charities or similar bodies in other countries, membership of national or international review boards and funding bodies, editorship of international journals and election to international bodies.

Community based subjects include epidemiology, public health research, health services research, primary care, psychiatry and related subjects, for example health psychology, medical sociology and Biostatistics
Adapted from the pages of RAE 2001 (United Kingdom Research Assessment Exercise)

## Conclusions

We think it is interesting that in the UK a system based on peer review, which may be open to favouritism and nepotism, has gained wide acceptance, whereas in Germany the assessment of research appears to be mechanistically based on the journal impact factor, which is known to be of limited value in this context. It would be worthwhile to perform a systematic, comparative review of approaches taken in different countries, which could inform discussions on how best to assess the quality of publicly funded research. Clearly, in any country, an open debate on procedures and criteria for the evaluation of research, and transparency during and after completion of the process, are required to prevent the type of unhappiness expressed by Decker and colleagues (2004). However, those who are opposed to one approach to evaluation will have to come up with a credible alternative for the accountable allocation of the large sums that society invests in the publicly funded research enterprise.

**Marcel Zwahlen, Christoph Junker, and Matthias Egger**

Soz.- Präventivmed. 49 (2004) 19–22

## References

*Adams D* (2002). The counting house. Nature *415*: 726–9.

*Brooks TA* (1985). Private acts and public objects: an investigation of citer motivations. J Am Soc Inform Sci *36*: 223–9.

*Decker O, Beutel ME, Brähler E* (2004). Deep impact – evaluation in the sciences. Soz Praventiv Med *49*: 10–4.

*Egger M, Dickersin K, Smith DG* (2001a). Problems and limitations in conducting systematic reviews. In: Egger M, Smith DG, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. London: BMJ Books: 43–68.

*Egger M, Smith DG* (1998). Meta-analysis: bias in location and selection of studies. BMJ *316*: 61–6.

*Egger M, Smith DG, O'Rourke K* (2001b). Rationale, potentials and promise of systematic reviews. In: Egger M, Smith DG, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. London: BMJ Books: 23–42.

*Flanders WD, O'Brien TR* (1989). Inappropriate comparisons of incidence and prevalence in epidemiologic research. Am J Public Health *79*: 1301–3.

*Garfield E* (1972). Citation analysis as a tool in journal evaluation. Science *178*: 471–9.

*Garfield E* (1984). How to use citation analysis for faculty evaluations, and when is it relevant? Essays of an information scientist. Philadelphia: ISI Press: 354–72.

*Garfield E* (1996). How can impact factors be improved? BMJ *313*: 411–3.

Institute for Scientific Information. http://sunweb.isinet.com/isi/search/glossary/

*Juni P, Holenstein F, Sterne J, Bartlett C, Egger M* (2002). Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol *31*: 115–23.

*Last JM* (2001). A dictionary of epidemiology. New York: Oxford University Press.

*Lee KP, Schotland M, Bacchetti P, Bero LA* (2002). Association of journal quality indicators with methodological quality of clinical research articles. JAMA *287*: 2805–8.

*Schubert A, Braun T* (1996). Cross-field normalization of scientometric indicators. Scientometrics *36*: 311–24.

*Seglen P* (1992). The skewness of science. J Am Soc Inform Sci *43*: 628–38.

*Seglen PO* (1997). Why the impact of journals should not be used for evaluating research. BMJ *314*: 497–502.

*Tapia Granados JA* (1997). On the terminology and dimensions of incidence. J Clin Epidemiol *50*: 891–7.

*Tomlinson S* (2000). The research assessment exercise and medical research. BMJ *320*: 636–9.

United Kingdom Research Assessment Exercise. http://www.hero.ac.uk/rae

**Address for correspondence**

**Prof. Matthias Egger**
**Department of Social and Preventive Medicine**
**University of Berne**
**Finkenhubelweg 11**
**CH-3012 Berne**
**E-mail: egger@ispm.unibe.ch**