Vol. 8: 93–102, 2008
doi: 10.3354/esep00084

ETHICS IN SCIENCE AND ENVIRONMENTAL POLITICS
Ethics Sci Environ Polit

Printed June, 2008
Published online February 12, 2008

Contribution to the Theme Section 'The use and misuse of bibliometric indices in evaluating scholarly performance'

OPEN ACCESS

Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results

Lutz Bornmann[1,*], Rüdiger Mutz[1], Christoph Neuhaus[1], Hans-Dieter Daniel[1,2]

[1]ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Zähringerstr. 24, 8092 Zurich, Switzerland
[2]University of Zurich, Evaluation Office, Mühlegasse 21, 8001 Zurich, Switzerland

ABSTRACT: With the ready accessibility of bibliometric data and the availability of ready-to-use tools for generating bibliometric indicators for evaluation purposes, there is the danger of inappropriate use. Here we present standards of good practice for analyzing bibliometric data and presenting and interpreting the results. Comparisons drawn between research groups as to research performance are valid only if (1) the scientific impact of the research groups or their publications are looked at by using box plots, Lorenz curves, and Gini coefficients to represent distribution characteristics of data (in other words, going beyond the usual arithmetic mean value), (2) different reference standards are used to assess the impact of research groups, and the appropriateness of the reference standards undergoes critical examination, and (3) statistical analyses comparing citation counts take into consideration that citations are a function of many influencing factors besides scientific quality.

KEY WORDS: Citation counts · Research evaluation · Citation distribution · Reference standards · Lorenz curve · Gini coefficient · Box plots

INTRODUCTION

The publication of a research paper serves to disseminate the results of the research and at the same time 'invites' other scientists to use the findings in their own research (McClellan 2003). When other scientists use these findings, they indicate this in their own publications by means of a formal citation. As the citations are meant to show that a publication has made use of the contents of other publications (research results, others' ideas, and so on), citation counts (the number of citations) are used in research evaluation as an indicator of the impact of the research: 'The *impact* of a piece of research is the degree to which it has been useful to other researchers' (Shadbolt et al. 2006, p. 202; see also Bornmann & Daniel 2007a). Following van Raan (2004), 'citation-based bibliometric analysis provides indicators of international impact, influence. This can be regarded as, at least, one crucial aspect of scientific quality' (p. 27; see also Martin & Irvine 1983). According to the REPP (2005), there is an emerging trend to regard impact, the measurable part of quality, as a proxy measure for quality in total.

In research evaluation, citation counts are being used for evaluation and comparison of the research performance of individual researchers, departments, and research institutions (Garfield et al. 1978, Adam 2002) as well as the scientific impact of nations (May 1997, King 2004). Citation counts are attractive raw data for the evaluation of research output. Because they are 'unobtrusive measures that do not require the cooperation of a respondent and do not themselves contaminate the response (i.e. they are non-reactive)' (Smith 1981, p. 84), citation rates are seen as an objective quantitative indicator for scientific success and are held to be a valuable complement to qualitative methods for research evaluation, such as peer review (Garfield & Welljamsdorof 1992, Daniel 2005).

Most commonly, the main resource for citation analysis are the citation indexes produced by Thomson

*Email: bornmann@gess.ethz.ch

Scientific (Philadelphia, Pennsylvania, USA). For many years, the citation indexes had a unique position among bibliographic databases because of their multidisciplinary nature and indexing of cited references. For this reason, the Thomson Scientific citation indexes provide an obvious starting point in assessing research performance. However, their coverage is restricted to a set of mainly internationally oriented journals, with the exception of some highly-cited book series and conference proceedings. While these journals tend to be the highest-impact peer-reviewed journals (according to Shadbolt et al.'s 2006 definition, cited above), they represent only a fraction of scientific work that is documented. Consequently, citation analysis based on the Thomson Scientific citation indexes is less applicable in fields such as computer science, engineering, and mathematics, where the journal literature plays an inferior role in the scholarly communication system (Moed 2005).

However, Thomson Scientific is no longer the only database offering citation indexing. Recently, discipline-oriented databases such as Chemical Abstracts (Chemical Abstracts Services), MathSciNet (American Mathematical Society), and PsycINFO (American Psychological Association) have enhanced their records with cited references. New abstract and citation databases such as Scopus (www.scopus.com; Elsevier), Google Scholar (http://scholar.google.com/), and CiteSeer (http://citeseer.ist.psu.edu/) have emerged (for an overview, see Neuhaus & Daniel 2008). The availability of citation data in additional bibliographic databases opens up the possibility of extending the data sources for performing citation analysis, and particularly for including other document types of written scholarly communication, such as books, chapters in edited books, and conference proceedings. The inclusion of other document types may contribute to the validity of bibliometric analysis when evaluating fields in which the internationally oriented scientific journal is not the main medium for communicating research findings (Moed 2005).

