Deep impact – evaluation in the sciences

Prof. Dr. Elmar Brähler; director of the Department of Medical Psychology and Medical Sociology, University Medial School in Leipzig. Dr. Manfred Beutel is professor of the Department of Psychosomatic and Psychotherapy, University Medical School in Gießen. Dr. Oliver Decker is assistant at the Department of Medical Psychology and Medical Sociology, University Medical School in Leipzig

Summary

The purpose of the paper is to evaluate the psychometric properties of the impact factor as an assessment procedure. Detailed criteria regarding theoretical underpinnings, test administration, scoring and interpretation are applied. The impact factor appears to be of limited use for deciding which journals to subscribe. It is not suitable for evaluating achievements of individual scientists and research groups. The impact factor contains serious sources of errors and flaws resulting in strong biases against culture- and language-bound medical subspecialties and non-Anglo-American countries. Possible alternatives are discussed.

The imperative "publish or perish" has become the selection criterion for scientific evolution: who publishes, stays, who does not publish, disappears. This requirement has characterized the evaluation of scientific quality for a long time: the number counts. A great number of publications makes scientists visible. Scientists have to pay for recognition, and their currency are publications. Society has implicitly and explicitly formulated expectations for achievement as an exchange for the providing funding. Expecting scientists to render an account for their funding can be neither surprising nor offending. However, in a society based on equivalence, what cannot be compared remains suspect. Value is ascribed only to what can be converted into Euros and Cents. Thus neither the effort to evaluate scientific publications nor the societal reference of evaluation is new. However, the development of electronic data processing has produced a new measure in the last years which is to be tested for its suitability: the impact factor.

Soz.- Präventivmed. 49 (2004) 10–14
0303-8408/04/010010–5
DOI 10.1007/s000038-003-2108-5

Particularly in German medicine, the impact factor has acquired a dominant role in the evaluation of scientific achievements. It is calculated by the Institute for Scientific Information (ISI) as part of the Thomson Company (Frankfurter Allgemeine Zeitung, 27.02.2002, p. N3). It is basically a by-product of computerized data bases of the Social Science Citation Index (SSCI) and of the Science Citation Index (SCI; Brähler & Decker 2002). The Association of the Scientific-Medical Societies (AWMF) has based its recommendations to evaluate scientific achievement on this impact factor (Frömter et al. 1999). In current practice, the impact factor is also used to evaluate the qualification of scientists – against the original intention of the ISI (Garfield 1995) – and to guide the distribution of funding to institutes and clinics. Here, Germany takes a prominent role. Only Finland has gone further where distribution of funding according to impact factors has already become part of the legislation (Adam 2002).

As the impact factor has acquired critical importance, this paper evaluates the impact factor as an assessment procedure: What are its psychometric properties? Criteria for the assessment of the impact factor refer to the theoretical underpinnings, test administration, interpretation and evaluation (Schumacher & Brähler 1997: Tab. 1).

1. Theoretical underpinnings

a) *Diagnostic purposes:* The purpose of the impact factor is to evaluate the significance of a journal for the scientific community. Thus the factor should guide decisions of libraries which journals to subscribe. The factor is not to be used as an instrument for guiding research activities and for evaluating achievements of individual persons (Garfield 1985).

b) *Theoretical assumptions:* The impact factor is based on the notion that scientists indirectly judge the quality of a journal by quoting papers from it. When papers from a journal are quoted frequently, the significance of this journal must be high. High acceptance of a research group in turn is manifested in the quality of the journal in which they can publish.

c) *Explicitness of test construction:* About 2 000 scientific journals are assessed for the SCI and the SSCI. Recorded are authors, titles of the papers and other papers quoted in the manuscripts. From the papers quoted, first authors, titles and journals are documented. Based on these quotations, the impression a journal makes in a scientific field is calculated by the frequencies of its quotations by authors of other papers. The papers for which ISI calculates an impact factor must be listed in the SCI or SSCI. The impact factor of a journal is based on the number of quotations in the year x, in papers listed by ISI, only counting papers in the two previous years (x-1 and x-2; Brähler et al. 2001).

2. Test administration

a) *Objectivity of administration and* b) *transparency:* The ISI is a private company. The decision for listing a journal rests with the ISI. The data processing is done by the ISI, for control purposes the quotations listed are compared with the original and the review papers from the original journals listed. Quotations which cannot be assigned to a specific paper are not checked further.

