Challenges in Ranking of Universities

Invited paper for the First International Conference on World Class Universities, Shanghai Jaio Tong University, Shanghai, June 16-18, 2005

Anthony F.J. van Raan
Centre for Science and Technology Studies
Leiden University
Wassenaarseweg 52, P.O.Box 9555
2300 RB Leiden, The Netherlands

Abstract

In this paper we discuss recent developments in rankings of universities and the impact of these rankings on academia in the context of international benchmarking and evaluation. We focus on technical and methodological problems behind these rankings, particularly those based on bibliometric methods, with special attention to the social sciences and humanities. We criticize the recent expert-based rankings by showing that the correlation of expert score and bibliometric outcomes is practically zero. This finding casts severe doubts on the reliability of these expert-based rankings. New approaches are proposed on the basis of advanced bibliometric methods. It is argued on the basis of preliminary results that, probably due to 'finite size' considerations, a league of outstanding universities worldwide will not have much more than around 200 members. Finally, we discuss the challenges for further research and practical applications.

1. Context and Central Questions

Scientific research has always been an international enterprise. The growth in the last few decades of worldwide R&D activities as well as the higher education sector in developed and developing countries reinforced established academic institutions and created many new ones. At the same time, the number and intensity of student and researcher exchange programs, international collaboration, and working stays outside the own country rapidly increased, intensified by the ever-growing worldwide mobility. An important parallel development, stimulated too by the growth of the higher education sector, is the strong demand for accountability, evidence of quality, and 'value for money'. This latter development gave rise to an expanding *evaluation culture*. This evaluation culture is based on elements present for already a very long time in the scientific world such as judgment of research quality by peer review. All these developments led to an increasing competition for financial support and for the best students and researchers between universities within nations and worldwide. Universities strive for belonging to the top of their country, or even to the world top. This implies the existence of some kind of a league (a national one, or an international one), to which one can only be admitted on the basis of performance. The higher the performance, the better chances a university has to become a member of an elite league and to reach a high ranking position in this league. Clearly, the basic question is: *How can we identify the best universities in the world?*

From this basic question, we derive several research questions that have to be answered in the context of ranking procedures. First, to which activities are the leagues related? In the case of universities, to teaching or to research? Or to both of these main academic tasks? Second, how do we measure the performance with respect to the chosen activity in an international context? Third, is it possible to conduct these performance measurements for, in fact, all universities in the world in a relatively easy but still reliable and valid way? Fourth, is it possible to express this performance in just one numerical value? If this is the case, then, in principle, a ranking can be made. This, however, immediately implies a next, fifth, question: what is the statistical significance of the difference in, for instance, ranking position 10 and 15. The sixth question is, like in a football (soccer) league: how many members can or should be covered by a specific league? In other words, at what position of the ranking should the league be ‘closed’?

The first question can be answered by making a decision. This paper deals with scientific research. But this choice does not mean that teaching has become irrelevant. Academic research is closely interwoven with teaching, and strength of research is particularly important at the M.Sc. and Ph.D. level. Thus, in the international competition for talented students, scientific performance of a university does matter. The second question leads us to the problem of performance measurement. As discussed above, this problem is not new. Peer review procedures and bibliometric analysis are the main methods to evaluate research performance of universities. So the most crucial question is the third one: *how much effort is a reliable evaluation of an entire university, and, as a consequence, will such an evaluation be possible for all universities in the world, in a short period of time, and against reasonable costs?*

Let me give an example. Since the early 1990’s, universities in the Netherlands are evaluated by discipline (VSNU 2002). Each year the focus is on three to four disciplines, for instance chemistry, biology, psychology, civil engineering. In the next year, three to four other disciplines are evaluated, and so on. The whole procedure has a ‘cycle time’ of about five years, so after five years there is new evaluation round for chemistry. In these nation-wide, discipline-specific evaluations, all departments and research groups in a specific discipline in all thirteen universities in the Netherlands are subject to an international peer review. In most natural and life sciences also an extensive bibliometric analysis is involved. The international peer committee of between five and ten members meets in the Netherlands, there are discussions with heads of departments, in some cases site visits are organized, reports have to be written, and so on. Staff members of the VSNU (the Association of the Universities in the Netherlands) have to prepare these evaluations, organize the peer committee, commission the bibliometric studies, etcetera. It is not difficult to imagine that these evaluations cost in the order of fifty thousand euros per discipline. With about 20 major disciplines the entire evaluation procedure (peer review, bibliometric analysis) per university within one cycle of five years is in the order of hundred thousand euros.

Clearly, this kind of thorough evaluation, meant as a basis for institutional research management, cannot be applied in a relatively short time, on a worldwide scale. Is this type of evaluation too ‘fine-tuned’ and could it be possible to evaluate universities on a broader level, with less attention for details, and more focus on the ‘overall quality’ of a university? Would a survey of, say, 1000 scientists worldwide provide us a reasonably reliable, broad picture of the research performance of a great number of universities all over the world? Would a carefully designed bibliometric analysis be able to do that? And, if both approaches were taken, would both approaches converge to the same results? Are rankings a reliable means of benchmarking universities against a global standard? In other worlds, is it possible to establish, on the basis of a survey, or with bibliometric analysis, or a combination of both, just one mark for each university -which means, all disciplines are taken together, or at least all natural and life sciences (including medicine). Or should we try to qualify universities at the level of major parts of science, such as natural sciences, life sciences, engineering, social sciences, humanities. And if we choose for this differentiation, would an aggregation of the (weighted) marks for the different disciplines to one final mark yield a meaningful outcome?

This paper addresses pitfalls and challenges in the worldwide ranking of universities. Our approach is based on the principle that performance must be reflected in observable evidence. We will discuss the conditions under which it is possible to evaluate in a reliable and valid way the research strengths of universities with highly automated procedures, within a reasonable time perspective and with reasonable costs. In this discussion, bibliometric elements such as numbers of publications and citations to these publications, play a crucial role.

The structure of this paper is as follows. In Chapter 2 we discuss recent developments in rankings and their impact on academia. Chapter 3 addresses technical and methodological problems behind these rankings. In Chapter 4 we discuss new approaches, particularly those based on bibliometric methods, with special attention to the social sciences and humanities. Also a first attempt to estimate the number of worldwide top-universities is made. In Chapter 5 further observations and outlook to near future developments are presented.

2. Recent Developments in Rankings

In the last few years rankings of universities, though controversial, have become increasingly popular: US News rankings, the UK Sunday Times University Guide and the Guardian's Guide to Universities, the Canadian Maclean's University Ranking in Canada, the German CHE University Ranking, the Asia Week's Best Universities in Asia, and others. Recently, two ranking publications widely attracted the attention of policy makers, the scientific world and the public media: the rankings published by the *Jiao Tong University* in Shanghai ('Academic Ranking of World Universities', SJTU 2003; 2004; 2005) and the rankings published by *Times Higher Education Supplement* ('World University Rankings', THES 2004, 2005).

These and other rankings suggest a similar simplicity for the evaluation of scientific performance as in the case of a football league. The immediate observation that the well-known US top-universities take the lead reinforces these suggestions. Universities responded enthusiastically, particularly if they felt their position was worth publishing and that quality of research is indeed a key contributing factor to these rankings (Umass 2004). Although things are not so simple and the various methodologies used in these ranking still have to be discussed thoroughly, the influence of these rankings is striking. Rankings have become unavoidable, and they will remain part of the academic life. General ideas about the international reputation and position of universities will be influenced considerably by rankings, high ranking universities will ‘advertise’ their position on the rankings as final truths about their research quality, and so these ranking may considerably guide the choice of young scientists and research students (Wheeler 2005). Ranking lists are changing the worldwide academic landscape.

The Times Higher Education Supplement (THES) is published by TSL Education Ltd of London, a leading educational publisher in the United Kingdom. The company is a subsidiary of News International Publishers Limited, which prints The Times and The Sunday Times. In the THES top-200 universities worldwide rankings, the opinions of scientists worldwide play a crucial role. Around 1,300 researchers in 88 countries were asked to mention the best universities in the geographic regions and the fields in which they considered themselves sufficiently qualified to judge the scientific standing of universities. These ‘peer review’ assessments counted for fifty per cent in the total score of a university. Four further criteria were used. Research impact in terms of citations per faculty member and staff student ratios, each of which account for twenty per cent of the score; the percentage of students and staff recruited internationally, each at five per cent of the total. Thus, in the THES rankings the bibliometric element counts for 20 per cent. These bibliometric data are derived from a commercial product, the Essential Science Indicators database produced by Thomson Scientific (the former Institute for Scientific Information, ISI). We notice that without knowing details of indicator design and calculation, one works in fact with a 'methodological black box'. Finally, the scores used in the ranking were normalized against a score of 1,000 for top-ranked Harvard University.

The crucial difference between the THES rankings and rankings produced by the Shanghai Jiao Tong University is that the latter do not include 'peer review'. The Shanghai rankings are based on four criteria. First, 'quality of education' over a long period, in terms of the number of alumni[1] of an institution winning Nobel Prizes and Fields Medals. This criterion counts for 10 per cent in the ranking. Next, two 'quality of faculty' measures, one non-bibliometric, namely staff[2] of an institution winning Nobel Prizes and Fields Medals, counting for 20 per cent, and a bibliometric measure, the number of highly cited researchers in 21 broad subject categories, also counting for 20 per cent. The following two criteria concern research output, and thus are bibliometric measures too: the number of articles published in Nature and Science in the period 1999-2003 (20 per cent)[3], and the number of articles[4] covered by the Science Citation Index (expanded version) and the Social Science Citation Index[5] for the year 2003 (also counting for 20 per cent in the ranking). Finally, the Shanghai group uses a kind of normalization criterion[6], the 'academic performance' with respect to the size of an institution (counting for 10 per cent). Thus, 60 per cent of the total score is based on bibliometric data. For each indicator, the highest scoring institution is assigned a score of 100, and other institutions are calculated as a percentage of the top score. Scores for each indicator are weighted to obtain overall scores for each institution.