Whereas in the past only specialists were able to work with the bibliometric data from different databases, Thomson Scientific and other providers have recently begun to offer ready-to-use tools with which users can quite easily generate bibliometric indicators for evaluation purposes (Weingart 2005, Steele et al. 2006). 'A recent trend which has raised some concern is the increased application of 'amateur bibliometrics' (REPP 2005, p. 6). 'The 'danger' is … organizations greedy to buy ready-to-go indicators without any competence to understand what is measured. It is my experience that authoritative persons in organizations still cultivate the unreasonable "please press the button and I have the numbers that I want to have" mentality' (van Raan 2005a, p. 133–134). It is for this reason that we present here standards of good practice for the analysis of bibliometric data and for presentation and interpretation of the results that should be adhered to when assessing and comparing the research performance of research groups. Only when the following standards are met can valid statements be made as to the output of research groups:

(1) Bibliometric analyses for evaluation and comparison of research performance usually use the arithmetic mean value as a measure of central tendency (Kostoff 2002, van Raan 2004). That means that the more frequently the publications of a research group are cited on the average over all publications, the higher the group's performance is rated. But there are dangers in the use of this measure: in the face of non-normal distributed citation data, the arithmetic mean value can give a distorted picture of the kind of distribution. Furthermore, there is huge empirical evidence that a small number of researchers account for a large proportion of the total research output. 'Thus, average figures do not say much about the small core of highly productive researchers on whom so much of the total research effort rests. Thus, rather than simply computing departmental mean performance scores, one might gain more by identifying the prolific researchers' (Daniel & Fisch 1990, p. 358). In 'Explorative data analysis in bibliometrics' below, we present statistical tools of explorative data analysis (Tukey 1977) that should be used to describe the distribution of data in bibliometric analyses.

(2) How frequently the publications of a research group have been cited says little on its own (see Schubert & Braun 1996, Kostoff 2002). The publication activity and citation habits of different scientific disciplines are too varied to allow evaluation of research success on the basis of absolute numbers. Rather, the assessment of research performance is relative to a frame of reference in which citation counts are interpreted. The comparison with reference standards makes it possible to assign meaning to citation counts and to place them in an adequate context. According to Schubert & Braun (1996), there are basically 3 approaches to setting reference standards for the comparative assessment of research performance. Reference standards may be established on the basis of (a) fields of research, (b) journals, or (c) related records. The evaluation of the relative research performance of a research group (as compared to other research groups) depends decisively on the selection of an appropriate reference standard (see 'Reference standards' below).

(3) Eugene Garfield, the founder of the Institute of Scientific Information (ISI, now Thomson Scientific) already pointed out in the early 1970s that citation

counts are a function of many variables besides scientific quality (Garfield 1972). Since then, a number of variables that generally influence citation counts have emerged in bibliometric studies that must be considered in the statistical analysis of bibliometric data. Lawani (1986) and other researchers established, for example, that there is a positive relation between the number of co-authors of a publication and its citation counts: a higher number of co-authors is usually associated with a higher number of citations. Based on the findings of these bibliometric studies, the number of co-authors and other general influencing factors that are mentioned in 'Citations depend on many factors' below should — whenever possible — be controlled in the statistical analysis of bibliometric data.

EXPLORATIVE DATA ANALYSIS

Prior to conducting every statistical analysis of data, the question of the level of measurement (e.g. categorical, ordinal) must be clarified. The data used in bibliometric studies are counts (van Belle et al. 2004, Ross 2007). These counts result from the collapsing of repeated binary events (cited or not) on subjects (articles) measured over some time period to a single count (for example, number of citations of an article). Counts are therefore integer values that are always greater or equal to zero. The Poisson distribution, unlike the normal distribution (symmetric Gaussian probability distribution, or bell curve), is often used to model information on counts, especially in situations where there is no upper limit on how large an observed count can be. These increments of variable y are Poisson distributed with expectation $E(y) = \lambda$ and variance λ. A mean of $\lambda = 2$ means that per time period, 2 citations occur on average. Because with a Poisson distribution the mean equals the variance, a quick test reveals whether the data are Poisson distributed (mean/variance $\approx$ 1.0). Because bibliometric analyses are used to evaluate individual researchers or research groups at a certain point in time, it is the cumulative or mean citation counts of their publications for a certain particular time interval. These aggregated data are no longer Poisson distributed (compounding distribution). For this reason, in bibliometric analysis the negative binomial distribution is recommended (Schubert & Glänzel 1983). In view of the non-normal distributed citation data, examining only the arithmetic mean value under the assumption of a normal distribution can lead to a distorted picture of the citation frequencies when comparing different research groups. Taking the example of bibliometric analysis of the publication activity of research groups within the Department of Biochemistry at the University of Zurich (see Table 1), we can illustrate the limits of a mean value orientation and also show possible solutions from the point of view of Exploratory Data Analysis (Tukey 1977). To support the bibliometric analysis, 2 kinds of statistical tools for description of distributions of citations can be helpful: box plots and Lorenz curves with Gini coefficients.

Table 1. Descriptive statistics for the bibliometric analysis (citation counts) of 5 research groups in a biochemistry department. N_{res}: number of researchers; N_{art}: number of articles; M: mean citation counts of articles; VAR: variance; Mdn: Median; M_{est}: M-estimator

Research group	N_{res}	N_{art}	M	VAR	Mdn	M_{est}
1	3	34	23.1	329	21.0	21.4
2	8	30	15.4	560	8.5	11.6
3	9	30	12.9	167	11.0	11.9
4	9	106	34.2	1360	25.0	26.7
5	11	62	20.3	478	12.0	16.6

Box plots

A box plot or box and whisker diagram is a convenient way of visually summarizing a distribution; it consists of the smallest observation, lower quartile (Q1, 25% of all observations), median (50% of all observations), upper quartile (Q3, 75% of all observations), largest observation, and in addition 2 whiskers that represent 1½ times the length of the box (van Belle et al. 2004). Any observed value outside the ends of the whiskers is considered unusual, or an outlier. In bibliometric analyses outliers supply important information on very highly or very lowly cited papers, but they can not depict the entire output of a research group. Fig. 1 shows example box plots for the citation counts of all papers published by 5 research groups within the department of biochemistry in a certain time interval (1997 to 2001). The distribution of observations is shown next to each box plot. The normal distribution curve is also added to each of the distributions of observations. If the citations are normally distributed within the research groups, then this curve must be symmetric and bell-shaped. This is not the case for any of the 5 research groups. This means that the requirement for many parametrical statistical tests (such as analysis of variance for the comparison of 5 groups) is not met.