Table 1 Criteria for the evaluation of tests (based on Schumacher & Brähler 1997)

1. **Theoretical underpinnings**
   a) Diagnostic purposes
   b) Theoretical assumptions
   c) Explicitness of test construction
2. **Test administration**
   a) Objectivity of administration
   b) Transparency
   c) Degree of susceptibility to deception
   d) Susceptibility to other sources of error, e.g., sensitivity of trait procedures to current states of the person
3. **Scoring and interpretation**
   a) Objectivity of scoring and interpretation
   b) Reliability
   c) Validity
   d) Standardization
   e) Gain of information
   f) Benefit, e.g., indicators derived from the test
4. **Test evaluation**
   a) Cost-offset ratio (economical)
   b) Fairness, e.g., degree of systematic discrimination of specific groups of persons based on specific sociodemographic characteristics
   c) Comparability to other tests
   d) Utility

c) *Degree of susceptibility to deception:* Journals which are assessed as original journals by ISI can manipulate their factor by different means. Preferential quoting of papers from the own journal of the two previous years is one way, agreements about mutual quotations with other original journals another one. In addition, there are editorial possibilities to increase the impact factor: the number of published original papers is reduced, instead, for instance, manuscripts are published as letters to the editors.

3. Scoring and interpretation

a) *Objectivity of scoring and interpretation:* A source of error which questions objectivity lies in the basis for calculating the impact factor. The factor itself is a ratio. The numerator is based on quotations of papers from original journals of the past two years. As a denominator, original papers and review articles are listed which a journal has published in the respective time period. The ISI only counts original papers and reviews for the denominator, but accepts all quoted papers from this period for the numerator – including letters to the editors, editorials, congress reports or book reviews. Therefore journals are favoured which published papers in many categories. Thus the impact factor is presumably overestimated by up to 40% in some journals (Adam 2002) which have many categories of articles like "Nature", "Lancet" or "Science".

A study by Moed could further show that the systematic error quote of an SCI inquiry was around 7% (Moed 2002), based on errors of documentation. In certain countries, however, error ratios were up to 13 or 18% (Spain and China) due to uncommon vowels (ä, ö, ü etc.) or special letters (ù, ò, Ø, ï etc.).

b) *Reliability:* Different varieties of names (e.g., of printing unfamiliar vowels) make the procedure quite unreliable. Even journals frequently are not quoted with the abbreviations introduced by the ISI. For instance, the journal "Psyche – Zeitschrift für Psychoanalyse und ihre Anwendungen" and its papers are abbreviated as "Psyche". The ISI, however, only counted quotations with the abbreviations "Psyche – Z Psychoanal" (Brähler et al. 2001). The raw data must therefore be considered a major source of error. Even the correctness of citations cannot be controlled by the ISI. If the author of a paper makes erroneous quotations – which is probably not infrequent (Adam 2002) – the data basis gets wrong.

In general, the quality of a paper and consequently the achievement of a study group or an individual person cannot be inferred from the score of the impact factor. About 15% of the papers in the journals listed by ISI

count for 85% of quotations of these journals (Seglen 1997).

c) *Validity:* The impact factor favors papers in journals, and there is an additional bias selecting publications only from the two previous years. In a study, Krell (2002) found that 98.5% of quoted papers in selected articles were older than two years. This has been confirmed by our results. In medical specialties which are culture- and language-bound, frequently books and papers are quoted which have been published decades ago (Decker & Brähler 2001a). This is even true for the journal "Sozial- und Präventivmedizin" whose authors quoted mostly papers, but to a substantial proportion also quoted books in the year of 2001. Even clearer and likely due to the international orientation, the ratio of papers in English versus German language was 7:1. Even in SPM only six quotations per paper dated less than two years back (Tab. 2).

Clearly, the one-side bias toward more recent publications (only those are counted in the impact factor), disregards papers which are considered important by authors of Sozial- und Präventivmedizin.

The validity of the impact factor must be questioned for an additional reason: Of course, it is plausible that a paper is particularly good when it is quoted frequently. However, the difference between a plausible assumption and a true statement is known by each scientist. With the same justification the position can be taken that controversial papers are quoted with high frequency. In certain cases, even bad or faulty papers are likely to be quoted frequently.

The impact factor has also effects reflecting back on its validity. One effect is that more and more becomes published. When a lot of papers are published, a lot of papers are quoted: Thus, in the US one per two million papers have been recorded by the ISI from 1992 to 1996, much more than in Germany (0.2 million) or Great Britain (0.3 million). On the average, American papers have been quoted five times, German papers only three times and English papers four times. When more gets published, more gets quoted (Nature 1997).