[1] Alumni are defined as those who obtained a bachelor, master, or doctoral degree from the institution. The weight is 1 for alumni obtaining degrees in 1991-2000, 0.9 for alumni obtaining degrees in 1981-1990, 0.8 for alumni obtaining degrees in 1971-1980, and so on, and finally 0.1 for alumni obtaining degrees in 1901-1910. If a person obtains more than one degree from an institution, the institution is considered once only.

[2] Staff is defined as those who work at an institution at the time of winning the prize. Different weights are set according to the periods of winning the prizes. The weight is 1.0 for winners in 2001-2004, 0.9 for winners in 1991-2000, 0.8 for winners in 1981-1990, 0.7 for winners in 1971-1980, and so on, and finally 0.1 for winners in 1911-1920. If a winner is affiliated with more than one institution, each institution is assigned the proportional fraction. Nobel prizes are often shared by more than one scientist, winners are assigned proportional fractions.

[3] A weight of 1.0 is assigned to the affiliation of the 'corresponding author', 0.5 for first author affiliation (second author affiliation if the first author affiliation is the same as corresponding author affiliation), 0.25 for the next author affiliation, and 0.1 for all other author affiliations. Only publications of the 'normal article type' are included in the analysis. For institutions specialized in humanities and social sciences such as London School of Economics, Nature and Science publications are not considered, and the weight of this ranking element is reallocated to other indicators.

[4] Only publications of the 'normal article' type are included.

[5] In this paper we use the term 'CI' (Citation Index) for the Web of Science version of the Science Citation Index (SCI-Expanded), the Social Science Citation Index (SSCI), the Arts & Humanities Citation Index (AHCI). These indexes are produced and published by Thomson Scientific (the former Institute for Scientific Information, ISI) in Philadelphia.

[6] Total scores of the above five indicators are divided by the number of full-time equivalent academic staff. If the number of academic staff for institutions of a country could not be obtained, the weighted total scores of the above five indicators is used. For the ranking of 2005 (SJTU 2005), numbers of full-time equivalent academic staff are obtained for institutions in USA, Japan, China, Australia, and several European countries, for instance Italy, Netherlands, Sweden, Switzerland and Belgium.

From the above we can make first observations. Both the Shanghai and the THES ranking do not cover measures of performance related to 'earning' research income, such as grants from national research councils, form governmental programs and funding from industry. Smaller universities, and particularly those with an emphasis on social sciences and humanities, will have a better chance by the peer review element in the THES ranking as compared to the Shanghai study. A striking example is the difference in position of the London School of Economics in the THES and in the Shanghai ranking: rank 11 versus rank between 202-301, respectively. In general, studies based on bibliometric analyses will always have difficulties to cover the social science and humanities properly, given the inherent limitation of the bibliometric methodology with respect to these disciplines. But also in peer-based analyses the problem is to find an adequate coverage of scientists in the relevant social science and humanities fields because of the many different 'schools of thought' in these fields. Quality of teaching, particularly in the M.Sc. and Ph.D. level, and with that, the international attractiveness for young people, is not included in both ranking approaches. Perhaps only the THES criterion concerning the percentage of students and staff recruited from overseas, relates to this quality of teaching element.

In Germany, the Centre for the Development of Universities (Centrum für Hochschulentwicklung, CHE) applies a broader range of criteria to judge the standing of German universities, with a focus on specific disciplines. Thus, CHE does not provide an 'overall ranking' (which university is the best), but applies a more differentiated approach: which universities are in the top for chemistry, for engineering, and so on. Criteria include research income, number of publications, number of patents, number of Ph.D. examinations, and a peer-based 'reputation survey' (CHE 2005). Also in Germany, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) published a ranking of all German research institutions on the basis of funding allocation. In this DFG study also other performance measures such as international scholars, publications and bibliometric impact indicators were taken into account (DFG 2004).

Rankings have strong 'de-equalizing' effects. This is not new, as all distributions functions related to performance (in economics: income; in bibliometrics: number of publications, citations) are very skew, often of a power-law type. This is a fundamental principle, and it is well described and documented (van Raan 2001; 2005). Rankings, however, directly confront scientists, university board members, policy-makers, politicians, journalists, and the interested man-in-the-street with this inequality. Rankings strengthen the idea of the academic elite, and institutions use the outcomes of rankings, no matter how large the methodological problems are, in their rivalry with other institutions.

As discussed in Chapter 1, the recent ranking publications in fact reinforce the 'evaluation culture'. Evaluations are necessary to account for the received financial support by society. Also academic researchers cannot withdraw anymore from these responsibilities as in earlier times, simply because nowadays the academic enterprise has grown to such an extent that it consumes a considerable part of the public means. Moreover, governments and national research councils are more and more inclined to use the outcomes of evaluations for the distribution of finances over the institutions, and university boards are willing to use the same outcomes for re-distribution of money allocation within their institutions, and even for salary increases of individual researchers.

Ranking positions do already have their effects. Recently, Deutsche Telekom wanted to finance two professorial chairs in German universities. Almost all vice-chancellors of the universities chosen as a candidate for the chairs, put forward their ranking positions as evidence of their research performance. Whether these ranking are based on a sound methodology or not, universities that consider themselves as 'winners', have already begun to cash in on these rankings

(Spiewak 2005). Clearly, the recent rankings shift evaluations to a ‘global’ platform which give them a lot of publicity and, with that, make them much more powerful.

3. Problems in Ranking Procedures

3.1 Reputation versus Contemporaneous Performance

After the publication of the Shanghai ranking a bizarre phenomenon arose: the question to which university Nobel Prize winners belong. Does Albert Einstein ‘belong’ to the Free University (FU) Berlin, or to the Humboldt University (HU) in Berlin? Both Berlin universities contested each other for 'the right' to adorn itself with the pre-1940 Nobel Prize Winners of the then University of Berlin. In the first Shanghai ranking these old Nobel Prizes were attributed to the FU. Later on, in the second, after ‘intervention’ of the HU, they were attributed to the HU, which caused a dramatic backsliding in the ranking of the FU of about 100 positions (Spiewak 2005)! Until now, the Humboldt University and the Free University can not agree on how to divide these ‘old’ Nobel Prizes won by the former Berlin University. This led to the quite dramatic decision of the Shanghai ranking team to leave both universities out of their recent 2005 ranking (SJTU 2005).

This event illustrates an interesting difference between the concepts of ‘reputation’ and ‘contemporaneous performance’. On the one hand, any serious peer review committee responsible for the evaluation of university research would not even think of using long-time-ago Nobel Prize Winners for the judgement of the (nowadays) quality of a university. If they would do, they would disqualify themselves as evaluators (Spiewak 2005). But on the other hand, the history of a university with ‘grand old men’ does count considerably in reputation. Universities publish in their presentations an honorary list of Nobel Prize Winners (if they have) and other famous scientists. Thus, the historical tradition of scientific strengths of particularly the older, classic universities is a strong asset in their present-day reputation. In this sense, the criterion of the Shanghai group for assessing ‘established reputation’ is to the point. The problem, however, is: what do we want to measure, and how does ‘established reputation’ relates to ‘contemporaneous performance’. For the moment we leave this question to further research, but we remark that established reputation is not necessarily the same as ‘past glory’. Often we see that institutions with an established reputation are strong in maintaining their position. They simply have the best possibilities to attract the best people, and this mechanism provides these renowned institutions with a cumulative advantage to further reinforce their research performance.

Another peculiar problem is related to the Nobel Prizes. The affiliation of scientists *at the time of winning* the Nobel prize is an important element in the Shanghai ranking. But where was the winning work done? A scientist may have an (emeritus) position at University A at the time of the award (which is the criterion in the Shanghai study), but the prize-winning work was done (often much earlier) at University B. One of my own teachers, the 1999 physics Nobel Laureate Martin Veltman, is a striking example (A= University of Michigan, Ann Arbor; B = University of Utrecht).

3.2 Can Expert Survey be considered as Peer Review?

‘The greatest influence has been exerted by those in the position to judge: the academics’, says John O’Leary, editor of the Times Higher Education Supplement (THES 2004). Indeed, judgement by knowledgeable colleague-scientists, known as peer review, is the principal procedure of assessing research performance, notwithstanding its shortcomings and disadvantages (Moxham and Anderson 1992; Horrobin 1990). In most cases, peer review is applied on a relatively small scale, ranging from the review of a submitted paper or a research proposal by two or three referees, the review of the record of candidates for a professorship by, say, five experts in the field, and the assessment of research groups and research programs within a specific discipline by between five and ten peers. The above examples illustrate the cases that we consider as typical peer review in science. In these cases the ‘cognitive distance’ between the individual peers and the object to be evaluated (paper, proposal, person, group, program) is small. This implies two important things. First, the peer can be regarded as an expert with respect to the quality of the object. Second, the object to be evaluated has a ‘size’ which is comparable with the normal direct working environment of the peer, namely a research group or a research program and, thus, surveyable for individual peer judgement. In the case, however, of scientists who have to judge (much) larger entities, the cognitive distance to the object to be evaluated increases, and they will have less and less as expertise with respect to all the specific parts of which the larger entity is composed. In the THES rankings and also in the CHE rankings, experts are requested to judge the quality of entire disciplines, ranging from, for instance, biology (in the case of the CHE rankings) to all natural sciences or all biomedical and life sciences (in the case of the THES rankings).