Whereas for Research Groups 1 and 3 the mean citation counts of their articles is largely in agreement with the median, for research group 4 the mean value (M = 34.2) and median (Mdn = 25.0) are very different (see Table 1). This deviation is explained by the strong sensitivity of mean values to the extreme values in the dis-

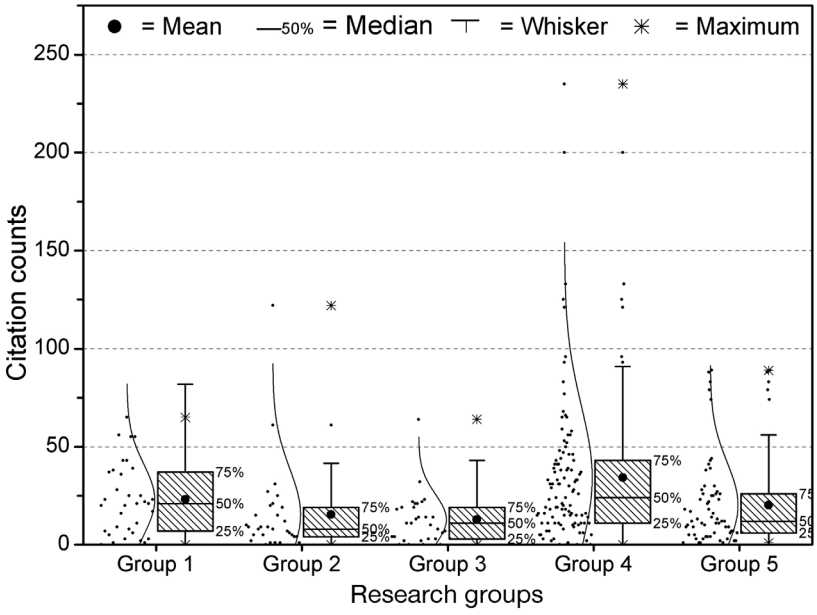

Fig. 1. Box plots for citation counts for 5 research groups in a biochemistry department. Small dots and curves show distribution of data

tribution (outliers). For instance, some articles published by research group 4 were cited more than 100 times (maximum = 235 citations). The whiskers make it possible to recognize the number of these outliers immediately. A robust alternative to the arithmetic mean that is less sensitive to outliers is Huber's M-estimator (Huber 1981). Huber's M-estimator does not eliminate the influence of outliers completely but minimizes it through a weighting, as Table 1 with the descriptive statistics shows: the mean value for Research Group 4 is now reduced from M = 34.2 (mean) to M_{est} = 26.7 (M-estimator) and approaches the median.

Research groups are frequently compared using the mean citation counts of their articles. If mean values are examined by appearances alone, Table 1 seems to show that there are distinct differences in the mean values between Research Group 4 and the other groups but also between Research Group 1 and Research Groups 2 and 3. In fact, however, these are only meaningful if the difference in means of the different research groups is statistically significant. A difference in the means is significant, if the variance of mean values between groups (signal) is distinctly higher than the variance within groups (noise). For testing, it is routine to use the analysis of variance (ANOVA). However, ANOVA requires normal distributed data and homogeneity of variance (same variance within groups), which is seldom the case in bibliometric analyses. As in Table 1 variance and mean value do not agree for all of the research groups, a negative binomial distribution must be assumed instead of a Poisson distribution. The assumption of a negative binominal distribution makes it necessary to calculate a generalized linear model (PROC GENMOD, SAS) to test for differences between the groups. The results of the analysis show that overall there are statistically significant differences between research groups in the mean values ($\chi^2(4) = 30.05$, $p < 0.05$). *A posteriori* contrasts with Bonferroni correction were tested in order to determine what pair-wise differences in mean values are statistically significant. These pair-wise tests show that of all of the research groups, only Research Group 4 differs statistically significantly from Research Groups 2, 3, and 5. But no significant differences could be found between Research Group 1 and Research Groups 2 and 3, which a mere look at the mean values had suggested. That is why assessing the differences in mean values only by looking at the values (signal) and without taking into account the variability within each research group (noise) can lead to misjudgments of the actual citation-impact differences between research groups.

Lorenz curve and Gini coefficient

Average figures do not say much about the small core of highly cited articles, or highly productive researchers. When the research group is the unit of analysis, some measures of concentration should be computed in order to distinguish between research groups with 'collective strength' and groups with 'individual strength' (Daniel & Fisch 1990, Burrell 2006). In the Lorenz curve, the cumulative proportion of articles (*x*-axis) is plotted against the cumulative proportion of their total citations on the *y*-axis (Damgaard & Weiner 2000). Fig. 2 shows Lorenz curves for the cumulative percentage of articles and that of authors against the cumulative percentage of total citations for 3 research groups.