This fact may explain the long-known, so-called Matthew effect in citations (Merton 1968) which has been confirmed in analyses from past years (Bonitz & Scharnhorst 2002). This concept refers to the effect that papers from a small number of countries with a high expectancy of quotations are quoted even more frequently than expected. Papers from the majority of countries, however, are quoted less frequently than expected. Winners are e.g., the USA, Great Britain and Switzerland. Losers are the People's Republic of China and Turkey. Bonitz and Scharnhorst attribute this effect to marketing of publication practice: Authors who want to gain a high result for their research select the papers for presenting their results which have a high impact factor. Aiming at a certain horizon of recipients determines presentation of results and research practice: Which product will be successful on the market of sciences, becomes the crucial issue. The impact factor structures the landscape of science by the financial means associated. Primarily, the impact factor reflects which scientists are able to correctly evaluate the horizon of recipients of a journal and which research areas are particularly well established in order to publish a lot and quote a lot.

d) *Standardization:* There is no standardization of the impact factor according to countries or scientific fields. Specialties are compared independently of their sizes. However, the factor varies according to countries and scientific fields. Journals from the US have a mean factor of 1.44 in the category "Other studies and professions allied to medicine", in Germany, the mean score is only 0.73 (Adams 1998). The difference with regard to sociology is even more pronounced: Here, journals produced in the US have a mean factor of 1.39, in Germany, they have a score of 0.49 (Adams 1998). As there is no standardization, the impact factor is also denoted as "the poor man's citation analysis" (Adams 2002). As an exception the AWMF recommends weighing book chapters and books according to a certain key in order to compensate for differential habits of citation.

e) *Gain of information:* The number of publications in listed journals correlates highly with the total number of publications (Kotiaho et al. 1999): Scientists who publish a lot, publish a lot in English and also publish in books (Decker & Brähler 2001b). The additional gain of the impact factor as compared to a simple comparison of the absolute number of publication is close to zero.

Table 2 Average references per original paper in the journal "Sozial- und Präventivmedizin" in the year 2001 (without supplements)

Books vs. journals	11 (s = 6) vs. 19 (s = 10)
English vs. German	22 (s = 12) vs. 7 (s = 7)
Before 1999 vs. after 1999	24 (s = 10) vs. 6 (2 = 5)

4. Test evaluation

a) *Cost-offset ratio:* The high correlation raises the question if the procedure is economical (see 3c).

b) *Fairness:* The experience of applying the impact factor shows that the procedure produces unfairness not only by its faulty usage, but also by the neglect of specificities of

culture and language. The erroneous practice of dealing with unfamiliar vowels (Brähler & Decker 2002) can only be understood as a discrimination against scientists from non-Anglo-American countries. The comparison of different scientific fields based on the impact factor leads to a one-sided preferential funding of large disciplines. Counting only articles from the last two years places social science journals at a disadvantage, as papers unfold their scientific impact only after longer time periods in these fields. When the less recent papers are quoted, they do not count for the calculation of the impact factor in the publishing journal. The calculation of the impact factor lacks transparency. The ISI is a commercial institute and the decision to count a journal as an original journal cannot be controlled based on scientific reasoning. Necessary criteria for acceptance are known, but sufficient criteria are not published. Therefore the process of decision-making is not accountable.

d) *Utility:* Based on applying the impact factor the overall evaluation is negative. The test procedure is only partially suitable in order to serve as a basis for the decision which journals should be subscribed. It should be kept in mind that only journals from one field are related to one another as an interdisciplinary comparison is not possible. It must be further taken into account that only a very small group of journals is considered for the calculation of the impact factors and that selection criteria are rather diffuse. The test procedure is not at all suitable for evaluating the achievement of individual persons and research groups. For one, the impact factor of a journal can only be deduced from a small proportion of articles in this journal, and for another reason, the procedure is fraught with an extremely high error rate. Thus, it particularly disadvantages culture- and language-bound medical specialties.

The procedure is more demanding than other possible techniques and contains systematic errors. These could be partially compensated by a standardization (e.g., bias by language and country effects). Partially they cannot be controlled as the data bases (quotations) cannot be evaluated regarding their correctness. New media, particularly based on the internet, are not regarded in the analysis of quotations, as these are not referenced according to the usual abbreviations. A "Euro impact", based on a scientific, democratic institution could be a solution. Another problem remains; the market orientation does have an impact on the science landscape. Effects occur here similar to those which lead to cyclical crises of the real economy. Recently observed in bond transactions of the new market, "exchange value" and "utility" of products may diverge. The increase of publications with a high exchange value (high impact factor) is increasingly contrasted to a decrease of utility of publications. This can even lead to making up publications. This is another consequence of the market orientation of science. There is still no other way to evaluate the quality of scientific papers, but to read them.