‘…Peer review, the most trusted method for university comparison…’ says Martin Ince (THES 2004). It is however questionable whether all the individual academics involved in such large-scale surveys can be regarded as knowledgeable experts in all those parts of the evaluated entities that really matter. As indicated above, the ‘cognitive distance’ between evaluating person and evaluated object is becoming too large. In such cases, the ‘experts’ will more and more tend to judge on the more general basis of established reputation, instead of their own actual knowledge (if they have!) of recent past performance. The probability that these ‘experts’ -who have to judge the quality of all the life sciences at all major universities in the world- are aware of important, recent breakthroughs in a specific field, decreases dramatically. This awareness, however, is precisely what a peer must have. It is also this recognition of recent past performance that forms the strength of bibliometric analysis. Indeed, bibliometric indicators can be seen as the aggregate of typical peer review. Well-informed colleague-scientists play their role as a member of an ‘invisible peer review college’ by referring (i.e., giving citations) in their own work to earlier work of other scientists. And as this happens for all publications of a university in many disciplines, the outcomes of a bibliometric analysis on the level of a university will be statistically very significant.

There is ample empirical evidence that the judgements of typical, ‘short cognitive distance’ peer review correlate substantially with the results of advanced bibliometric analysis (Rinia *et al* 1998). Both are focused on recent past performance. And recent past performance is the best predictor of future success. Given the above considerations, we expect that the correlation between the outcomes of the THES expert surveys and the outcomes of bibliometric analysis will be very low. In Section 3.4 we will see that this is indeed the case.

Dependence of the outcomes on the choice of experts is one of the major problems. This dependence may cause biases in fields of expertise. After publication of the first THES ranking, The Sydney Morning Herald gloried the presence of not less than six Australian universities in the top 50, while the German news agency DPA bemoaned the absence of the country in the top positions (THES 2004). Six universities in the top-50 universities counts for 12%, whereas Australia contributes for about 2.5 % to the worldwide scientific impact. Germany contributes around 8 % to worldwide scientific impact. Most of these Australian ‘top’ universities score low to very low in citations. These strange discrepancies between the results of an expert survey and bibliometric findings suggest that most probably there are strong geographical biases, particularly an Asian one, in the expert survey of THES. The expert survey of THES was produced by a London-based company specialized in MBA and graduate recruitment. This may very well cause a positive bias for universities with large institutes of economics and schools of management, and a negative bias for universities without a strong emphasis on these disciplines. As THES comments to the (very) low ranking in citation positions of the Australian 'top' universities: 'But the Australian universities are popular in our peer review and do especially well in our rankings of international success. They are among the world's most enthusiastic recruiters of international staff and students, with years of recruiting in Asia and beyond now visibly paying off' (THES 2004).

Next to the above-discussed general problems of expert surveys, there are many methodological questions to be answered. What is the non-response rate? What are the specific characteristics of this non-responding set of experts as compared to the responding set? How is the sample size determined in relation to the reliability of the measurement? What is the standard deviation in the scores for the universities? A large standard deviation is an important indication that the sample size is not large enough. Also the entire distribution function of scores for a specific university is important to understand statistical properties, e.g., are there remarkable outliers in the scores, and how are these related to the chosen sample? What is the statistical significance of a ranking difference of, for instance, 5 positions?

Furthermore, the experts were probably not asked to compare all the universities in the ranking. How many universities could they nominate? Did they receive a worldwide list of universities to make their selection, or did they have to nominate universities 'by heart'? The experts know only a limited part of all universities involved. This put severe limits to the statistical reliability of the expert scores for the entire ensemble of universities. For instance, do all reviewers use the same judgement scales for the same number of universities, and how large is this number? How can it be validated that the differences in scores between peers with their very many different field-specific backgrounds have a reliable meaning? Moreover, for external validation of the experts we need controlling variables such as own institution, geographic distribution, fields of expertise, and own scientific status. So far, all these methodological questions remain, to the best of our knowledge, unanswered. We will discussion a suggestion for a pragmatic improvement of the expert surveys in Chapter 5.

3.3 Pitfalls of bibliometric analysis

3.3.1 Basic assumptions

Bibliometric assessment of research performance is based on one central assumption: scientists, who have to say something important, do publish their findings vigorously in the open, international journal (serial) literature. This assumption introduces unavoidably a 'bibliometrically limited view of a complex reality'. For instance, journal articles are not in all fields the main carrier of scientific knowledge. They are not equivalent elements in the scientific process. They differ widely in importance. And they are challenged as the 'gold standard' by new types of publication behaviour, particularly electronic publishing. However, the daily practice of scientific research shows that inspired scientists in most cases, and particularly in the natural sciences and medical research fields, 'go' for publication in the better and -if possible- the best journals. A similar situation is developing in the social and behavioural sciences (Glänzel 1996; Hicks 1999), engineering and, to a lesser extent, in the humanities. We discuss the limitations and possibilities of bibliometric analysis for the social sciences and the humanities in Section 4.1.

A first and good indication whether bibliometric analysis is applicable to a specific field is provided by the publication characteristics of the field, in particular the role of international, refereed journals. If international journals are the dominating or at least a major means of communication in a field, then in most cases bibliometric analysis is applicable. Therefore it is important to study first the ‘publication practices’ of a research group, department, or institute, in order to establish whether bibliometric analysis can be applied. A practical measure here is the share of CI-covered publications in the total research output.

Work of at least some importance provokes reactions of colleague-scientists. Often, these colleague-scientists play their role as a member of the invisible college by referring (i.e., giving citations) in their own work to earlier work of other scientists. Thus, citation analysis is based on reference practices of scientists. The motives for giving (or not giving) a reference to a particular article may vary considerably (Brooks 1986; MacRoberts and MacRoberts 1988; Vinkler 1998). So undoubtedly the process of citation is a complex one, and it certainly not provides an ideal monitor on scientific performance (MacRoberts and MacRoberts 1996). This is particularly the case at a statistically low aggregation level, e.g., the individual researcher. There is, however, sufficient evidence that these reference motives are not so different or randomly given to such an extent that the phenomenon of citation would lose its role as a reliable measure of impact (van Raan 1998). Therefore, application of citation analysis to the entire work, the oeuvre of a group of researchers as a whole over a longer period of time, does yield in many situations a reliable indicator of scientific performance.

3.3.2 Technical problems

The most central technical process, on which citation analysis is based, is the matching of *citing* publications with *cited* publications. In a publication (the ‘citing publication’) a reference is given to another publication (the ‘cited publication’), and this reference has to be identified as -an earlier- ‘source publication’ in the citation indexes. A wide variety of errors may occur in this citing-cited matching process leading to a ‘loss’ of citations to a specific publication. In average, the number of non-matching references -although they are citation index covered source papers- is about 7% of the citations matched. Frequently occurring non-matching problems relate to publications written by consortia (large groups of authors), to variations and errors in author names particularly -but not only- authors from non-English speaking countries, errors in journal volume numbers, errors in initial page numbers, discrepancies due to journals with dual volume-numbering systems or combined volumes, or to journals applying different article numbering systems. Thus, these non-matching citations are highly unevenly distributed in specific situations, which may cause an increase of the percentage of lost citations up to 30% (Moed 2002). So if the citation indexes are used for *evaluation purposes*, all these possible errors have to be corrected as much as possible.

The second major technical problem relates to the attribution of publications -and with that, of the citations to these publications- to specific organizations such as institutes, university departments, and even on a high aggregation level to the main organization, for instance universities. It is often thought the citation indexes can simply be scanned in order to find ‘all’ publications of University X. This assumption is based on the argument that all these publications mention somewhere in the address data of the publication clearly 'University X' as the main affiliation of the authors. But this assumption is wrong.

Next to *variations* in the name and abbreviations of the same university, departments and institutes (in many variations) are to a non-negligible extent mentioned *without proper indication* of the university. For instance, in the commercial database on highly cited scientists we find five variants for (parts of) Leiden University: Leiden University, Universiteit Leiden, Leiden Observatory, Leiden University Medical Center, Leids Universitair Medisch Centrum. Furthermore, groups or institutes of a national research organization (such as the French CNRS) are quite often mentioned instead of the university where the research actually takes place. For instance, if a research group uses as an address only ‘CNRS, Illkirch’, it is very probably a CNRS-financed research group working at the University of Strasbourg 1 (or: Université Louis Pasteur; Illkirch is a suburb of Strasbourg). Similar problems occur for graduate schools. Even the name of a city will not always be sufficient to identify an institution. Parts of universities may be located in suburbs. The change of spelling of Chinese city names may cause considerable difficulties in identifying the same university. If a university is not properly *unified*, it may loose a substantial part of its citation score.

There are major differences in research systems between countries affecting the definition of a university. For instance, the University of London is not a university in the usual sense. It is an ‘umbrella organization’ covering several different virtually autonomous universities. Similar problem occur with multi-campus university systems in the United States, for instance the University of California. In France we deal with autonomous universities in the same city that were part of originally one ‘mother-university’. As a consequence, it is very cumbersome to distinguish between departments of these different universities within one and the same city. In some cases these problems are so large (e.g., Vrije Universiteit Brussel and the Université Libre de Bruxelles, both are indexed as 'Free University (of) Brussels') that it is impossible to distinguish both universities on CI-based address data only. Similar problems occur for Leuven and Louvain-la-Neuve. Next to splitting also merging of institutions may occur. Last but not least, various types of address errors appear in the CI database.