Lorenz curves capture the degree of inequality or concentration. If each article had equal value in its shares of the total citations, it would plot as a straight diagonal line, called the perfect equality line (see the dotted line in Fig. 2). If the observed curve deviates from the perfect equality line, the articles do not contribute equally strongly to the total number of citations (Fig. 2a,b). For example, for Research Group 1, 20% of

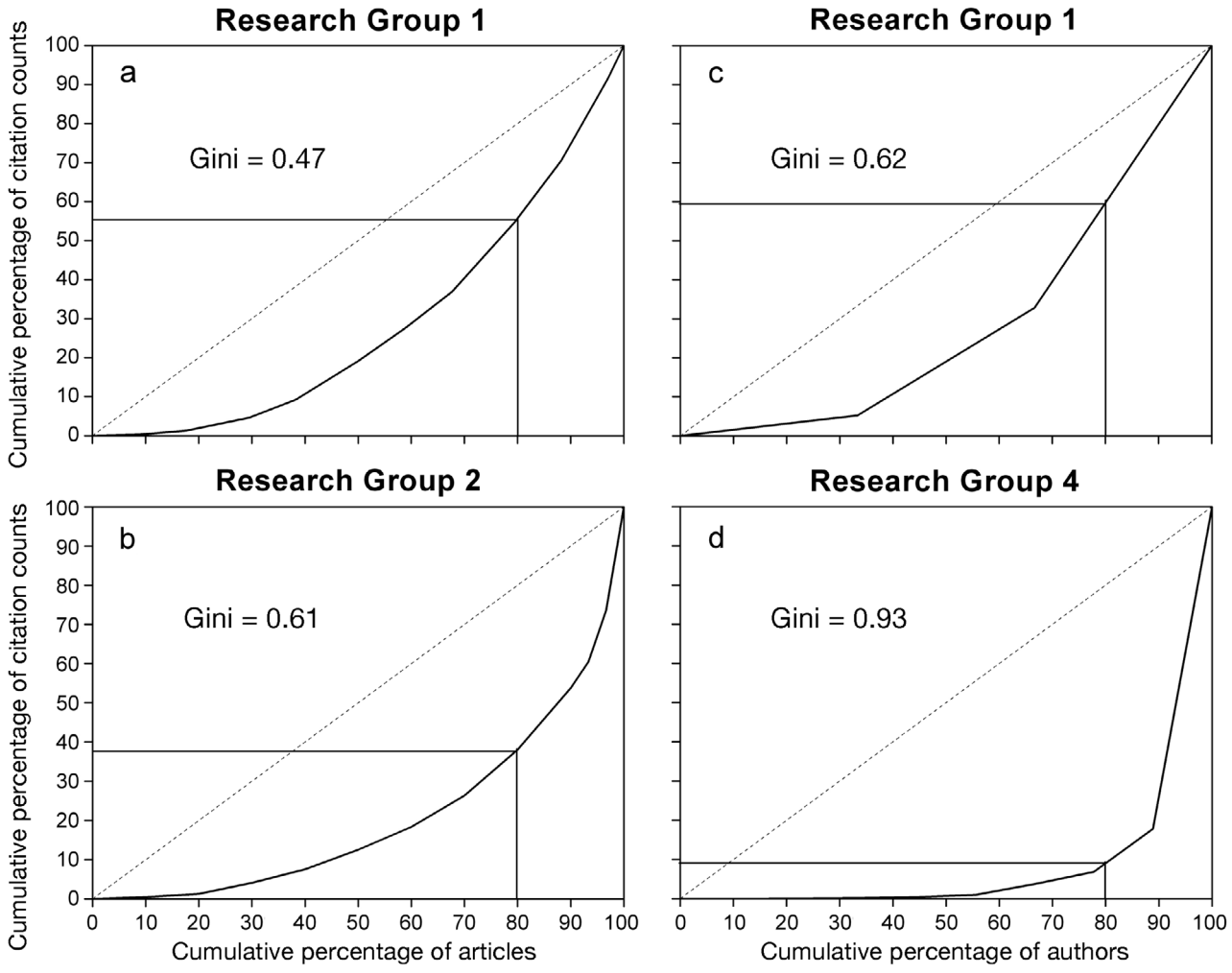

Fig. 2. Lorenz curves with Gini coefficients for (a,b) articles and citation counts and (c,d) authors and citation counts. Example: for Research Group 4, 20% of the authors receive approximately 90% of the total citations

the articles receive about 45% of the total citations, and for Research Group 2 the inequality is even greater: 20% of the articles receive >60% of the total citations. The degree of the concentration of the citations received by a few articles can also be expressed with the Gini coefficient (Atkinson 1970), a mathematical summary (area ratio) of inequality based on the Lorenz curve (ranging from 0.0 = perfect equality to 1.0 = complete inequality), which at 0.61 is plainly higher for Research Group 2 than for Research Group 1 at 0.47.

If the total citation counts of whole research groups are compared, the implicit assumption is that each individual researcher in a research group has contributed an equal share in total citations. But Lorenz curves and Gini coefficients calculated for the authors of the articles (Fig. 2c,d) show clearly that this must not be the case: whereas in Research Group 1 20% of the authors receive 40% of the total citations, in Research Group 4, 20% of the authors receive a full 90% of the total citations. Research impact as measured by total citations is concentrated in very few members of the Research Group 4 in comparison to Research Group 1.

REFERENCE STANDARDS

In research evaluation, a widely used approach is to compare the average number of citations to the oeuvre of a research group with that of the average number of citations to the field(s) in which the research group has published its papers (van Raan 2004, van Leeuwen 2007). The definition of research fields is based on a classification of journals into subject categories developed by Thomson Scientific. Each journal is classified as a whole to one or several subject categories. In general, this journal classification scheme proves to be of great value for research evaluation. But its limitations

become obvious in the case of (1) multidisciplinary journals such as *Nature* or *Science* (see, for example, Glänzel et al. 1999) and (2) highly specialized fields of research (e.g. Schubert & Braun 1996; Kostoff 2002). These limitations are illustrated with some examples below.