Oliver Decker, Manfred E. Beutel and Elmar Brähler

Zusammenfassung

Ziel des Manuskriptes ist die Evaluation der psychometrischen Eigenschaften des Impact-Faktors als ein Testinstrument. Detaillierte Kriterien bezüglich Testgrundlage, -durchführung, -auswertung und -evaluation werden angewendet. Der Impact-Faktor erscheint von bedingtem Nutzen für die Auswahl des Abonnements einer Zeitschrift. Er ist nicht geeignet für die Bewertung von Leistungen individueller Wissenschaftler und Forschergruppen. Der Impact-Faktor enthält ernste Fehlerquellen, die zu einer starken Benachteiligung kultur- und sprachgebundener medizinischer Fächer und nicht-angloamerikanischer Länder führt. Mögliche Alternativen werden diskutiert.

Résumé

Le but de ce manuscrit est l´évaluation des caractéristiques psychométriques du facteur d'impact comme outil de test. Des critères détaillés sont appliqués, quant à la base du test, sa procédure, sa validation et son évaluation. Le facteur d'impact semble d´utilité limitée pour le choix d´un abonnement à un magazine. Il n´est pas approprié pour la validation des performances des scientifiques ou groupes de chercheurs individuels. Le facteur d'impact contient des sources d´erreurs graves menant à une défavorisation des disciplines liées à certaines cultures ou langues et des pays non anglo-américains. Des alternatives possibles seront discutées.

Soz.- Präventivmed. 49 (2004) 10–14

References

Adams J (1998). Benchmarking international research. Nature *396*: 615-8.

Adam D (2002). The counting house. Nature *415*: 726-9.

Bonitz M, Scharnhorst A (2002). Wissenschaft und Ökonomie – wissenschaftsmetrische Bemerkungen. In: Parthey H, Spur G, eds. Wissenschaft und Innovation. Berlin: Gesellschaft für Wissenschaftsforschung: 85–95.

Brähler E, Decker O, Borkenhagen A (2001). Das Wahre, Schöne, Gute oder schöne, gute Ware? Psyche *54*: 1245–52.

Brähler E, Decker O (2002). Der Impaktfaktor: Erbsenzählerei oder valides Instrument zur Leistungsmessung? Journal der Deutschen Gesellschaft für Plastische und Wiederherstellungschirurgie *14*: 9–12.

Decker O, Brähler E (2001a). Von Büchern und Zeitschriften – Diskussion der Bewertung wissenschaftlicher Leistungen in den kultur- und sprachgebundenen Fächern in der Medizin. Zeitschrift für Klinische Psychologie, Psychiatrie und Psychotherapie *49*: 235–46.

Decker O, Brähler E (2001b). Veröffentlichungspraxis als Bewertungsmassstab – am Beispiel der Lehrstuhlinhaber in der Medizinischen Psychologie und in der Psychosomatik. Psychosomatik, Psychotherapie, medizinische Psychologie *51*: 288–95.

Frömter E, Brähler E, Langenbeck U, Meenen NM, Usadel KH (1999). Das AWMF-Modell zur Evaluierung publizierter Forschungsbeiträge in der Medizin. Dt Med Wochenschr *124*: 910–5.

Garfield E (1985). Use and misuse of citation frequency. Curr Contents *43*: 3–9.

Kotiaho JS, Tomkins J, Simmons L (1999). Unfamiliar citations breed mistakes. Nature *400*: 307.

Krell F-T (2002). Why impact factors don't work for taxonomy. Nature *415*: 957.

Merton RK (1968). The Matthew effect in science. Science *159*: 56–63.

Nature (1997). News: EU eliminates citation gap with America. Nature *387*: 537.

Moed HF (2002). The impact-factors debate: the ISI's uses and limits. Nature *415*: 731–2.

Schumacher J, Brähler E (1997). Testdiagnostik in der Psychotherapie. In: Senf W, Broda M, eds. Praxis der Psychotherapie: theoretische Grundlagen von Psychoanalyse und Verhaltenstherapie. Stuttgart: Thieme: 47–56.

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

Address for correspondence

Prof. Dr. Elmar Brähler
Universitätsklinikum Leipzig
Stephanstr. 11
D-04103 Leipzig
e-mail: brae@medizin.uni-leipzig.de