Very problematic is capturing the *medical research* of universities, as often only the medical school, and/or the name of the hospital without mentioning the university is indicated. One explicitly needs the names of one or more hospitals in a specific city -and also in the suburbs of a city- that are in fact university hospitals. So if a hospital clearly has the role of a university hospital (because it is known as such and it is indicated as such in, for instance, the website of the university’s medical faculty), but an article mentions only (for instance) ‘Radcliffe Hospital’, this article must be added to the ‘broad definition’ of the university (in this example: Oxford). Otherwise a university would miss a major part of its basic and, particularly, clinical medical research, as compared to universities where the hospital has the word ‘University’ in its name. It is clear that this is not always a simple procedure, as university-related hospitals are often autonomous organizations that may have relations with other research institutes or other institutes of higher education (for instance with polytechnics, in the case of health care). Again, large efforts in cleaning and re-building of the original citation indexes are necessary to solve this problem by proper definition and unification of a specific institution. These unifications has to be done as good as possible, otherwise universities with medical research cannot be compared anymore in a reliable way. Based on our experiences in dozens of research performance assessments of scientific institutions conducted during the past two decades, we summarize our above observations with five main statements concerning address and affiliation problems in the text box below.

1. Data on institutional author affiliations of CI-covered articles are often found to be incomplete and/or inaccurate. Not all articles list institutional addresses of their authors. Names of many organizations may appear in large numbers of variations.
2. The CI-database producer reformats and unifies (de-duplicates) institutional addresses to some extent. This is particularly true for US institutions. In these internal procedures errors are made. A typical example is the conversion of two distinct Belgian Universities into one single entry ‘Free University Brussels’, so that publications from these two institutions cannot be separated merely on the basis of addresses appearing in the CI. Such errors cannot be easily detected.
3. It is not always clear how an institution must be properly defined. Particularly the role of ‘affiliated’ institutes or umbrella/parent organizations is problematic. For instance, in some countries ‘academic’ hospitals are a part of the parent university, whereas in other countries they are separate entities.

4. Authors do not always give full information on their affiliations. Authors from research institutes that operate under the umbrella of a parent research council, in some case give only the name of the institute, and in other cases only the name of parent organization, and sometimes both names. A similar situation occurs for authors in research institutes of a national research council located in a university, a common phenomenon in for instance France and Italy.
5. It is in numerous cases extremely difficult to capture all variations under which an institution's name may appear in addresses in scientific publications. Although for instance our institute, CWTS, has put a large effort in unifying institution names, it cannot be claimed that our unification is free of error, nor do we include the most recent name variations and acronyms.

The increasing use of bibliometric data in evaluation procedures and particularly in rankings underlines the vital importance of a clear, coherent and effective presentation to the outside world of universities in their publications. For instance, King's College, University of London (KCL), introduced a code of practice to ensure that all publications are properly attributed to the College. This is in light of recent evidence that up to 25% of citations from KCL academics in the last two years were missed due to failure to use 'King's College London' in the address mentioned in the publication heading (Wheeler 2005).

3.3.3 Methodological problems

Methodology directly relates to the aims of a study. The objective of both the Shanghai as well as the THES study is to obtain a worldwide ranking of universities in terms of their 'scientific strength'. The crucial point is then, with which indicator, or weighted combination of indicators, such rankings have to be constructed.

An important bibliometric ranking element in the Shanghai study is the number of highly cited researchers in a university for the life sciences, medicine, physical sciences, engineering and social sciences. The Shanghai group used the Thomson Scientific Highly Cited Scientists database. These individuals are the 250 most highly cited researchers within each of the 21 broad fields of science[7] for the period of 1981-1999. The Shanghai group here depends completely on how the database producer identifies the highly cited scientists and calculates citation rates. Methodologically this means the use of a 'black box'. Problems with the identification of individual scientists by their names are even larger than the problem discussed above with the identification and definition of institutions. Moreover, in our opinion the 21 'broad fields, are too large. Publication and citation characteristics vary substantially between fields of science within these broad fields. Because of these differences, proper field-specific normalization is necessary, which is not done in the highly cited scientists database. Field-normalization must also take the size of the field into account. The Thomson Scientific broad fields differ considerably in size (e.g., immunology covering about 70 journals, versus molecular biology and genetics covering about 150 journals!).

Furthermore, the time period 1981-1999 does not reflect the state-of-the-art of the research front at present. These highly cited scientists are determined by 'life time citation counts', which enhances the 'old boys effect' and does not specifically focus on the impact at the research front of today, particularly by younger researchers. As one of the 14 highly cited scientists of our university told me: 'I really should not be anymore on this list'. Thus, we reject the use of *highly cited scientists* - and only those who happen to be covered in a specific commercial database. We discuss in Chapter 4 an alternative approach in which *all highly cited papers* of a university are used for benchmarking of international scientific excellence (Tijssen *et al* 2002).

[7] Broad fields are defined as combinations of smaller fields, and these latter are defined as sets of journals, called 'journal categories'. The coverage of these journal categories can be found in the Thomson Scientific Web of Science website. Although this classification is certainly not perfect, it is at present the only classification that properly fits the multidisciplinary character of the Citation Indexes.

In the identification of highly cited papers, it is important to distinguish between the various article types: normal articles, letters, notes, and reviews. There are large differences between reviews and normal articles both on the 'citing side' (the number of references, for reviews this number is usually high to very high) and on the 'cited side' (the number of citations received, for reviews in many cases again high). Therefore, review papers may constitute a considerable part in the collection of highly cited scientists. However, reviews are in most cases not original scientific work, the authors present state-of-the-art overviews of developments in their field. So it is necessary to take article type into account in all normalization procedures (i.e., comparison of reviews with reviews, etc.) in the calculation of impact-indicators used for the ranking procedure.

Definition of fields on the basis of international journals and application of citation data depends on the role of journal articles in the different fields and the coverage of these journals by the citation index system. This coverage is particularly problematic for engineering, social and behavioural sciences, and certainly for the humanities. Thus, the strength of a university in engineering, in the social and the behavioural sciences or in the humanities may contribute little -or even hardly- to the position of that university in a ranking based on bibliometric data. For a further discussion we refer to Section 4.1.

A next point concerns the time dimension in the indicator framework. This relates to the earlier discussed controversy between established reputation and contemporaneous performance. In the Shanghai 2005-study different time horizons are used for the various indicators: back to 1910 for Nobel Prizes, the period 1981-2003 for the highly cited scientists, the period 2000-2004 for the number of Nature and Science papers, and the year 2004 for articles covered in the SCI-expanded. It still need to be explained how these very different time horizons contributes to the notion of 'who belongs now to the top', which is undoubtedly the question that drives the users of such rankings.

A further important methodological problem is the influence of biases in the citation index system. In rankings based on bibliometric indicators, US universities and research institutions tend to dominate the top-positions of the ranking. No doubt that the US top-universities are institutions of undisputed world-class. There is, however, also the effect of US dominance in the overall publication- and citation traffic. It is not easy to find out to what extent this phenomenon affects impact assessments, and to correct accurately for it. More methodological work is necessary to come to grips with this problem. Before this problem is solved, we have to be careful with the interpretation of worldwide rankings as they suggest that, for instance, Europe is scientifically far beyond the US. This may be true in a number of situations, but certainly not to the extent as is suggested purely on the basis of bibliometric rankings.

A more concrete bias is related to publication language. Recent work (Grupp *et al* 2001; van Leeuwen *et al* 2001) shows that the utmost care must be taken in interpreting bibliometric data in a comparative evaluation of national research systems (May 1997). The measured value of impact indicators of research activities at the level of an institution and even of a country strongly depends upon whether one includes or excludes publications in CI-covered journals written in languages other than English, particularly French and German. Generally the impact of publications of these French and German-language journals is very low. Thus, in the calculation of impact indicators, these publications count on the output side, but they contribute very little, if any, on the impact side. Therefore, such non-English language publications considerably 'dilute' the measured impact of a university or a department. We have empirical evidences (van Leeuwen *et al* 2001) that the use of German-language journals covered by the citation indexes may lead to about 25% lower measured impact. Simply by removing the publications in these German-language journals and only using the English-language journals (which is fair in an international comparison and certainly in a comparison with, for instance, the United States and the UK), the measured impact will 'improve' with this 25% for a whole medical faculty of a university! No doubt that there will be the same effect for French-language journals. It is clear that such a 25% difference may substantially affect ranking positions. Therefore, in less advanced bibliometric analyses, particularly Germany and France will 'suffer'. These findings clearly illustrate again that indicators need to be interpreted against the background of their inherent limitations such as, in this case, effects of publication language, even at the 'macro-level' of entire countries, but certainly at the level of institutions.

On the basis of exactly the same data and exactly the same technical and methodological starting points, still different types of impact-indicators can be constructed, for instance one focusing entirely on normalized impact, and another in which also scale (size of the institution) is taken in to account. Rankings based on these different indicators are not the same, as proven by a recent CWTS study for the European Commission (Van Raan and Van Leeuwen 2001).

3.4 Correlation between peer review results and bibliometric findings

In earlier studies comparisons are made between bibliometric results and the judgement of scholars or experts on the quality of research (Anderson *et al* 1978; Bayer and Fulger 1966; Chang 1975; Cole and Cole 1967; Cole *et al* 1978; Martin and Irvine 1983; Nederhof 1988; Nederhof and Van Raan 1987, 1989). These studies revealed a reasonable to significant correspondence between the results of bibliometric analyses on the one hand, and judgements of scientific quality by peers on the other. It is important to find general patterns in cases where bibliometric results do correspond, and where -and why- they do not correspond with judgment by peers. For instance, a poor correlation was found between bibliometric indicators and the originality of research proposals in application-oriented research judged by peers (Van den Beemt and Van Raan 1995).

Studies of larger-scale evaluation procedures in which empirical material is available with data on both peer judgment as well as bibliometric indicators are quite rare. Recently, detailed empirical results on the correlation between peer judgment and indicators have become available. Rinia *et al* (1998) found a significant correlation between peer judgment and indicator values for 56 research programs in condensed matter physics in the Netherlands. This study covers 10 years of about one third of the total volume of physics in the Netherlands, with 5,000 publications and 50,000 citations to these publications. More recently, we investigated the correlation between peer judgments and indicators values for all university research groups in chemistry and chemical engineering in the Netherlands. The time period covered is 1991-2000. In total, the analysis covers 18,000 publications and 240,000 citations of 147 chemistry groups (VSNU 2002). Also in this chemistry and chemical engineering study we find a significant correlation between the outcomes of peer review and bibliometric analysis.