Multidisciplinary journals

Because papers published in multidisciplinary journals are not assigned to a specific field but classified as multidisciplinary, reference standards based on journal classification schemes yield an incomplete picture of the research output of a given field. As a result, a considerable portion of the relevant literature is not captured (Rinia et al. 1993). In some cases, papers with the highest impact in a field are published in multidisciplinary journals and not in disciplinary ones. For example, take a research group in experimental immunology at the University of Zurich. The publication list of the research group contains 195 papers in the period from 1996 to 2001. On average, the group's papers published in journals classified as 'immunology' in the Essential Science Indicators — an analytical tool offering citation data for scientists, institutions, countries, journals and fields, published by Thomson Scientific — were cited 20.3 times (n = 94 papers), in journals classified as 'clinical medicine' 36.6 times (n = 43), and in journals classified as 'multidisciplinary' as many as 41.9 times (n = 29). Expectedly, in terms of mean impact, the research group rates highest with papers published not in immunology journals but in multidisciplinary journals.

The example of the research group in experimental immunology shows that reference standards based on journal classification schemes (such as the journal set 'immunology') are based on only a fraction of papers effectively published in a given field. Consequently, the 'true' value of the reference standard, based on all papers published in the field of immunology, can not be established by the journal set approach.

Highly specialized fields of research

The limitations of journal classification schemes can be illustrated taking another example, this time a highly specialized research group in the Department of Neurology at the University of Zurich. Investigating the vestibular and ocular motor system in humans, the research group is active in a field that has a small scientific community. In the period from 1999 to 2003 the research group published 48 papers that accumulated a total of 164 citations over a period of 3 yr after publication. We compared the average number of citations per publication, CPP = 164/48 = 3.42, with a reference standard based on (1) the journal classification scheme (journal sets) of the Journal Citation Reports produced by Thomson Scientific, and (2) the Medical Subject Headings (MeSH) of the bibliographic database MEDLINE (US National Library of Medicine). In contrast to the journal classification scheme, MeSH index terms are assigned on a paper-by-paper basis. Furthermore, MeSH terminology is arranged in a hierarchical structure and distinguishes fields of research at a much lower level. For a comparison of research groups, the subject classification approach based, for example, on MeSH index terms is therefore more appropriate than the journal classification approach. Both reference standards were calculated for papers published in 2003 and their citations over a 3 yr period, including self-citations.

For the neurology research group investigating the vestibular and ocular motor system in humans, the reference standard based on the journal classification scheme of the Journal Citation Reports varies between 2.31 for the journal set 'otorhinolaryngology' and 18.90 for 'multidisciplinary sciences'. The weighted reference standard is FCSm = 8.31, the weights being determined by the number of papers published by the research group in the respective field. As is usual in bibliometric analysis, we calculated the reference standards as the arithmetic mean of citations. The impact of the research group, CPP/FCSm = 3.42/8.31 = 0.41, lies far below the international standard of the field (see van Raan 2004).

For the reference standard based on the subject classification scheme, we retrieved all research articles indexed with the MeSH index terms 'eye movements' or 'reflex, vestibulo-ocular' from the MEDLINE database (n = 527) and searched their citations in the SCISEARCH database (Science Citation Index Expanded, produced by Thomson Scientific) at the online database host STN International (www.stn-international.de). The value for the reference standard amounts to $FCSm_{MeSH}$ = 3.71; therefore, the impact of the research group with $CPP/FCSm_{MeSH}$ = 3.42/3.71 = 0.92 is about the international standard of the field.

In order to learn something about the exactness of this estimate, we estimated the confidence intervals (CI) for both of the bibliometric indicators. The result was a 95% CI of 2.58 to 4.25 for the CPP indicator and of 3.28 to 4.23 for the $FCSm_{MeSH}$ indicator. The area in which contains 95% of all possible location parameters is therefore considerably larger with the CPP indicator than with the $FCSm_{MeSH}$ indicator. The different exactness of the 2 estimates depends on the size of the sample; increasing the size of the sample leads generally to a smaller CI. However, the $CPP/FCSm_{MeSH}$ indicator

does not take into account the different exactness of the 2 estimates. Since citation counts are not normally distributed but follow a negative binomial distribution (Schubert & Glänzel 1983), resampling methods (such as bootstrap) can be used for deriving estimates of CIs.

The average citation rate is in addition strongly influenced by individual, highly cited papers. In a statistical sense these highly cited papers can be viewed as outliers. As mentioned in 'Explorative data analysis', we calculated the M-estimator, resulting in a location parameter of 3.15 for CPP and 2.65 for $FCSM_{MeSH}$. Accordingly, the impact of the research group is $CPP/FCSM_{MeSH}$ = 1.19, and thus considerably higher than the impact as calculated by the non-robust arithmetic mean, which amounts to 0.92.

The example shows that the evaluation of the research performance of a research group depends decisively on the selection of the reference standard. Especially the comparison with papers published in journals belonging to the journal set 'multidisciplinary sciences' is dubious in many cases, since multidisciplinary journals publish papers in a wide range of fields that have very different expected or average citation rates. The journal set 'multidisciplinary sciences' has a reference value of 18.90, which is very high in comparison with other journals. If research groups working in small and highly specialized fields of research are measured according to that journal set, this leads inevitably to invalid conclusions. In highly specialized fields there is only a small number of researchers that can potentially cite research findings, even if they are published in high-impact multidisciplinary journals. It is therefore reasonable to scrutinize the appropriateness of a reference standard as the case arises, especially when bibliometric analysis is used to inform decisions on the allocation of funds to research groups, for instance. Certainly, the level of aggregation is an important criterion for the selection of a reference standard. For citation analysis at a macro level (e.g. nations or universities), reference standards based on journal classification schemes may be a good choice, whereas for citation analysis at the meso or micro level (e.g. institutes, research groups or individual scientists), reference standards based on subject classification schemes may reveal a more differentiated picture (see also Schubert & Braun 1996).

CITATIONS DEPEND ON MANY FACTORS

The research activity of a research group, publication of their findings, and citation of the publications by colleagues in the field are all social activities. This means that citation counts for the group's publications are not only an indicator of the impact of their scientific work on the advancement of scientific knowledge. They also reflect (social) factors that are unrelated to the accepted conventions of scholarly publishing (Bornmann & Daniel 2008). 'There are 'imperfections' in the scientific communications system, the result of which is that the importance of a paper may not be identical with its impact. The 'impact' of a publication describes its *actual* influence on surrounding research activities at a given time. While this will depend partly on its importance, it may also be affected by such factors as the location of the author, and the prestige, language, and availability of the publishing journal' (Martin & Irvine 1983, p. 70). Bibliometric studies published in recent years have revealed the general influence of this and a number of other factors on citation counts (see Peters & van Raan 1994, Bornmann & Daniel 2008). In order to control for these factors, further independent variables in addition to the variable of actual interest (see 'Explorative data analysis' above) should be considered in the statistical analysis of bibliometric data (see e.g. Bornmann & Daniel 2006).

Field dependent factors

Citation practices vary between natural sciences and social sciences fields (Hurt 1987, Bazerman 1988, Braun et al. 1995a,b, Hargens 2000, Ziman 2000) and even within different areas (or clusters) within a single subfield (Lewison & Dawson 1998, Klamer & van Dalen 2002). According to Podlubny (2005), 'one citation in mathematics roughly corresponds to 15 citations in chemistry, 19 citations in physics, and 78 citations in clinical medicine' (p. 98). As the chance of being cited is related to the number of publications (and the number of scientists) in the field (Moed et al. 1985), small fields attract far fewer citations than more general fields (King 1987). For this reason, bibliometric comparisons of research groups should be conducted only within a field, or the fields in which the research groups work (or in which the publications appeared) should be included in the statistical analysis (see 'Reference standards' above).

Journal dependent factors

There is some evidence that the order in which an article falls within a journal issue matters considerably for the influence that the article gathers (Laband & Piette 1994, Smart & Waldfogel 1996). More precisely, the first article (Ayres & Vars 2000) or a lead paper (Hudson 2007) in a scientific journal tends to produce more citations than other articles (these order-factors

might become less relevant in the e-journal era). Furthermore, journal accessibility, visibility, and internationality (Vinkler 1987, Yue & Wilson 2004) — as well as the impact, quality, or prestige of the journal — may influence the probability of citations (Cronin 1984, Moed et al. 1985, Seglen 1989, Tainer 1991, Meadows 1998, Boyack & Klavans 2005, van Dalen & Henkens 2005). This means that in the statistical analysis of citation counts, the Journal Impact Factor (provided by Thomson Scientific) of the journal in which the cited publications appeared should be included.

Article dependent factors

Citation counts of methodology articles, review articles, research articles, letters, and notes (Shaw 1987, Cano & Lind 1991, MacRoberts & MacRoberts 1996, Aksnes 2006) differ considerably. For instance, review articles are generally expected to be cited more often than research articles. There is also a positive correlation between the citation frequency of publications and (1) the number of co-authors of the work (Lawani 1986, Baldi 1998, Beaver 2004), and (2) the number of the references within the work (Peters & van Raan 1994). And, as longer articles have more content that can be cited than do shorter articles, the sheer size of an article influences whether it is cited (Laband 1990, Stewart 1990, Abt 1993, Baldi 1998, Leimu & Koricheva 2005, Bornmann & Daniel 2007b, Hudson 2007). The document type, number of co-authors, number of references, and number of pages of the cited publication should accordingly be included in the statistical analysis as independent variables.

Author/reader dependent factors

The language in which a paper is written influences the probability of citations (Cronin 1981, Liu 1997, Kellsey & Knievel 2004, van Raan 2005b). English-language publications generally are expected to be cited more frequently than papers published in other languages. Results from Mählck & Persson (2000), White (2001), and Sandström et al. (2005) show that citations are affected by social networks: authors cite primarily works by authors with whom they are personally acquainted. Cronin (2005) finds this hardly surprising, as it is to be expected that personal ties become manifest and strengthened, resulting in greater reciprocal exchange of citations over time. Considering these findings, therefore, the analysis of bibliometric data should also control for possible language and network effects.

CONCLUSIONS

'Measurement of research excellence and quality is an issue that has increasingly interested governments, universities, and funding bodies as measures of accountability and quality are sought' (Steele et al. 2006, p. 278). Weingart (2005) notes that a general enthusiastic acceptance of bibliometric figures for evaluative purposes or for comparing the research success of scientists can be observed in institutions and government bodies today. The UK, for instance, is planning to base allocation of government funds for university research to a large extent on bibliometric indicators: 'The Government has a firm presumption that after the 2008 RAE [Research Assessment Exercise] the system for assessing research quality and allocating 'quality-related' (QR) research funding to universities from the Department for Education and Skills will be mainly metrics-based' (UK Office of Science and Technology 2006, p. 3).

With the easy availability of bibliometric data and ready-to-use tools for generating bibliometric indicators for evaluation purposes, there is a danger of improper use. We therefore recommend that the standards of good practice for analysis of bibliometric data and presentation and interpretation of the results presented here should always be considered. Conclusions comparing the research performance of research groups are valid only if (1) the scientific impact of the research groups and their publications is examined in a differentiated way using box plots as well as Lorenz curves and Gini coefficients (that is, in a way that goes beyond the typically used arithmetic mean value), (2) different reference standards are used to assess the impact of research groups and the appropriateness of the reference standards is examined critically, and (3) the fact that citations are a function of many influencing factors besides scientific quality is taken into consideration in the statistical analysis of citation counts for the publications of the group in question (e.g. Bornmann & Daniel 2006).