Thus, if we find such a correlation for two major disciplines within the natural sciences, we can be sure that for all natural science and also for all life sciences as a whole, similar correlations will appear if we have a sufficiently large group of peers that is reasonably well spread over the different disciplines. In Figure 1 we show the results of the correlation analysis between expert scores and citation-analysis based scores in the THES ranking (THES 2004).

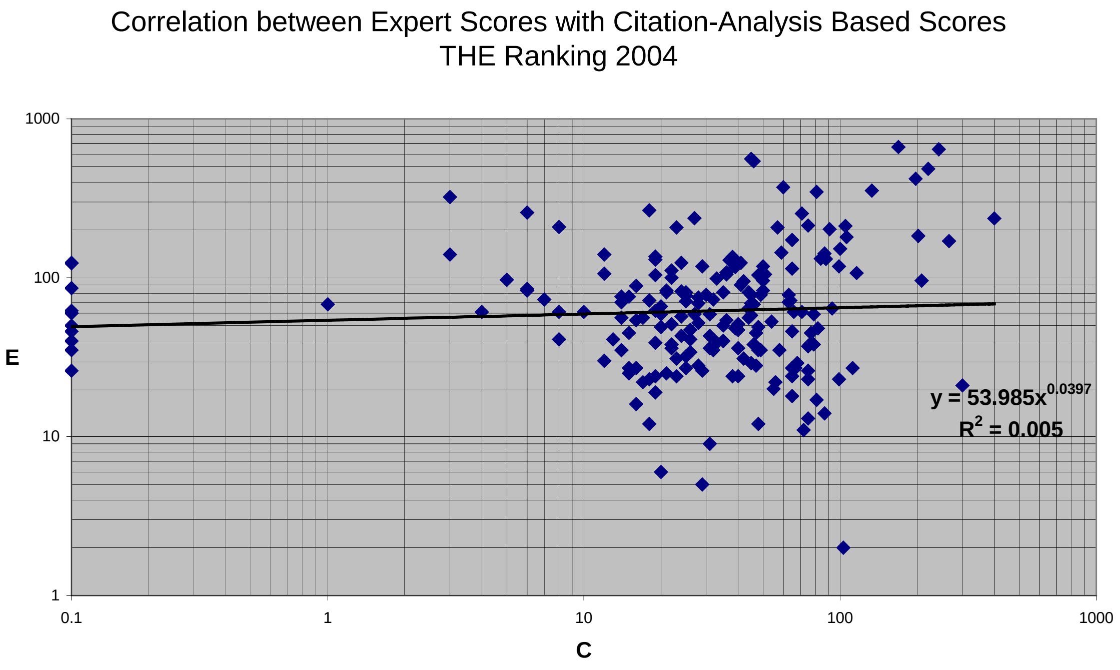

*Fig. 1: Correlation between expert scores (**E**) and citation-analysis based scores (**C**) in the THES 2004 ranking.*

We immediately observe the absence of any correlation between the judgement of the experts involved in the THES ranking study and the citation-analysis based results (coefficient of determination of the fitted regression $R^2 = 0.005$!). This result is sufficient to seriously doubt the value of the THES ranking study.

Finally, we remark that recently a significant correlation was found between the *funding* by the German Research Foundation (the funding ranking study mentioned in Chapter 2) and bibliometrically measured impact of German universities (DFG 2004).

4. New Bibliometric Approaches

4.1 Specific remarks concerning the social sciences and the humanities

Before presenting new approaches based on advanced bibliometric methods, we first discuss a number of important issues concerning the application of quantitative evaluation methods in the social sciences and humanities. In many parts of the social sciences and humanities, scientific publication practices are less standardized than they are in the natural and life (including biomedical) sciences. Particularly, the role of international peer-reviewed journals is less important. English is not always a dominant language, and journals may even be multi-lingual. Generally, the social sciences disciplines, in which the structure of the communication system is similar to that in the natural and life sciences, such as psychology and economy, are rather well covered by the Citation Index system.

Although many databases are publicly available that cover scholarly output in disciplines of the social and behavioural sciences and the humanities, these databases have several properties that make them less suitable or even useless for calculating bibliometric indicators:

- It may be unclear which bibliographic *sources* are processed; criteria for selection of sources may be unclear;
- Databases may have national or other *geographical biases*;
- Large fractions of processed documents do *not* have institutional *affiliations* of publishing authors;
- Even if documents do contain addresses of publishing authors, the database producer may *not include* these in the database;
- Important data elements - even journal titles and country names - may not be *standardized*;
- To the best of our knowledge none of the major databases include cited references; this fact makes these databases *completely useless for impact measurements* and for all other citation-related analyses;
- Many databases are available only through host computers that offer only limited counting and other statistical *facilities*. Their use (on-line data retrieval) may be expensive, and particularly if data are needed on a large-scale, the costs may be exceptionally high.

Given these limitations, the following approach provides a possibility to get at least a first indication of output size of institutions in the social sciences and the humanities. First, for all social science disciplines the adequacy of coverage of the Citation Index has to be examined. Preliminary empirical studies conducted at CWTS suggest that the CI-coverage of psychology and economics is probably reasonably good. Other databases than the Citation Index have to be explored (for instance: ECONLIT, Psychological Abstracts, and Sociological Abstracts). One could start with disciplines with a low CI-coverage, for instance language and linguistics, law, political science, sociology, and educational sciences. Next, a specification of the types of publications that must be taken into account in the calculation of publication counts is necessary. This also depends upon the classification system applied by the various databases producers. As a general principle, however, it can be stated that monographs and multi-authored books are important sources of written communication in many disciplines of the social sciences and the humanities. Therefore, they should not be omitted in a comprehensive assessment of publication output. From the sixth point mentioned above it is clear that these non-CI databases cannot be used for bibliometric impact analysis, i.e., citation analysis.

4.2 Institutional definition and unification of universities

In Section 3.3.2 we discussed the technical problems concerning addresses and affiliations. The current version of the CWTS bibliometric data-system contains unified names of the vast majority of major public research institutions worldwide as well as many of the large science-based/R&D-intensive companies. However, in most cases this unification deals only with the main organizational level. Furthermore, representatives of the main organizations involved do not yet systematically check our unification results. Our experiences lead to the conclusion that a reliable, useful bibliometric assessment of a research institute cannot be merely based on an analysis of addresses *as given in* the publications. An appropriate identification scheme of an institution’s publication output must involve detailed background knowledge provided and/or thoroughly checked by the institutions themselves. We are convinced that verification by the representatives of institutions is indispensable for obtaining outcomes that are sufficiently accurate for evaluation studies. This approach will enable representatives of institutions to indicate what they believe are variants of institutional names that correspond to their organization. Thus, they may highlight particular pitfalls, for instance special characteristics related to publication practices of their research workers, identification of their research papers, or the institutional definition of their organization. We are currently working on these affiliation-related problems in an extensive project supported by the European Commission[8] aiming at the measurement of the performance of European universities and expect the first results by the end of this year.

4.3 Top-10% approach

As soon as reliable definitions of institutions, based on procedures described in the foregoing section are available, it is possible to collect all relevant publications of an institution. The next step is to apply advanced bibliometric indicator algorithms to the publication sets of all institutions involved.

Citation analysis is a crucial part the bibliometric methodology. It is more than just counting. Valid indicators have to be constructed, including field-specific normalization, because publication and citation habits in the many fields of science can be very different. In the Appendix we give a short presentation of the main elements of the bibliometric methodology developed by CWTS and an overview of the resulting standard indicators (van Raan 1996, 2000). We regard the *internationally field-normalized impact indicator* ***CPP/FCSm*** (see Appendix, Section A3) as our 'crown' indicator. This indicator enables us to observe immediately whether the performance of a research group, institute or university is significantly far below (indicator value < 0.5), below (between 0.5 and 0.8), around (between 0.8 and 1.2), above (between 1.2 and 1.5), or far above (>1.5) the international impact standard related to all fields involved. We stress, however, that for the interpretation of the measured impact value one has to take into account the *aggregation level of the entity* under study. The higher the aggregation level, the larger the volume of publications and the more difficult it is to have an impact significantly above the international level. Based on our long-standing experiences, we can say the following. At the level of an entire institution (e.g., a university) with about 500 or more publications per year, a ***CPP/FCSm*** value above 1.2 means that the institution's impact as a whole is significantly above the international average (composed of the weighted averages of all field involved, see Appendix). With a ***CPP/FCSm*** value above 1.5, the institution can be considered to be scientifically strong, with a high probability of finding very good to excellent groups[9].

In this paper we focus on a more specific indicator of *scientific excellence:* the number of publications that are *cited within the top 10%* of the worldwide impact distribution of each of the fields within an institution (see Appendix, for further details we refer to Noyons *et al* 2003). In the calculation of this indicator the entire, field-specific citation distribution function has to be taken into account, thus providing a better statistical measure than those based on mean values.

Standard statistical techniques relate to quantities that are distributed approximately 'normally'. Many characteristics of research performance, particularly those based on citation analysis, are not normally, but very skewly distributed. Thus statistical averages can be misleading. For larger samples, such as the entire oeuvre of a research group over a period of years, and certainly for entire universities, the central limit theorem says that whatever the underlying distribution of a set of independent variables (provided that their variance is finite), the sum or average of a relatively large number of these variables will be a random variable with a distribution close to normal. On the basis of these considerations we

[8] The ASSIST project (Analysis & Studies and Statistics & Indicators on Science and Technology) is part of the European Commission (EC) Sixth Framework Programme for Research, Technology and Development.