LITERATURE CITED

Abt HA (1993) Institutional productivities. Publ Astron Soc Pac 105:794–798

Adam D (2002) The counting house. Nature 415:726–729

Aksnes DW (2006) Citation rates and perceptions of scientific contribution. J Am Soc Inf Sci Technol 57:169–185

Atkinson AB (1970) Measurement of inequality. J Econ Theory 2:244–263

Ayres I, Vars FE (2000) Determinants of citations to articles in elite law reviews. J Legal Stud 29:427–450

Baldi S (1998) Normative versus social constructivist processes in the allocation of citations: a network-analytic model. Am Sociol Rev 63:829–846

Bazerman C (1988) Shaping written knowledge. The genre and activity of the experimental article in science. University of Wisconsin Press, Madison, WI

Beaver DB (2004) Does collaborative research have greater epistemic authority? Scientometrics 60:399–408

Bornmann L, Daniel HD (2006) Selecting scientific excellence through committee peer review — a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. Scientometrics 68:427–440

Bornmann L, Daniel HD (2007a) Functional use of frequently and infrequently cited articles in citing publications. A content analysis of citations to articles with low and high citation counts. In: Torres-Salinas D, Moed HF (eds) Proc 11th Int Conf Int Soc Scientometrics Informetrics. Spanish Research Council (CSIC), Madrid, p 149–153

Bornmann L, Daniel HD (2007b) Multiple publication on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine. J Am Soc Inf Sci Technol 58:1100–1107

Bornmann L, Daniel HD (2008) What do citation counts measure? A review of studies on citing behavior. J Doc 64:45–80

Boyack KW, Klavans R (2005) Predicting the importance of current papers. In: Ingwersen P, Larsen B (eds) Proc 10th Int Conf Int Soc Scientometrics Informetrics. Karolinska University Press, Stockholm, p 335–342

Braun T, Glänzel W, Grupp H (1995a) The scientometric weight of 50 nations in 27 science areas, 1989–1993. 1. All fields combined, mathematics, engineering, chemistry and physics. Scientometrics 33:263–293

Braun T, Glänzel W, Grupp H (1995b) The scientometric weight of 50 nations in 27 science areas, 1989–1993. 2. Life sciences. Scientometrics 34:207–237

Burrell QL (2006) Measuring concentration within and co-concentration between informetric distributions: an empirical study. Scientometrics 68:441–456

Cano V, Lind NC (1991) Citation life-cycles of ten citation-classics. Scientometrics 22:297–312

Cronin B (1981) Transatlantic citation patterns in educational psychology. Soc Sci Inf Stud 24:48–51

Cronin B (1984) The citation process; the role and significance of citations in scientific communication. Taylor Graham, Oxford

Cronin B (2005) Warm bodies, cold facts: the embodiment and emplacement of knowledge claims. In: Ingwersen P, Larsen B (eds) Proc 10th Int Conf Int Soc Scientometrics Informetrics. Karolinska University Press, Stockholm, p 1–12

Damgaard C, Weiner J (2000) Describing inequality in plant size or fecundity. Ecology 81:1139–1142

Daniel HD (2005) Publications as a measure of scientific advancement and of scientists' productivity. Learned Publishing 18:143–148

Daniel HD, Fisch R (1990) Research performance evaluation in German university sector. Scientometrics 19:349–361

Garfield E (1972) Citation analysis as a tool in journal evaluation: journals can be ranked by frequency and impact of citations for science policy studies. Science 178:471–479

Garfield E, Welljamsdorof A (1992) Citation data: their use as quantitative indicators for science and technology evaluation and policy-making. Current Contents 49:5–13

Garfield E, Malin MV, Small H (1978) Citation data as science indicators. In: Elkana Y, Lederberg J, Merton RK, Thackray A, Zuckerman H (eds) Toward a metric of science: the advent of science indicators. John Wiley, New York, p 179–207

Glänzel W, Schubert A, Czerwon HJ (1999) An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. Scientometrics 44:427–439

Hargens LL (2000) Using the literature: reference networks, reference contexts, and the social structure of scholarship. Am Sociol Rev 65:846–865

Huber PJ (1981) Robust statistics. Wiley, New York, NY

Hudson J (2007) Be known by the company you keep: citations: quality or chance? Scientometrics 71:231–238

Hurt CD (1987) Conceptual citation differences in science, technology, and social sciences literature. Inf Process Manag 23:1–6

Kellsey C, Knievel JE (2004) Global English in the humanities? A longitudinal citation study of foreign-language use by humanities scholars. Coll Res Libr 65:194–204

King DA (2004) The scientific impact of nations: what different countries get for their research spending. Nature 430:311–316

King J (1987) A review of bibliometric and other science indicators and their role in research evaluation. J Inf Sci 13:261–276

Klamer A, van Dalen HP (2002) Attention and the art of scientific publishing. J Econ Methodol 9:289–315

Kostoff RN (2002) Citation analysis of research performer quality. Scientometrics 53:49–71

Laband DN (1990) Is there value-added from the review process in economics? Preliminary evidence from authors. Q J Econ 105:341–352

Laband DN, Piette MJ (1994) Favoritism versus search for good papers: empirical evidence regarding the behavior of journal editors. J Polit Econ 102:194–203

Lawani SM (1986) Some bibliometric correlates of quality in scientific research. Scientometrics 9:13–25

Leimu R, Koricheva J (2005) What determines the citation frequency of ecological papers? Trends Ecol Evol 20: 28–32

Lewison G, Dawson G (1998) The effect of funding on the outputs of biomedical research. Scientometrics 41:17–27

Liu ZM (1997) Citation theories in the framework of international flow of information: new evidence with translation analysis. J Am Soc Inf Sci 48:80–87

MacRoberts MH, MacRoberts BR (1996) Problems of citation analysis. Scientometrics 36:435–444

Mählck P, Persson O (2000) Socio-bibliometric mapping of intra-departmental networks. Scientometrics 49:81–91

Martin BR, Irvine J (1983) Assessing basic research: some partial indicators of scientific progress in radio astronomy. Res Policy 12:61–90

May RM (1997) The scientific wealth of nations. Science 275:793–796

McClellan JE (2003) Specialist control: the publications committee of the Academie Royal des Sciences (Paris) 1700–1793. Transactions of the American Philosophical Society, Vol. 93, American Philosophical Society, Philadelphia, PA

Meadows AJ (1998) Communicating research. Academic Press, London

Moed HF (2005) Citation analysis in research evaluation. Springer, Dordrecht

Moed HF, Burger WJM, Frankfort JG, van Raan AFJ (1985) The use of bibliometric data for the measurement of university research performance. Res Policy 14:131–149

Neuhaus C, Daniel HD (2008) Data sources for performing citation analysis: an overview. J Doc (in press)

Peters HPF, van Raan AFJ (1994) On determinants of citation scores: a case study in chemical engineering. J Am Soc Inf Sci 45:39–49

Podlubny I (2005) Comparison of scientific impact expressed by the number of citations in different fields of science. Scientometrics 64:95–99

REPP (Research Evaluation and Policy Project) (2005) Quantitative indicators for research assessment: a literature review. REPP Discussion paper 05/1, Research Evaluation and Policy Project, Research School of Social Sciences, The Australian National University, Canberra

Rinia EJ, De Lange C, Moed HF (1993) Measuring national output in physics: delimitation problems. Scientometrics 28:89–110

Ross SM (2007) Introduction to probability models. Elsevier, London

Sandström U, Wadskog D, Karlsson S (2005) Research institutes and universities: Does collaboration pay? In: Ingwersen P, Larsen B (eds) Proc 10th Int Conf Int Soc Scientometrics Informetrics. Karolinska University Press, Stockholm, p 690–691

Schubert A, Braun T (1996) Cross-field normalization of scientometric indicators. Scientometrics 36:311–324

Schubert A, Glänzel W (1983) Statistical reliability of comparisons based on the citation impact of scientific publications. Scientometrics 5:59–74

Seglen PO (1989) From bad to worse — evaluation by journal impact. Trends Biochem Sci 14:326–327

Shadbolt N, Brody T, Carr L, Harnad S (2006) The open research web: a preview of the optimal and the inevitable. In: Jacobs N (ed) Open access: key strategic, technical and economic aspects. Chandos, Oxford, p 195–208

Shaw JG (1987) Article-by-article citation analysis of medical journals. Scientometrics 12:101–110

Smart S, Waldfogel J (1996). A citation-based test for discrimination at economics and finance journals. NBER Working Paper, No. 5460. National Bureau of Economic Research, Cambridge, MA

Smith LC (1981) Citation analysis. Libr Trends 30:83–106

Steele C, Butler L, Kingsley D (2006) The publishing imperative: the pervasive influence of publication metrics. Learned Publishing 19:277–290

Stewart JA (1990) Drifting continents and colliding paradigms: perspectives on the geoscience revolution. Indiana University Press, Bloomington, IN

Tainer JA (1991) Science, citation, and funding. Science 251:1408

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, MA

UK Office of Science and Technology (2006). Science and innovation investment framework 2004–2014: next steps. UK Office of Science and Technology, London

van Belle G, Fisher LD, Heagerty PJ (2004) Biostatistics. Wiley, New York

van Dalen HP, Henkens KE (2005) Signals in science: on the importance of signaling in gaining attention in science. Scientometrics 64:209–233

van Leeuwen TN (2007) Modelling of bibliometric approaches and importance of output verification in research performance assessment. Res Eval 16:93–105

van Raan AFJ (2004) Measuring science. Capita selecta of current main issues. In: Moed HF, Glänzel W, Schmoch U (eds) Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems. Kluwer, Dordrecht, p 19–50

van Raan AFJ (2005a) Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics 62:133–143

van Raan AFJ (2005b) For your citations only? Hot topics in bibliometric analysis. Measurement 3:50–62

Vinkler P (1987) A quasi-quantitative citation model. Scientometrics 12:47–72

Weingart P (2005) Das Ritual der Evaluierung und die Verführbarkeit. In: Weingart P (ed) Die Wissenschaft der Öffentlichkeit: Essays zum Verhältnis von Wissenschaft, Medien und Öffentlichkeit. Velbrück, Weilerswist, p 102–122

White HD (2001) Authors as citers over time. J Am Soc Inf Sci Technol 52:87–108

Yue W, Wilson CS (2004) Measuring the citation impact of research journals in clinical neurology: a structural equation modelling analysis. Scientometrics 60: 317–332

Ziman J (2000) Real science: what it is, and what it means. Cambridge University Press, Cambridge

Editorial responsibility: Howard Browman, Storebø, Norway and Konstantinos Stergiou, Thessaloniki, Greece

Submitted: August 29, 2007; Accepted: November 18, 2007
Proofs received from author(s): January 28, 2008