[9] At the *research group level* a ***CPP/FCSm*** value above 2 indicates a very strong group, and above 3 the research group can be, generally, considered to be excellent and comparable to the top groups at the best US universities. If the threshold value for the ***CPP/FCSm*** indicator is set at 3.0, excellent groups can be identified (van Raan 2000).

are confident that, for instance, our ‘crown indicator, the field-normalized international impact indicator ***CPP/FCSm*** (see Appendix) does provide a useful measure. This can be demonstrated by the strong correlation of ***CPP/FCSm*** and the ‘top 10%’ indicator in which the distribution function is taken into account (Noyons *et al* 2003).

A ranking analysis based on the top-10% approach can be executed in the following two steps (van Raan and Van Leeuwen 2001):

(1) Identification, on the basis of a bibliometric data-system with an as good as possible definition and unification of institutions, of the 250 most active (in terms of publications) institutions worldwide (which implies institutions with number of publications ***P*** about 500 and more per year);

(2) Collection of all publications (in all fields relevant to the intended ranking procedure) of these 250 institutions, and all citations (corrected for self-citations) to these publications;

(3) Ranking of these 250 institutions in the following three modalities:
(a) by impact indicators for the *entire oeuvre* (number of publications ***P***) of the institutions, particularly by ***CPP***, and by the field-normalized impact indicator ***CPP/FCSm*** (thus, these two different impact indicators provide sub-modalities);
(b) by the same impact indicators, but now exclusively for the *top-10% part of the entire oeuvre*, i.e., the publications within the *top-10%* (for each of the fields within an institution, ***P(top)*** in total), ***CPP(top)*** and the field-normalized impact indicator ***CPP/FCSm(top)*** (again the two sub-modalities);
(c) by the field-normalized impact indicator ***CPP/FCSm(top)*** (i.e., only for the publications within the *top-10%* of each of the fields within an institution) multiplied by this number of publications: ***CPP/FCSm(top)*** * ***P(top)***.

Ranking *(a)* provides a benchmarking of all 250 universities based on the average impact and the average field-normalized impact, respectively, of their entire oeuvre. The indicators are normalized for size (in terms of numbers of publications produced by the institutions) as well as for field of science. Thus, smaller institutions do not ‘suffer’ in this ranking, neither do institutions that are characterized by fields of science with a relatively low absolute citation level. In ranking *(b)* the performance assessment focuses exclusively on the high-impact part (the top-10% of the distribution), with the same type of indicators. Hence, scientific excellence is the central aspect in this ranking. Finally, ranking *(c)*, also based on the top-10% field-normalized performance, provides a ‘brute force’ benchmarking by removing the size-normalization (multiplication with number of publications). These three modalities (with sub-modalities for the first two) offer the possibility to rank *on the basis of the same methodological approach and the same collection of data* according to different indicators that focus on different aspects of international scientific impact, thus allowing an effective robustness check of the rankings.

The final research question formulated in Chapter 1 relates to an important element in the debates on ranking of universities and on ‘leagues’ of outstanding universities in general: how many members can or should be covered by a specific league? In other words, at what position of the ranking should the league be ‘closed’? In order to find an answer to this question based on empirical grounds, we analysed statistical properties of our earlier ranking exercise on worldwide top-universities in the life (biomedical) sciences (van Raan and Van Leeuwen 2001). Particularly, we investigated the relation between impact of the institutions (indicator ***CPP***) and ranking position (***r***). The (preliminary) results of this analysis are shown in Figure 2. We observe a quite surprising phenomenon: it appears clearly that the group of outstanding universities will not be larger than around 200 members.

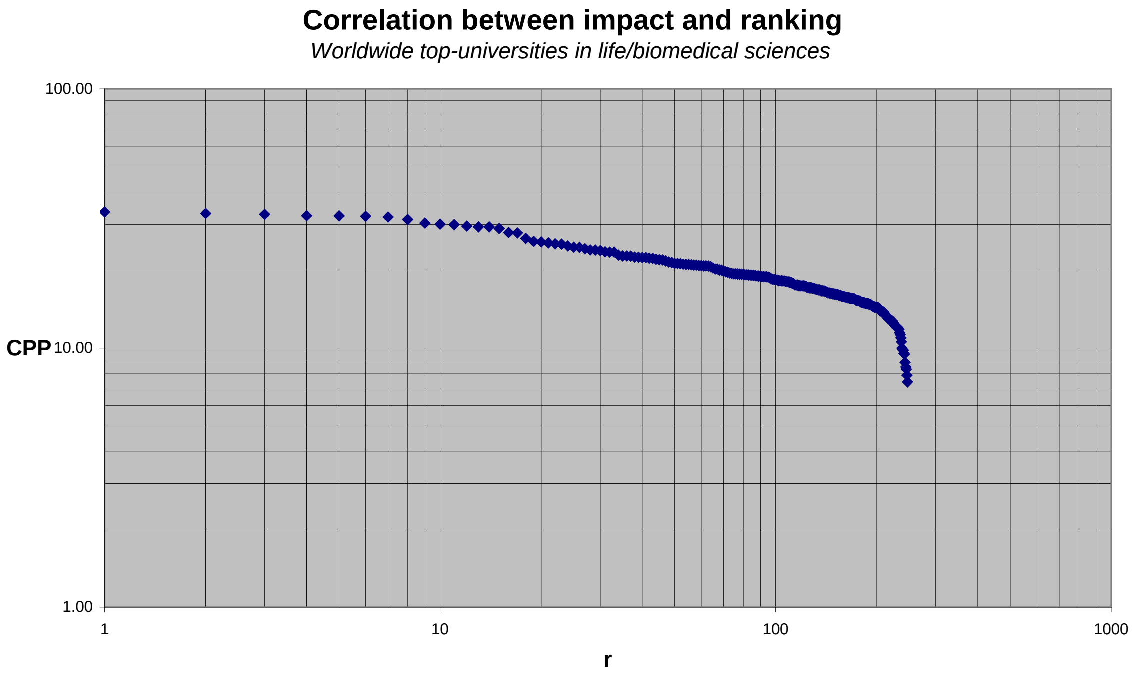

*Fig. 2: Correlation between impact of top-universities in the life and biomedical sciences (**CPP**) and ranking position (**r**).*

Our explanation of this phenomenon is simple and based on 'finite size' considerations. Most of the top-universities are large, broad research universities. They have attracted on the basis of their reputation for already a long time the best students and scientists. These universities are the 'natural attractors' in the world of science, and, apparently, around 200 of these institutions are able to acquire and hold on the vast majority of top-scientists. After ranking position 200 or so, there will certainly be smaller universities with excellent research in specific field of science. There is, however, no more room for further 'power houses of science' because no more excellent scientists are available worldwide.

5. Further Observations and Outlook

We discussed in our paper that both expert-based analysis as well as bibliometric analysis provide us, in principle, with possibilities to benchmark universities in a worldwide perspective. However, we showed that the reliability of simple 'expert-based' rankings is very doubtful. From our discussions it is clear that we need further research on several important themes such as evaluation methodology and, particularly, statistical methods to compare and to rank. Next, we need to study the influence of ranking on institutional management and public policy, and implications for developing countries. In this final chapter we discuss a few selected themes that are in our opinion crucial for substantial progress in evaluation and ranking procedures.

Basically, the methodological problem of determining the quality of a subject is still far from solved, as illustrated by the results of re-review of previously granted research proposals, see Nederhof (1988) and the studies mentioned by Weingart (2004). *Re-iteration* of the expert-survey with renowned scientists from the first-round 200 top-universities is a first, necessary step in the direction of judgment by highly qualified reviewers. It would be very interesting to see what differences will emerge from such a first-iteration expert round as compared to the original ranking and, by that, how robust the first ranking is. We expect that already in the first re-iteration the expert judgments will converge to the citation-based impact measures and that after, say, two re-iterations, there will be a high correlation between expert-based and citation-based impact.

Next, we have a fundamental point concerning the availability of correct addresses. Bibliometric analysis for evaluation purposes has to be more than just using what is readily available in databases. This is a crucial element of advanced bibliometric research, development and practical application. Scientists write their publications for communication and knowledge dissemination. They are not responsible for the use of the their publications for a completely different goal, namely evaluation of scientific performance. So ‘bibliometric evaluators’ have their own, *specific responsibility* that is different from and beyond that of an individual author, particularly as far the attribution of publications concerns. Precisely this attribution is one of the most crucial elements in the whole evaluation process! This is a matter of ethical conduct as an evaluator.

In the political debates on evaluations and rankings often the basic question is: what are the characteristics of successful universities? There must be ‘sufficiently enough’ quality, but what does that mean, and how can it be measured? On the basis of our long-standing experiences in bibliometric research performance analysis, we have an answer that is clear and can be well operationalized. Successful universities are those universities that succeed to perform significantly above the international average in more than half of their first, say, 20 largest (internationally standardized) fields, and in less then 10 % of these fields significantly below international average. To identify successful universities, standardized research profiles of universities have to be constructed along the methodological lines discussed in the Appendix. For a further discussion we refer to Van Raan (2004) where also an example of a research profile is given. In this paper, we present the research profile of Leiden University as an example, see Fig. 3 and the textbox below this figure.

In addition to research performance, various facilitating factors (important examples: library, housing for international guest researchers and students, quality of city life) are crucial to become and to remain a ‘world class’ university.

Finally, we plead for evaluation procedures in which the involvement of the research units to be evaluated will be limited to the absolutely necessary. Thorough evaluations make great demands on the time of scientists involved in these evaluations, both on the side of the evaluators as well as on the side of those to be evaluated. An intensification of research performance assessment procedures may very well damage the academic potential, as they will destroy the already scarcely available rest for thinking and developing new ideas. Some scientists even foresee a period of ‘science destruction’ if the evaluation-hype continues to explode. Advanced bibliometric methodology provides the opportunities to carry out effective evaluations with a low burden for the objects of the evaluation. One of the ‘products’ of advanced bibliometric methodology is an extensive, high-quality worldwide benchmarking of universities by ranking.

Acknowledgements

The author would like to thank his CWTS colleagues Henk Moed, Robert Tijssen en Thed van Leeuwen for discussions on evaluation and ranking methodology.

Research profile

Output and impact per field

2000 - 2003

Leiden University

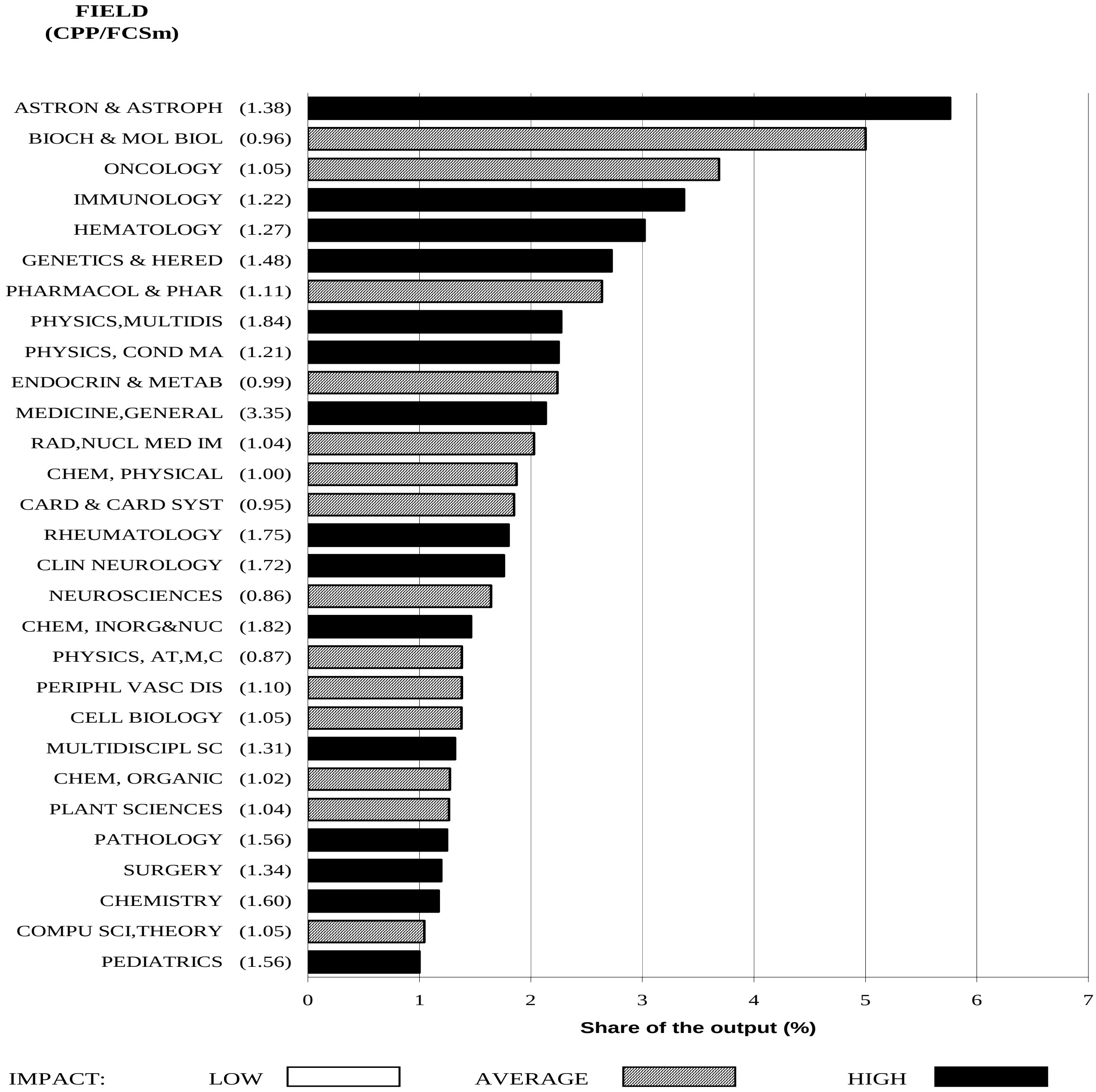

Fig. 3: Research profile of Leiden University, 2000-2003

A research profile is the *breakdown* of the research output of a university into research fields. The profile is constructed by ranking fields according to size (in terms of numbers of publications) in a graphical display. The university has a performance significantly above the international average in fields with black bars (‘high’, indicator value > 1.2), about international average in fields with shaded bars (‘average’, value between 0.8 and 1.2), and significantly below international average in fields with white bars (‘low’, value < 0.8). Field-normalized impact values (indicator ***CPP/FCSm***) are given directly behind the name of the fields. The length of the bars represents the relative output (in % of the total number of publications of the university) for the indicated field.

References

Anderson, R.C., Narin, F. and P. McAllister. (1978). Publication ratings versus peer ratings of universities. *Journal of the American Society for Information Science* 29, 91-103

Bayer, A.E. and F.J. Fulger. (1966). Some correlates of a citation measure of productivity in science. *Sociol. Educ.* 339, 381-390.

van den Beemt, F.C.H.D. and A.F.J. van Raan. (1995). Evaluating research proposals. *Nature* 375, 272.

Brooks, T.A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science* 37, 34-36.

Chang, K.H. (1975). *Evaluation and survey of a subfield of physics: magnetic resonance and relaxation studies in The Netherlands,* FOM-Report No. 37175, Utrecht.

Cole, S. and J.R. Cole. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review 32,* 377-390.

Cole, S., Rubin, L., and J.R. Cole. (1978). *Peer Review in the National Science Foundation*. National Academy of Sciences, Washington, DC.

CHE (2005). Das CHE Forschungsranking deutscher Universitäten 2004. Centrum für Hochschulentwicklung, http://www.che.de/index.php, retrieved May 2005.

DFG (2004). *Funding Ranking 2003: Institutions-Regions-Networks. DFG Approvals and Other Basic Data on Publicly Funded Research*. Bonn: Deutsche Forschungsgemeinschaft. Online version available at http://www.dfg.de/en/ranking/index.html.

Glänzel, W. (1996). A bibliometric approach to social sciences, national research performances in 6 selected social science areas, 1990-1992. *Scientometrics* 35, 291-307.

Grupp, H., U. Schmoch, and S. Hinze (2001). International alignment and scientific regard as macro-indicators for international comparisons of publications. *Scientometrics* 51, 359-380

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics* 44, 193-215.

Horrobin, D.F. (1990). The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association* (*JAMA*) 263, 1438-1441.

van Leeuwen, Th.N., H.F. Moed, R.J.W. Tijssen, M.S. Visser, and A.F.J. van Raan (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics* 51, 335-346.

MacRoberts, M.H. and B.R. MacRoberts (1988). Author motivation for not giving citing influences- a methodological note. *Journal of the American Society for Information Science* 39, 432-433.

MacRoberts, M.H. and B.R. MacRoberts (1996). Problems of citation analysis. *Scientometrics* 36, 435-444.

Martin, B.R. and Irvine, J. (1983). Assessing basic research. Some partial indicators of scientific progress in radio astronomy. *Research Policy* 12, 61-90.

May, R.M. (1997). The scientific wealth of nations. *Science* 275, 793-796.

Moed, H.F. (2002). The impact factors debate: the ISI's uses and limits. *Nature* 415, 731-732.

Moxham, H. and J. Anderson (1992). Peer review. A view from the inside. *Science and Technology Policy* 7-15.

Nederhof, A.J. and van Raan, A.F.J. (1987). Peer review and bibliometric indicators of scientific performance: A comparison of cum laude doctorates with ordinary doctorates in physics. *Scientometrics* 11, 333-350.

Nederhof, A.J. (1988). *The validity and reliability of evaluation of scholarly performance*. In: Van Raan, A.F.J. (ed.) (1988). Handbook of Quantitative Studies of Science and Technology. Amsterdam: Elsevier/North-Holland, pp.193-228 (ISBN 0-444-70537-6).

Nederhof, A.J. and van Raan, A.F.J. (1989). A validation study of bibliometric indicators: The comparative performance of cum laude doctorates in chemistry. *Scientometrics* 17, 427-435.

Noyons, E.C.M., R.K. Buter, A.F.J. van Raan, U. Schmoch, T. Heinze, S. Hinze, and R. Rangnow (2003). *Mapping excellence in science and technology across Europe* (Part 1: *Life sciences*, Part 2: *Nanoscience and nanotechnology*). Report to the European Commission by the Centre for Science and Technology Studies (CWTS), Leiden University, and the Fraunhofer Institute for Systems and Innovation Research (Fraunhofer-ISI), Karlsruhe.

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. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics* 43, 129-139.

van Raan, A.F.J. (2000). The Pandora's Box of Citation Analysis: Measuring Scientific Excellence, the Last Evil? In: B. Cronin and H. Barsky Atkins (eds.). *The Web of Knowledge. A Festschrift in Honor of Eugene Garfield*, Ch. 15, p. 301-319. Medford (New Jersey): ASIS Monograph Series, 2000. ISBN 1-57387-099-4.

van Raan, A.F.J. (2001). Two-step competition process leads to quasi power-law income distributions. Application to scientific publication and citation distributions. *Physica A* 298, 530-536.

van Raan, A.F.J. and Th.N. van Leeuwen (2001). *Identifying the fields for mapping RTD excellence in the life sciences -a first approach.* Report to the European Commission, Brussels; contract nr. COPO-CT-2001-00001.

van Raan, A.F.J. (2004). Measuring Science. Capita Selecta of Current Main Issues. In: H.F. Moed, W. Glänzel, and U. Schmoch (eds.). *Handbook of Quantitative Science and Technology Research.* Dordrecht: Kluwer Academic Publishers, p. 19-50.

van Raan, A.F.J. (2005). Statistical Properties of Bibliometric Indicators Research Group Indicator Distributions and Correlations. *Journal of the American Society for Information Science and Technology*, to be published.

Rinia, E.J. Th.N. van Leeuwen, H.G. van Vuren, and A.F.J. van Raan (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria. Evaluation of condensed matter physics in the Netherlands. *Research Policy* 27, 95-107.

SJTU (2003). *Academic Ranking of World Universities -2003*. Shanghai Jiao Tong University, Institute of Higher Education, see http://ed.sjtu.edu.cn/rank/2003/2003main.htm.

SJTU (2004). *Academic Ranking of World Universities -2004*. Shanghai Jiao Tong University, Institute of Higher Education, see http://ed.sjtu.edu.cn/rank/2004/2004Main.htm.

SJTU (2005). *Academic Ranking of World Universities -2005*. Shanghai Jiao Tong University, Institute of Higher Education, see http://ed.sjtu.edu.cn/rank/2005/ARWU2005Main.htm.

Spiewak, M. (2005). Neues aus der Ranking-Schmiede (News from the Ranking-Smithy), *Die Zeit* Nr. 8, February 17, 2005, http://www.zeit.de/2005/08/B-Uni-Ranking, retreived May 2005.

THES (2004). World University Rankings: The World's Top 200 Universities. *The Times Higher Education Supplement*, November 5, 2004; University rankings cause a global stir. *The Times Higher Education Supplement*, November 12; 2004; The best of the bright sparks. *The Times Higher Education Supplement*. December 10, 2004.

THES (2005). World University Rankings. Who is Number One? *The Times Higher Education Supplement*, February 24, 2005; April 29, 2005; May 2, 2005, see http://www.thes.co.uk/worldrankings/ .

Tijssen, R.J.W., M.S. Visser, and T.N. van Leeuwen (2002). Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference? *Scientometrics* 54 (3), 381-397.

UMass (2004). Research Reputation Spurs World Ranking, University of Massachusetts Research ACCESS, *Expanding the Capacity for Research & Innovation*, Vol. 1 (5), December 2004, http://www.umass.edu/research/access/Vol1_Iss5_12-04.html, retrieved May 2005.

Vinkler, P. (1998). Comparative investigation of frequency and strength of motives toward referencing, the reference threshold model- comments on theories of citation? *Scientometrics* 43, 107-127.

VSNU (2002). *Quality Assessment of Research. Chemistry and Chemical Engineering Research in the Netherlands*. Utrecht: VSNU. The bibliometric part of this report is available via the CWTS website: www.cwts.nl.

Weingart, P. (2004). *Impact of bibliometrics upon the science system: inadvertent consequences?* In: H.F. Moed, W. Glänzel and U. Schmoch (eds). Handbook on Quantitative Science and Technology Research. Dordrecht (The Netherlands): Kluwer Academic Publishers, 2004.

Wheeler, J. (2005). Making sense of university world league tables. King's College London International News E-zine, Nr. 2, March 2005, http://www.kcl.ac.uk/depsta/sreo/ezine, retrieved May 2005. For the code of practice on publications, see www.kcl.ac.uk/college/policyzone/ .

Appendix

A.1 Publication output and impact indicators

During the past decades, CWTS has developed a number of bibliometric output and impact indicators. These indicators are calculated from the Thomson Scientific Citation Indexes as described in Chapter 2 of this paper. In the following sections we describe the main features of the CWTS bibliometric methodology. We give the symbols of the indicators as they are used in our work.

The *first* indicator is the total number of papers published by a research unit (e.g., a research group, department, institute, university) during a specific period ***(P)***. We consider only papers classified as *normal articles*, *letters*, *notes*, and *reviews*. Meeting abstracts, corrections, and editorials are *not* included. If a paper is published in a journal for which no citation data are available, or that is not assigned to a CI field (see footnote 7 in this paper), this paper will not be considered in the calculation of the indicators presented below.

The next two indicators are the total number of citations received, *without* (***C***) and *with* self-citations (***Ci***). A self-citation is a citation given in a publication of which at least one author (first author or co-author) is also an author of the cited paper (first author or co-author). As an indication of the self-citation rate we present the percentage of self-citations ***(% Selfcits***), relative to the total number of citations received. The *fourth* indicator is the average number of citations per publication calculated while self-citations are not included (***CPP***). A *fifth* indicator is the percentage of articles not cited during the time period considered (***%Pnc***), excluding self-citations.

A.2 International citation reference values

Next, two international reference values are computed. A first value represents the mean citation rate of the journals in which the research unit has published (***JCSm,*** the ***m***ean **J**ournal **C**itation **S**core). The ***JCSm*** takes into account both the type of paper (e.g., normal article, letter, review), as well as the specific years in which the papers were published. For example, the number of citations received during the period 2000 - 2004 by a *letter* published by a research unit in 2000 in journal X is compared to the average number of citations received during the same period (2000 - 2004) by all *letters* published in the same journal (X) in the same year (2000). Generally, a research unit publishes its papers in several journals rather than one. Therefore, we calculated a weighted average ***JCS*** indicated as ***JCSm***, with the weights determined by the number of papers published in each journal. Self-citations are excluded from the computation of ***JCSm***.

A unit U that has published two articles in journal Y in 2000 (***JCS*** = 3), and one letter in journal X in 2001 (***JCS*** = 0.3) obtains a ***JCSm*** of (3 + 3 + 0.3)/(1+ 1 +1) = 2.1 citations per publication.

The second reference value presents the mean citation rate of the fields in which the research unit publishes its papers (***FCSm,*** the mean **F**ield **C**itation **S**core). In calculating ***FCSm***, we used the same procedure as the one we applied in the calculation of ***JCSm,*** with journals replaced by fields. In most cases, a research unit is active in more than one field. In these cases we calculate a weighted average value, the weights being determined by the total number of papers the research unit has published in each field.

Suppose that journal X belongs to field Z, and that all 2001 letters in field Z are cited 1.5 times on average in 2001 - 2004, while journal Y belongs to field A where all 2000 articles are cited 0.6 times on average in 2000- 2004. Then, the unit U mentioned before obtains an ***FCSm*** score of (1.5 + 1.5 + 0.6)/(1 + 1 + 1) = 1.2 citations per publication.

When a journal is classified in multiple fields, citation scores are computed as follows. Basically, a paper in a journal classified in N fields is counted as 1/N paper in each field, and so are its citations and ***FCSm*** scores.

A.3 Main citation impact indicators

The two most important indicators compare the average number of citations received by the oeuvre of a research unit (***CPP***) with the two international reference values, namely the corresponding journal-based and the field-based mean citation scores (***JCSm*** and ***FCSm***, respectively), by calculating the ratio for both. Self-citations are excluded in the calculation of the ratios ***CPP/FCSm*** and ***CPP/JCSm***, to prevent that ratios are affected by divergent self-citation behaviour.

If the ratio ***CPP/JCSm*** is above 1.0, the mean impact of a research unit's papers exceeds the mean impact of all articles published in the journals in which the particular research unit has published its papers. A limitation of this indicator is that low impact publications published in low impact journals may get a similar score as high impact publications published in high impact journals.

The ***CPP/FCSm*** indicator is free from this limitation, because it takes the impact level of all journals in a specific field into account. Therefore, it seems the most suitable indicator of the international position of a research unit. If the ratio ***CPP/FCSm*** is above (below) 1.0, this means that the oeuvre of the research unit is cited more (less) frequently than an 'average' publication in the field(s) in which the research unit is active. ***FCSm*** constitutes a *worldwide field-specific average* in a specific (combination of) field(s). In this way, one may obtain an indication of the international position of a research unit, in terms of its impact compared to a world average. This world average is calculated for the total population of articles published in CI journals assigned to a particular field. As a rule, about 80 percent of these papers are authored by scientists from the United States, Canada, Western Europe, Australia and Japan. Therefore, this world average is dominated by the Western world. We apply a statistical test to establish whether the average impact of a research unit's publication oeuvre ***(CPP)*** differs significantly from the average impact of all papers in the research unit's journal set ***(JCSm)*** or from the worldwide field average ***(FCSm)*** in the field(s) in which the research unit is active

A.4 Top-of-the-distribution impact indicators

In the 'top-of-the-distribution' analysis, we first determine the citation distribution of the worldwide publication output at the level of scientific fields. Second, we compute the number of citations corresponding with specific percentile-thresholds, i.e. the number of citations required to reach the X% top-percentile of the field-specific citation distribution for a given (range of) publication year(s) and citation time-interval(s). We rank each publication according to the number of citations it received during a fixed period after publication, and identify those publications belonging to the X% most frequently cited papers in a given time-period. Threshold X is usually set at 1%, 5%, or 10%. Next, we calculate indicator ***Ptop***, the number of papers a research unit has within the worldwide top X% cited papers for a specific publication year, document type, and field. This method is a substantial advancement of the bibliometric methodology, as the rank assigned to papers is based on the actual impact distribution of *all* similar papers, and self-citations are excluded.

Finally, the ***A/E(Ptop)*** indicator marks the relative contribution to the X% most frequently cited papers. It is calculated as the ratio of the ***Ptop*** and ***E(Ptop)***, the latter is the statistically *expected* number of highly cited papers (i.e., papers with the top-X% of the citation distribution). Thus, a value above (below) one indicates a relatively high (low) contribution to the X% most frequently cited papers. This indicator directly shows whether the number of highly cited papers of a research group, department, institute, or university, is higher or lower than expected on the basis of the total publication output.